

MATH5805 Advanced Time Series Analysis

Session 1 2014

William T.M. Dunsmuir

May 21, 2014

Topic: General Information

Contact Details

- ▶ Lecturer: William Dunsmuir
- ▶ Room: RC-2057
- ▶ Phone: 9385 7035
- ▶ email: W.Dunsmuir@unsw.edu.au
- ▶ Consultation: Wednesday 4-5pm

Course Outline

Please read the Course Outline, including the attachment on School of Mathematics and Statistics policies, and let me know if you have any questions

This course will cover topics in the general area of nonlinear non-Gaussian time series modelling, filtering, smoothing and forecasting. The topics covered are applicable to modelling a single response time series with covariates. The response series can be discrete valued (e.g. Poisson, binary, binomial counts) or continuous valued (e.g. stochastic volatility, heavy tailed responses). Applications range from assessing the impact of covariates, particularly policy changes, relevant to policy evaluation, and modelling financial time series at various observation frequencies. Two major types of models are considered in the course: parameter driven (latent process) models in which the state process evolves according to a deterministic regression component plus a latent autocorrelated time series; and, observation driven models in which the latent process is replaced by a function of the history of the observed process. Approximately the first three quarters (9 weeks) of the course will be concerned with parameter driven models and the final quarter (3 weeks) will be concerned with observation driven models.

Topic: 1: Nonlinear non-Gaussian State Space Models

Nonlinear non-Gaussian State Space Models

Models with linear Gaussian signal

Exponential Family

Heavy-tailed Noise Distributions

Stochastic Volatility Models

Likelihood for the nonlinear non-Gaussian model

Software Resources for Course

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Importance sampling for smoothing and estimation

Nonlinear non-Gaussian State Space Models – 1

Let $Y_n = (y'_1, \dots, y'_n)'$ and $\alpha = (\alpha'_1, \dots, \alpha'_n)'$ where for $t = 1, \dots, n$:

$$y_t | \alpha \sim p(y_t | \alpha_t), \quad \alpha_{t+1} | \alpha \sim p(\alpha_{t+1} | \alpha_t), \quad \alpha_1 \sim p(\alpha_1) \quad (1.1)$$

Assume **conditional independence** for the response vectors given the state vectors:

$$p(Y_n | \alpha) = \prod_{t=1}^n p(y_t | \alpha_t) \quad (1.2)$$

Assume **Markov** state vector transitions:

$$p(\alpha) = \prod_{t=1}^{n-1} p(\alpha_{t+1} | \alpha_t) \quad (1.3)$$

Nonlinear non-Gaussian State Space Models – 2

Variety of Model Types:

- ▶ **linear, Gaussian:** All distributions Gaussian, relationships between y_t and α_t and α_t and α_{t-1} linear.
- ▶ **nonlinear, Gaussian:** All distributions Gaussian, at least one relationships between y_t and α_t or α_t and α_{t-1} is nonlinear.
- ▶ **linear, non-Gaussian:** all relationships between y_t and α_t and α_t and α_{t-1} linear.
- ▶ **nonlinear, non-Gaussian:** At least one distribution is non-Gaussian and at least one of the relationships is nonlinear.

Models with linear Gaussian signal

Define the '**signal**' $\theta_t = Z_t \alpha_t$ where Z_t is a matrix or vector to be specified

$$p(y_t | \alpha_1, \dots, \alpha_n, y_1, \dots, y_{t-1}) = p(y_t | \theta_t) \quad (1.4)$$

and the state vector evolves as

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \stackrel{\text{indep}}{\sim} N(0, Q_t) \quad (1.5)$$

The quantities Z_t , T_t , R_t and Q_t are non-random matrices.

$p(y_t | \theta_t)$ can be nonlinear and/or non-Gaussian.

Special case: $p(y_t | \theta_t)$ normal and θ_t is linear in y_t gives the linear Gaussian model (see later for details).

Exponential family response distributions

$$p(y_t|\theta_t) = \exp[y_t\theta_t - b_t(\theta_t) + c_t(y_t)], \quad (1.6)$$

where $b_t(\theta_t)$ is assumed to be twice differentiable throughout this course and $c_t(y_t)$ is a function only of y_t .

We will mainly focus on discrete distributions such as the Poisson, negative binomial, binary and binomial.

However, the gamma distribution arises in modelling time series of durations for example.

Signal observed with non-Gaussian noise

$$y_t = \theta_t + \epsilon_t, \quad \epsilon_t \stackrel{\text{indep}}{\sim} p(\epsilon_t). \quad (1.7)$$

and $p(\epsilon_t)$ is a non-gaussian density.

Examples include: linear regression with non-normal errors, Gaussian autoregressive signal plus regression with non-normal observation noise, models contaminated by (additive) outliers in some or all of the components of the observation vector.

Stochastically evolving variance

$$y_t = \mu + \exp\left(\frac{1}{2}\theta_t\right)\epsilon_t, \quad \epsilon_t \stackrel{\text{indep}}{\sim} p(\epsilon_t) \quad (1.8)$$

where the mean μ is fixed.

This is the stochastic volatility model and is the ‘parameter driven’ analogue of the ‘observation driven’ generalized autoregressive conditional heteroscedasticity (GARCH) model.

Nonlinear non-Gaussian State Space Models

Models with linear Gaussian signal

Exponential Family

General Specification

Poisson family

Binomial and Binary Responses

Binary Count

Negative Binomial and Multinomial

Heavy-tailed Noise Distributions

Stochastic Volatility Models

Likelihood for the nonlinear non-Gaussian model

Software Resources for Course

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Exponential Family Response Models – 1

In the exponential family density (1.6) let

$$\dot{b}_t(\theta_t) = \frac{\partial b_t(\theta_t)}{\partial \theta_t} \quad \ddot{b}(\theta_t) = \frac{\partial^2 b_t(\theta_t)}{\partial \theta_t \partial \theta'_t} \quad (1.9)$$

Note that

$$E(y_t) = \dot{b}_t(\theta) \quad \text{and} \quad \text{Var}(y_t) = \ddot{b}_t(\theta). \quad (1.10)$$

NOTE: \ddot{b} must be a positive definite matrix for non-degenerate distributions. This is very important in order that certain algorithms (the Kalman Filter and Smoother) can be used for computations of likelihoods for this class of models.

Exponential Family Response Models –2

Notes:

- ▶ Expressions for the mean and variance (1.10) follow by differentiating $\int p(y_t|\theta_t)dy_t = 1$ once and twice.
- ▶ Standard identities are:

$$E\left[\frac{\partial \log p(y_t|\theta_t)}{\partial \theta_t}\right] = 0, \quad (1.11)$$

$$E\left[\frac{\partial^2 \log p(y_t|\theta_t)}{\partial \theta_t \partial \theta'_t}\right] + E\left[\frac{\partial \log p(y_t|\theta_t)}{\partial \theta_t}\right] \frac{\partial \log p(y_t|\theta_t)}{\partial \theta'_t} = 0$$

EXERCISE: Prove the above.

Exponential Family Response Models – 3

Poisson Counts, log-link

Put

$$\theta_t = \log \mu_t, \quad b_t = \exp(\theta_t), \quad c_t(y_t) = -\log y_t! \quad (1.12)$$

giving the density of y_t given the signal θ_t as

$$p(y_t|\theta_t) = \exp[y_t\theta_t - \exp(\theta_t) - \log y_t!] \quad (1.13)$$

Note that

$$\mu_t = E(y_t) = \dot{b}_t(\theta_t) = \exp(\theta_t) \quad (1.14)$$

$$\text{Var}(y_t) = \ddot{b}_t(\theta_t) = \exp(\theta_t) = \mu_t \quad (1.15)$$

so that the **conditional** mean and variance are the same.

Exponential Family Response Models – 4

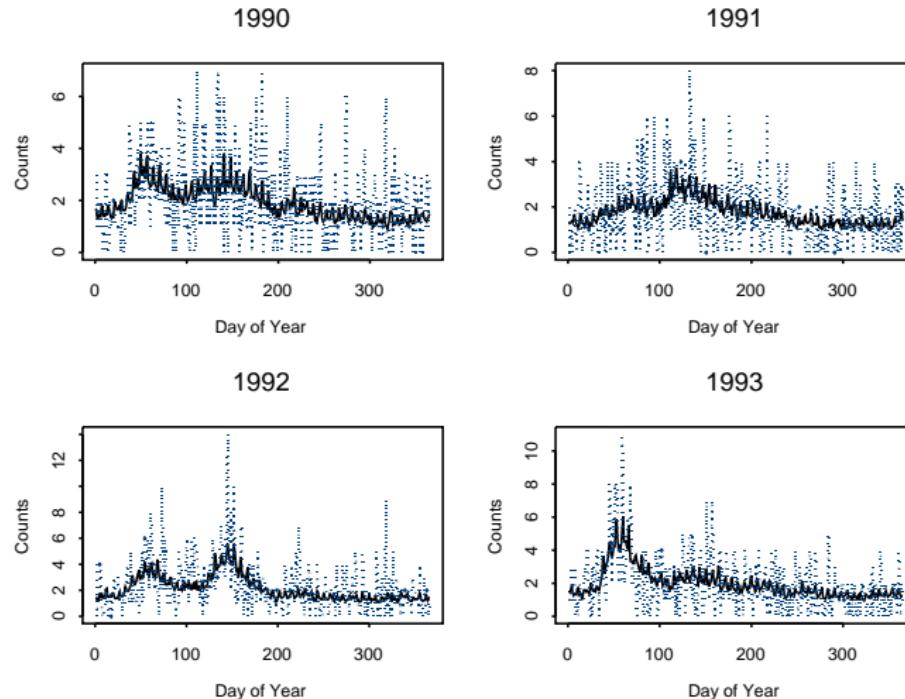
Poisson Counts, Examples

Examples of Poisson count time series Arise in areas including:

- ▶ Epidemiology/ public health: disease counts per unit of time (day, week, month etc.)
- ▶ Inventory management: demand for slow moving spare parts
- ▶ Risk management: injury and death counts due to accidents or self harm
- ▶ Financial transactions data: numbers of financial transactions per unit of trading time, price change in ticks in high frequency financial data

Exponential Family Response Models – 5

Poisson Counts, Daily asthma presentations at the Campbelltown Hospital Emergency Department



Exponential Family Response Models – 6

Binomial Counts

At time t there are k_t independent trials (experiments) in each of which there is success with probability π_t and the number of successes is y_t . Let

$$\theta_t = \log\left[\frac{\pi_t}{1 - \pi_t}\right], \quad b_t = k_t \log(1 + e^{\theta_t}), \quad c_t(y_t) = \log\binom{k_t}{y_t} \quad (1.16)$$

Then the density of y_t given the signal θ_t as

$$p(y_t|\theta_t) = \exp[y_t\theta_t - k_t \log(1 + \exp(\theta_t)) + \log\binom{k_t}{y_t}] \quad (1.17)$$

Note that

$$\mu_t = E(y_t) = k_t b_t(\theta_t) = k_t \frac{e^{\theta_t}}{1 + e^{\theta_t}} = k_t \pi_t \quad (1.18)$$

$$\text{Var}(y_t) = \ddot{b}_t(\theta_t) = k_t \left\{ \frac{e^{\theta_t}}{1 + e^{\theta_t}} - \left(\frac{e^{\theta_t}}{1 + e^{\theta_t}} \right)^2 \right\} = k_t \pi_t (1 - \pi_t) \quad (1.19)$$

Exponential Family Response Models – 7

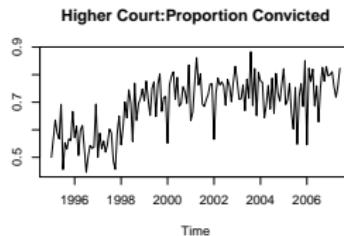
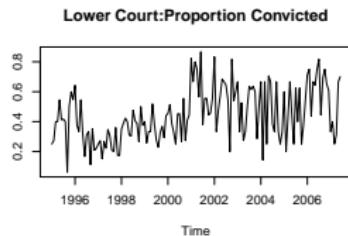
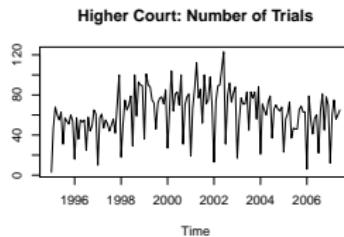
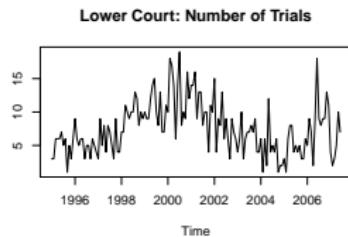
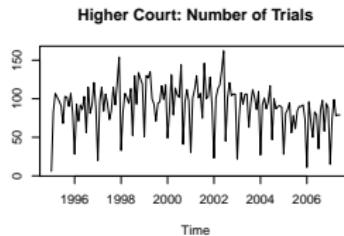
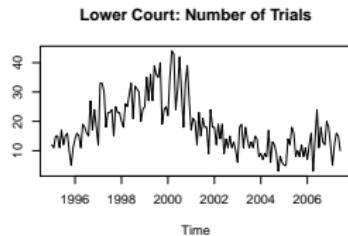
Binomial Counts, Examples

Examples of Binomial count time series: Arise in areas including:

- ▶ Meteorology: Numbers of rain days in each week or month
-DISCUSS INDEPENDENCE.
- ▶ Criminology: Numbers of convictions out of a given number of criminal trials per month for a category of crime.
- ▶ Musicology: Numbers of a panel of listeners who respond positively to musical features throughout a musical performance.

Exponential Family Response Models – 8

Binomial Counts, Monthly trials and convictions in higher and lower courts in NSW for Robbery



Exponential Family Response Models – 9

Binary Counts

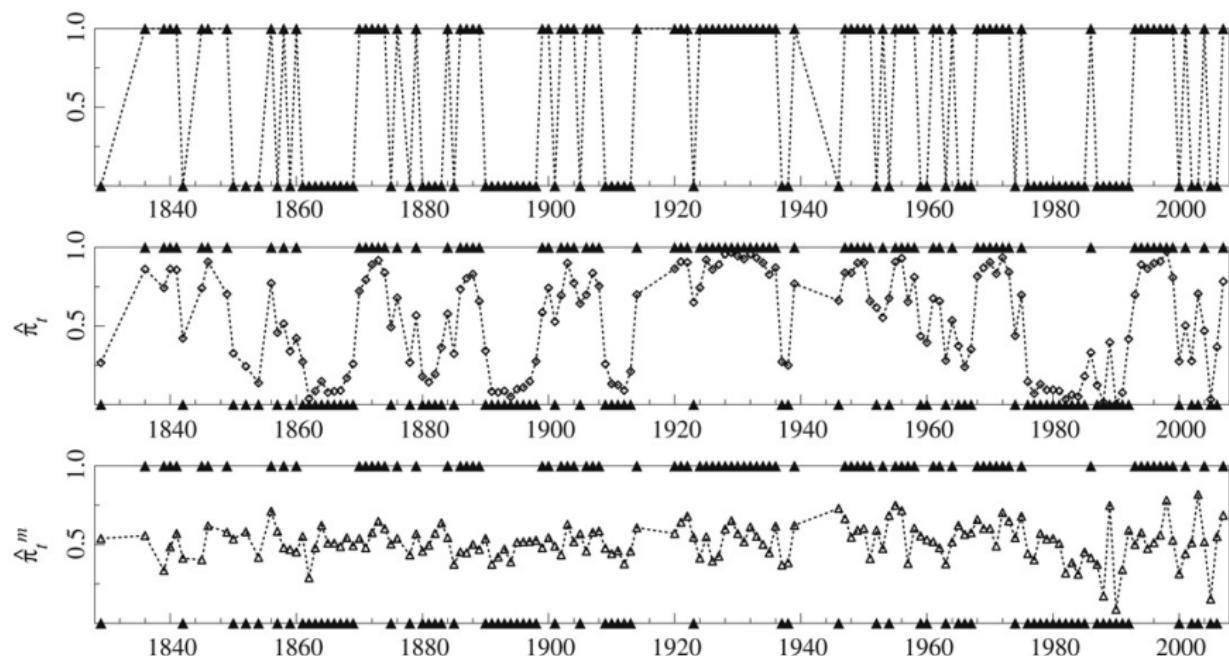
Here $k_t = 1$ for all t . Note that $c_t(y_t) = \binom{k_t}{y_t} = 1$ in this case.

Examples of Binary count time series: Arise in areas including:

- ▶ Meteorology: Rainfall occurrence, temperature exceedances, flooding etc.
- ▶ Musicology: Single listener responding positively to musical features throughout a musical performance.
- ▶ Sports performance: Winner of match between two teams - example Cambridge-Oxford Boat Race.
- ▶ Financial data: Does Price change at a single transaction? Does price rise or fall?
- ▶ Compliance/ positive test on drug testing for heroin addicts enrolled in a treatment program.

Exponential Family Response Models – 10

Binary Counts, Cambridge-Oxford Boat Race



Exponential Family Response Models – 11

Negative Binomial

Durbin and Koopman give details on the negative binomial response (Sec. 9.3.4):

- ▶ There are a number of ways to describe this. Durbin and Koopman present it as the number of independent trials needed to achieve a given number k_t of successes.
- ▶ The alternative is a gamma mixture of Poisson counts.
- ▶ Both lead to overdispersion where the variance is at least as large as the mean relative to the Poisson model.
- ▶ Examples include those that were listed under Poisson response above.

We will return to details of the negative binomial response as needed.

Exponential Family Response Models – 11

Multinomial

Durbin and Koopman give details on the multinomial response (Sec 9.3.5):

- ▶ Examples arise in opinion polling, studies of sleep patterns, numbers of ticks by which a stock price moves at each transaction.
- ▶ There are $h > 2$ possible outcomes (ordered or non-ordered) at each time point and the response y_t records which outcome is observed.
- ▶ Here the state is actually a vector of length $h - 1$ of the logits of the probability of being in a particular category relative to being in the others (or log odds). These could be cumulative logits designed for ordinal categories also.

We will return to these in detail are needed.

Nonlinear non-Gaussian State Space Models

Models with linear Gaussian signal

Exponential Family

Heavy-tailed Noise Distributions

t -distribution

Mixture of normals

Stochastic Volatility Models

Likelihood for the nonlinear non-Gaussian model

Software Resources for Course

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Heavy-tailed Noise Distributions

Consider the signal plus noise model (1.7) Some possibilities for $p(\epsilon_t)$ are:

- ▶ t-distribution
- ▶ Mixture of normals
- ▶ General error distribution

t-distribution

The logdensity is

$$\log p(\epsilon_t) = \log a(\nu) + \frac{1}{2} \log \lambda - \frac{\nu + 1}{2} \log(1 + \lambda \epsilon_t^2) \quad (1.20)$$

where ν is the degrees of freedom and, for $\nu > 2$ is **not** necessarily integer valued:

$$a(\nu) = \frac{\Gamma(\frac{\nu}{2} + \frac{1}{2})}{\Gamma(\frac{\nu}{2})}, \quad \lambda^{-1} = (\nu - 2)\sigma_\epsilon^2, \quad \sigma_\epsilon^2 = \text{Var}(\epsilon_t).$$

Notes:

- ▶ $E(\epsilon_t) = 0$ and $\text{Var}(\epsilon) = \sigma_\epsilon^2$ for all $\nu \in (2, \infty)$.
- ▶ ν , σ_ϵ^2 and, hence, λ can vary over time t .

Mixture of normals

Mixes a moderate variance normal (occurring λ^* of times) with a very large (by a factor of χ) variance normal (occurring $1 - \lambda^*$ of the time) to get density

$$p(\epsilon_t) = \lambda^* \frac{1}{(2\pi\sigma_\epsilon^2)^{\frac{1}{2}}} \exp\left(-\frac{\epsilon_t^2}{2\sigma_\epsilon^2}\right) + (1-\lambda^*) \frac{1}{(2\pi\chi\sigma_\epsilon^2)^{\frac{1}{2}}} \exp\left(-\frac{\epsilon_t^2}{2\chi\sigma_\epsilon^2}\right) \quad (1.21)$$

Again, if necessary for modelling, λ^* and χ can vary with t through a model of some type.

Nonlinear non-Gaussian State Space Models

Models with linear Gaussian signal

Exponential Family

Heavy-tailed Noise Distributions

Stochastic Volatility Models

Overview of Volatility Modelling

Basic Stochastic Volatility Model

Extensions: Stochastic Volatility Model

Other financial models

Likelihood for the nonlinear non-Gaussian model

Software Resources for Course

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Stochastic Volatility Modelling – General –1

Returns (changes between observations at successive times) of financial series are often uncorrelated from time to time but their squared values are not.

There are two key classes of models used to account for this phenomena:

- ▶ GARCH (generalized autoregressive conditional heteroscedastic models) in which the variance of the current observation is influenced by squares of past returns using a linear dynamic equation and variants of that. This is an **observation driven or single source of variability models**.
- ▶ Stochastic Volatility Models in which the changing variance is described by a latent (unobserved) serially correlated process. These are **parameter driven or double source of variability models**

Stochastic Volatility Modelling – General –1

Both can be related to continuous time processes which arise in theoretical financial modelling and both have a rich history of successful application in describing stylised facts about market returns (serially uncorrelated, heavier than normal kurtosis or ‘fat’ tails, volatility clustering).

For now we define the most basic stochastic volatility model. Later we will return to examining a variety of useful models as described in Durbin and Koopman Section 9.5.

Let $y_t = \log(P_t/P_{t-1}) = \log P_t - \log P_{t-1}$ be the first differences of the series of log prices of some financial asset.

Basic Stochastic Volatility Model –1

Observations:

$$y_t = \mu + \sigma \exp\left(\frac{1}{2}\theta_t\right)\epsilon_t, \quad \epsilon_t \stackrel{\text{indep}}{\sim} N(0, 1) \quad (1.22)$$

where the mean μ and σ are assumed fixed and unknown.

Signal: Unobserved log-volatility $\theta_t = Z_t\alpha_t$ where α_t is generated by (1.5).

Standard Autoregressive Signal Case: $\theta_t = \alpha_t$ where

$$\alpha_{t+1} = \phi\alpha_t + \eta_t, \quad \eta_t \stackrel{\text{indep}}{\sim} N(0, \sigma_\eta^2) \quad (1.23)$$

and $0 < \phi < 1$ with $\alpha_1 \sim N(0, \sigma_\eta^2/(1 - \phi^2))$

Basic Stochastic Volatility Model – 2

Conditional density of $y_t|\theta_t$:

$$\log p(y_t|\theta_t) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{1}{2}\theta_t - \frac{1}{2} \left(\frac{y_t - \mu}{\sigma \exp(\frac{1}{2}\theta_t)} \right)^2 \quad (1.24)$$

Conditional mean and variance:

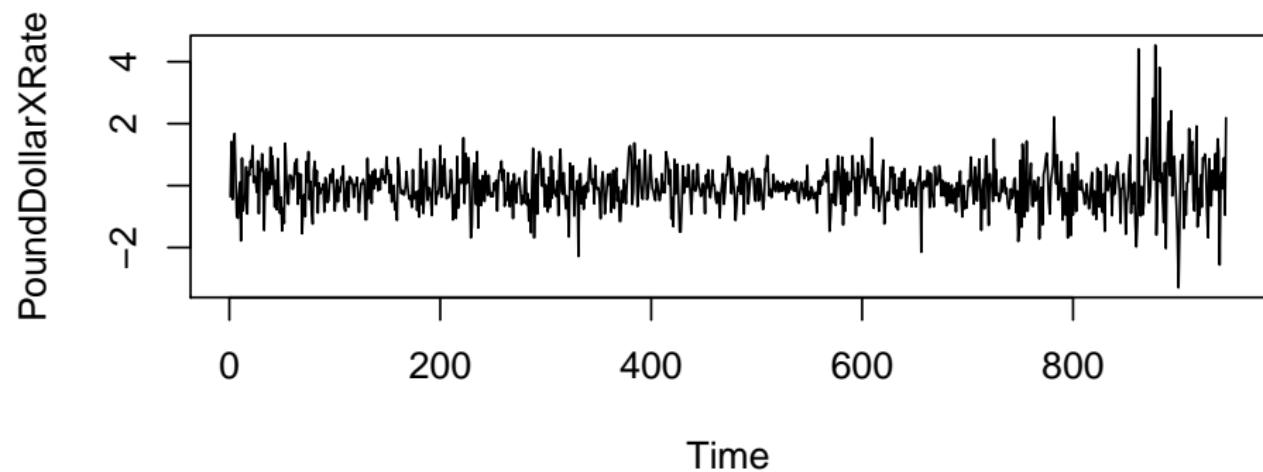
$$E(y_t|\theta_t) = \mu, \quad \text{Var}(y_t|\theta_t) = \sigma^2 \exp(\theta_t). \quad (1.25)$$

Unconditional mean and variance:

$$E(y_t) = \mu, \quad \text{Var}(y_t) = \sigma^2 E(e^{\theta_t}) = \sigma^2 \exp \left(\frac{1}{2} \frac{\sigma_\eta^2}{1 - \phi^2} \right). \quad (1.26)$$

Pound-Dollar Exchange Rate Data

From Durbin and Koopman



Extensions: Stochastic Volatility Model

- ▶ Multiple volatility factors
 - ▶ Regression and fixed effects
 - ▶ Heavy-tailed disturbances
 - ▶ Additive noise
 - ▶ Leverage effects
 - ▶ Stochastic volatility in the mean
 - ▶ Multivariate SV models
 - ▶ GARCH and SV models
- see Durbin and Koopman Section 9.5 for more details. We will return to these later.

Other Financial Models

A number of other financial modelling situations can be approached using the nonlinear non-Gaussian model class considered in this course. Examples include:

- ▶ Modelling durations between trades of a stock. Here the exponential or gamma response densities are relevant (both members of the exponential family of response distributions) together with a linear state equation in a function of the mean time between transactions. This is the parameter driven counterpart to the autoregressive conditional duration models promoted by Engle and Russell. Extensions to other 'survival' distributions are possible.
 - ▶ Trade frequencies counted in 'buckets' of time (10 seconds e.g.) and modelled using Poisson or negative binomial response distributions.
 - ▶ Credit risk models using binary, binomial and multinomial response models.
- see Durbin and Koopman Section 9.6 for additional details.

Nonlinear non-Gaussian State Space Models

Models with linear Gaussian signal

Exponential Family

Heavy-tailed Noise Distributions

Stochastic Volatility Models

Likelihood for the nonlinear non-Gaussian model

General form of Likelihood as an Integral

Importance Sampling

Need to study linear Gaussian State Space Models

Software Resources for Course

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Likelihood for the nonlinear non-Gaussian model – 1

Consider the general models with conditionally independent non-Gaussian response distributions and Markov state transitions defined by (1.1) and (1.3).

Let ψ denote the collection of all unknown parameters needed to specify the densities $p(Y_t|\alpha_t; \psi)$ and $p(\alpha; \psi)$. Usually the parameter vector ψ will separate into components for $p(Y_t|\alpha_t; \psi_1)$ and $p(\alpha; \psi_2)$ but we don't need to be concerned with those details at this stage.

Given observations y_t for $t = 1, \dots, n$ the likelihood for ψ is

$$L(\psi) = \int p(\alpha, Y_n; \psi) d\alpha = \int \prod_{t=1}^n p(y_t|\alpha_t; \psi) \prod_{t=1}^n p(\alpha_t|\alpha_{t-1}; \psi) d\alpha \quad (1.27)$$

which involves a potential massively high dimensional integral – not untypical these days for this to be 1000's or 10,000's.

Likelihood for the nonlinear non-Gaussian model – 2

Some key challenges for rapid and accurate estimation of ψ :

- ▶ How to do the integration needed to calculate the likelihood for a given ψ ?
- ▶ How to iterate in ψ -space to find the global maximizer $\hat{\psi}$ (the maximum likelihood estimator)?

A crude Monte Carlo approach would replace the integral by a sum over many simulated sample paths for the state process α_t but this is very inefficient.

Why? Simply put, most simulations from the α_t process will not be relevant to the particular one that generated the data Y_n we have.

What we want instead is a way to draw relevant samples.

Importance Sampling – 1

Let $g(\alpha, Y_n)$ be an approximating density to $p(\alpha, Y_n)$ selected in some way (details on good choices during course).

Rewrite the likelihood (1.27) as

$$\begin{aligned}L(\psi) &= \int \frac{p(\alpha, Y_n; \psi)}{g(\alpha, Y_n)} g(\alpha, Y_n) d\alpha \\&= g(Y_n) \int \frac{p(\alpha, Y_n; \psi)}{g(\alpha, Y_n)} g(\alpha | Y_n) d\alpha \\&= L_g(\psi) E_g[w(\alpha, Y_n; \psi)]\end{aligned}$$

where $L_g(\psi) = g(Y_n; \psi)$ is the marginal distribution for the observations in the approximating model,
 $w(\alpha, Y_n; \psi) = p(\alpha, Y_n; \psi)/g(\alpha, Y_n; \psi)$ are called ‘importance weights’ and the expectation E_g is computed using the conditional distribution $\alpha | Y_n \sim g(\alpha | Y_n; \psi)$.

Importance Sampling – 2

What is required to make this work:

- ▶ Good ways to choose the approximating model or distribution $g(\alpha, Y_n)$ which make the importance weights close to unity at least for ψ reasonably close to the MLE $\hat{\psi}$ and for most of the samples drawn from $\alpha|Y_n \sim g(\alpha|Y_n; \psi)$ the approximating conditional density of the state given the observations.
- ▶ $L_g(\psi) = g(Y_n; \psi)$ should be easy to compute – we aim to get this by integration in closed form.
- ▶ Fast ways to draw N samples $\alpha^{(k)}|Y_n$ efficiently from $g(\alpha|Y_n; \psi)$ to use in a finite sum approximation to:

$$E_g[w(\alpha, Y_n; \psi)] \approx \frac{1}{N} \sum_{k=1}^N w(\alpha^{(k)}, Y_n; \psi)$$

Importance of linear Gaussian State Space Models

In addition to obtaining the likelihood estimated of unknown parameters ψ we also may want to use Bayesian analysis and similar techniques can be used.

We are often interested in finding the 'smoothed' signal estimate (conditional mean or mode of $p(\alpha|Y_n)$) or in forecasting future values of this or the observed series Y_t for $t > n$.

Often the approximating conditional density $g(\alpha|Y_n)$ is a multivariate normal density which make drawing samples from it straightforward. However to find this density (or the approximating model) Kalman filtering and smoothing algorithms for linear state space models can be used.

Other approaches to getting an approximating multivariate normal density for $g(\alpha|Y_n)$ use Laplace approximation and the innovations algorithm from time series methodology or efficient importance sampling.

Nonlinear non-Gaussian State Space Models

Models with linear Gaussian signal

Exponential Family

Heavy-tailed Noise Distributions

Stochastic Volatility Models

Likelihood for the nonlinear non-Gaussian model

Software Resources for Course

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Importance sampling for smoothing and estimation

Software Resources for Course

- R
- ▶ Freeware, recommend you also get R Studio.
 - ▶ Will use generally for graphical displays and data analysis.
 - ▶ S+ uses almost identical commands and scripting language.
 - ▶ 'glarma' package available to install from CRAN servers – will use this for fitting observation driven models (last quarter of the course).

S+Finmetrics

- ▶ Copy available for student to use under our license agreement - sign a form.
- ▶ Will use the SsfPack to do linear Gaussian state space modelling.
- ▶ We (!) will write programs to use this to implement the nonlinear non-Gaussian model fitting.

Other specialist software - Davis & Rodrigues-Yam and others

Topic: 2: Local Level Model

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models



Objective

The objective of this lecture is to introduce the simplest non-trivial example of a state space model, **the local level model** and use this to introduce and illustrate all the key points of filtering, smoothing, initialization and forecasting.

This will provide an graceful springboard to understanding the more general linear state space model.

Reference: DK Chapter 2.

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models



Local Level Model

This is the Random Walk Signal Plus Noise:

$$y_t = \alpha_t + \epsilon_t, \quad \epsilon_t \stackrel{\text{indep}}{\sim} N(0, \sigma_\epsilon^2) \quad (2.1)$$

$$\alpha_{t+1} = \alpha_t + \eta_t, \quad \eta_t \stackrel{\text{indep}}{\sim} N(0, \sigma_\eta^2) \quad (2.2)$$

The sequences of random variables ϵ_t and η_t are mutually independent. That is any collection of ϵ_t and any collection of η_t are independent. Additionally they are independent of α_1 .

States and Observations

States: $\alpha_1, \dots, \alpha_n$.

Observations: y_1, \dots, y_n .

Objectives:

- ▶ Assuming the parameters σ_ϵ^2 and σ_η^2 and the distribution of the initial α_1 are known, determine the probabilistic properties of the states $\alpha_1, \dots, \alpha_n$ given n observations y_1, \dots, y_n or some partial subset of these (missing data setting).
- ▶ Estimate unknown parameters if required using the observations.

Comments on Assumptions

Initial State We assume to begin with that $\alpha_1 \sim N(a_1, P_1)$ where a_1 and P_1 are known. This will be relaxed later.

Normal Dist's Under the above assumptions that the ϵ_t and the η_t are normally distributed we perform the required inference using classical and Bayesian methods.

Other Dist's When normality is relaxed but variances are assumed to exist, the same formulae can be obtained using minimum variance linear unbiased estimation.

Iterated form of RW signal plus noise

First note that by iteration back in time of (2.2) we have

$$\alpha_t = \alpha_1 + \sum_{j=1}^{t-1} \eta_j \quad (2.3)$$

and substitution in (2.1) we have

$$y_t = \alpha_1 + \sum_{j=1}^{t-1} \eta_j + \epsilon_t, \quad t = 1, \dots, n \quad (2.4)$$

The Multivariate Normal Perspective

Let $Y_n = (y_1, \dots, y_n)'$ be the column vector of the n observations.

Let $\mathbf{1}$ denote the n -vector of 1's. Then

$$Y_n \sim N(\mathbf{1}a_1, \Omega), \quad \text{where } \Omega = \mathbf{1}\mathbf{1}'P_1 + \Sigma \quad (2.5)$$

and the matrix Σ has elements:

$$\Sigma_{ij} = \begin{cases} (i-1)\sigma_\eta^2, & i < j \\ \sigma_\epsilon^2 + (i-1)\sigma_\eta^2, & i = j \\ (j-1)\sigma_\eta^2, & i > j \end{cases} \quad (2.6)$$

Exercise Prove (2.4) (2.5) and (2.6).

Joint Distributions etc of α 's and y 's

Exercise.

- ▶ Let Y_n be as before and let $\alpha = (\alpha_1, \dots, \alpha_n)'$ be the vector of states. Find the mean vector, covariance matrix and joint multivariate distribution of $(Y'_n, \alpha')'$ and hence the conditional distribution of $\alpha | Y_n$.
- ▶ Interpret the results. In particular, what does conditioning on Y_n do to improving knowledge of the distribution of α ?

The result of this exercise is the classical multivariate analysis approach to the problem of estimating α given the observations. As n increases the computations in this approach become more and more burdensome. We seek **recursive methods** based on the structure of the above model to obtain the posterior mean (and mode in this case) of the signal given the observations.

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

The Kalman filter

Regression Lemma

Example: Nile River Flow at Aswan Dam

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Kalman Filter for local level model –1

Basic Notation

Let

$$Y_{t-1} = (y_1, \dots, y_{t-1})', \quad t = 2, 3, \dots$$

Assume, for known a_t and P_t , that

$$\alpha_t | Y_{t-1} \sim N(a_t, P_t).$$

Since all distributions are normal any conditional and marginal distributions are normal. Hence

Filtering: $\alpha_t | Y_t \sim N(a_{t|t}, P_{t|t}).$

1-step Prediction: $\alpha_{t+1} | Y_t \sim N(a_{t+1}, P_{t+1})$

Objective: Calculate $a_{t|t}$ (filtered estimate), $P_{t|t}$ (variance of filtered estimate), a_{t+1} (one-step predictor of α_{t+1}) and P_{t+1} (variance of one step prediction) as y_t is observed and used to update.

Kalman Filter for local level model –2

Prediction Errors

One step prediction error of y_t given Y_{t-1} is denoted

$$v_t = y_t - a_t, \quad t = 1, \dots, n$$

Exercise. Show that

- ▶ $E(v_t | Y_{t-1}) = 0$
- ▶ $\text{Var}(v_t | Y_{t-1}) = P_t + \sigma_\epsilon^2 =: F_t$
- ▶ $E(v_t | \alpha_t, Y_{t-1}) = \alpha_t - a_t$
- ▶ $\text{Var}(v_t | \alpha_t, Y_{t-1}) = \sigma_\epsilon^2$ for $t = 2, \dots, n$

Also denote the **Kalman Gain** by $K_t = P_t/F_t = P_t/(P_t + \sigma_\epsilon^2)$.

Kalman Filter - Key recursions

Filtering:

$$a_{t|t} = a_t + K_t v_t, \quad P_{t|t} = K_t \sigma_\epsilon^2$$

and hence

$$p(\alpha_t | Y_t) = N(a_{t|t}, P_{t|t})$$

Prediction:

$$a_{t+1} = a_t + K_t v_t, \quad P_{t+1} = K_t \sigma_\epsilon^2 + \sigma_\eta^2$$

and hence

$$p(\alpha_{t+1} | Y_t) = N(a_{t+1}, P_{t+1})$$

Exercise. Prove the above relationships for the conditional means and variances.

Summary: Kalman filter recursions for Local Level Model

Updating from time t to $t + 1$:

$$v_t = y_t - a_t, \quad F_t = P_t + \sigma_\epsilon^2, \quad (2.7)$$

$$a_{t|t} = a_t + K_t v_t, \quad P_{t|t} = P_t(1 - K_t), \quad (2.8)$$

$$a_{t+1} = a_t + K_t v_t, \quad P_{t+1} = P_t(1 - K_t) + \sigma_\eta^2 \quad (2.9)$$

for $t = 1, \dots, n$ and recall the Kalman gain

$$K_t = P_t/F_t = P_t/(P_t + \sigma_\epsilon^2)$$

is between 0 and 1.

Note: for $t = 1$, Y_{t-1} is deleted from above formulae.

REFLECTION: IS THIS INTUITIVELY REASONABLE? EXPLAIN THESE RECURSIONS IN PLAIN LANGUAGE

Other Derivations

- ▶ Using the regression lemma (conditional distributions for subcomponents of multivariate normal distributions) for bivariate normal distributions – see D&K Section 2.2.2.
- ▶ Bayesian derivation; α_t are treated as unknown parameters – see D&KSection 2.2.3.
- ▶ Minimum variance linear unbiased estimation derivation; normality is dispensed with – see D&KSection 2.2.4.

Example: Nile River Flow at Aswan Dam

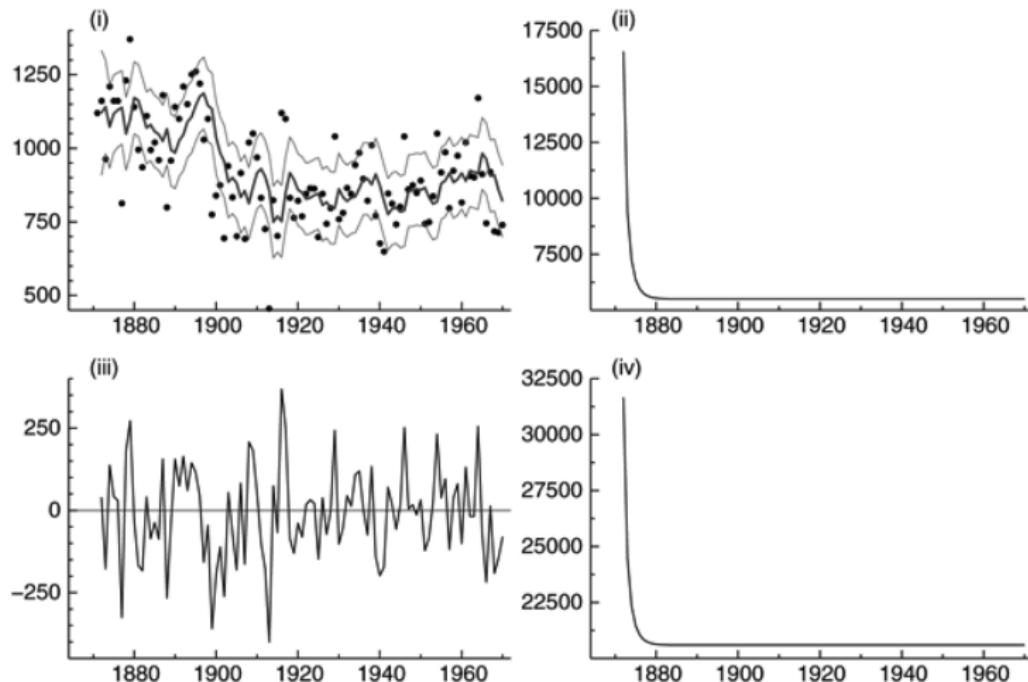


Figure : Fig. 2.1 Nile data and output of Kalman filter: (i) data (dots), filtered state at (solid line) and its 90% confidence intervals (light solid lines); (ii) filtered state variance $P_t | t$; (iii) prediction errors v_t ; (iv) prediction variance F_t .

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

- Cholesky decomposition

- Error recursions

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models

Forecast Error or Innovations

Note that $E(y_t | Y_{t-1}) = E(\alpha_t + \epsilon_t | Y_{t-1}) = E(\alpha_t | Y_{t-1}) = a_t$.

Also, $\text{Var}(y_t | Y_{t-1}) = \text{Var}(\alpha_t | Y_{t-1}) + \text{Var}(\epsilon) = P_t + \sigma_\epsilon^2 = F_t$.

Hence the error in forecasting y_t using past observations is

$v_t = y_t - a_t$ defined previously. Recall that v_t are zero mean, variance F_t normally distributed random variables.

Forecast errors v_t are also called **innovations** because they are the new part of y_t that is not predictable using past information Y_{t-1} .

Cholesky Decomposition –1

Applying (2.7) and (2.9) gives the recursions

$$v_1 = y_1 - a_1,$$

$$v_2 = y_2 - a_2 = y_2 - (a_1 + K_1(y_1 - a_1)) = y_2 - a_1 - K_1(y_1 - a_1),$$

$$v_3 = y_3 - a_1 - K_2(y_2 - a_1) - K_1(1 - K_2)(y_1 - a_1), \quad \text{etc.}$$

Note that the K_t depend only on P_1 (initial state variance) and the error variances σ_ϵ^2 and σ_η^2 .

Cholesky Decomposition -2

Hence the vector of innovations $v = (v_1, \dots, v_n)'$ can be expressed as the linear combination of Y_n :

$$v = C(Y_n - 1a_1)$$

where C is a lower triangular matrix with unit diagonals and lower triangular elements:

$$c_{i,i-1} = -K_{i-1}$$

$$c_{i,j} = -(1 - K_{i-1})(1 - K_{i-2}) \cdots (1 - K_{j+1})K_j$$

for $i = 2, \dots, m$ and $j = 1, \dots, i - 1$. Since $Y_n \sim N(1a_1, \Omega)$ it follows, using standard results for linear transformations of multivariate normals, that

$$v \sim N(\mathbf{0}, C\Omega C')$$

where Ω is defined in (2.5).

Cholesky Decomposition – 3

However, the innovations are zero mean and uncorrelated because for $s < t$

$$E(v_s v_t) = E[v_s E(v_t | Y_{t-1})] = E[v_s \times 0] = 0.$$

Hence v has a diagonal covariance matrix $F = \text{diag}(F_1, \dots, F_n)$ so that

$$\text{Cov}(v) = C\Omega C' = F$$

Hence, the Kalman filter applied to the local level model results in a **Cholesky factorization** (any lower triangular matrix which diagonalises a symmetric **matrix**) of the covariance matrix Ω for the observation vector Y_n

Note that $\Omega^{-1} = C'F^{-1}C$ a fact that is useful for likelihood estimation of unknown parameters – see later.

Error Recursions

Denote the **state estimation error** as:

$$x_t = \alpha_t - a_t, \quad \text{where} \quad \text{Var}(x_t) = P_t.$$

Note that the innovations are related to these by:

$$v_t = y_t - a_t = \alpha_t + \epsilon_t - a_t = x_t + \epsilon_t$$

so that the innovations can be thought of as a new observation process with states x_t evolving as

$$\begin{aligned} x_{t+1} &= \alpha_{t+1} - a_{t+1} \\ &= \alpha_t + \eta_t - (a_t + K_t v_t) \\ &= x_t + \eta_t - K_t(x_t + \epsilon_t) \\ &= L_t x_t + \eta_t - K_t \epsilon_t \end{aligned}$$

where $L_t = 1 - K_t = \sigma_\epsilon^2 / F_t$. Note $x_1 = \alpha_1 - a_1$.

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Smoothed state

Smoothed state variance

Example (continued)

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Smoothed State – 1

State Smoothing: Focus now on estimating the states $\alpha_1, \dots, \alpha_n$ given the complete set of observations Y_n .

Smoothed State $\hat{\alpha}_t = E(\alpha_t | Y_n) = E(\alpha_t | v)$

Smoothed State Variance $V_t = \text{Var}(\alpha_t | Y_n) = \text{Var}(\alpha_t | v)$

using the fact that the information in Y_n is the same as in $v = (v_1, \dots, v_n)'$. Note also this is equivalent to conditioning on $(Y'_{t-1}, v'_{t:n})'$ where $v_{t:n} = (v_t, \dots, v_n)'$.

Since all distributions are normal we have the conditional distribution of $(\alpha_t, v'_{t:n})' | Y_{t-1}$

$$\begin{bmatrix} \alpha_t \\ v_{t:n} \end{bmatrix} \sim N \left(\begin{bmatrix} a_t \\ 0 \end{bmatrix}, \begin{bmatrix} \text{Var}(\alpha_t | Y_{t-1}) & \text{Cov}(\alpha_t, v'_{t:n} | Y_{t-1}) \\ \text{Cov}(v_{t:n}, \alpha_t | Y_{t-1}) & \text{Cov}(v_{t:n} | Y_{t-1}) \end{bmatrix} \right) \quad (2.10)$$

Smoothed State – 2

Note that (see above) $\text{Cov}(v_{t:n}|Y_{t-1}) = F_{t:n}$ and hence using the general regression result for multivariate normal distributions we have

$$\alpha_t|Y_n \sim \alpha_t|v_{t:n}, Y_{t-1} \sim N(\hat{\alpha}_t, V_t) \quad (2.11)$$

where

$$\hat{\alpha}_t = a_t + \text{Cov}(\alpha_t, v'_{t:n}|Y_{t-1})F_{t:n}^{-1}v_{t:n} \quad (2.12)$$

and

$$V_t = \text{Var}(\alpha_t|Y_{t-1}) - \text{Cov}(\alpha_t, v'_{t:n}|Y_{t-1})F_{t:n}^{-1}\text{Cov}(v_{t:n}, \alpha_t|Y_{t-1}) \quad (2.13)$$

It remains to obtain recursive formulae for the conditional mean and variance.

NOTE changes to subscripts $v_{t:n}$, $K_{t:n}$

Smoothed State – 3

Because F^{-1} is a diagonal matrix the smoothed state in (2.12) can be rewritten as

$$\hat{\alpha}_t = a_t + \sum_{j=t}^n \text{Cov}(\alpha_t, v_j | Y_{t-1}) F_j^{-1} v_j \quad (2.14)$$

D&K(pp20-21) show that this smoothed state vector can be calculated by **backwards** recursion:

$$r_{t-1} = F_t^{-1} v_t + L_t r_t, \quad \hat{\alpha}_t = a_t + P_t r_{t-1}, \quad t = n, \dots, 1 \quad (2.15)$$

where $L_t = 1 - K_t = \sigma_\epsilon^2 / F_t$.

These are called the **state smoothing recursion**.

Smoothed State – 4

Also, because F^{-1} is diagonal and $P_t = \text{Var}(\alpha_t | Y_{t-1})$ the smoothed state variance in (2.13) can be rewritten as

$$V_t = P_t - \sum_{j=t}^n [\text{Cov}(\alpha_t, v_j | Y_{t-1})]^2 F_j^{-1} \quad (2.16)$$

D&K(pp21-22) show that these can be calculated using the (backward) state variance smoothing recursion:

$$N_{t-1} = F_t^{-1} + L_t^2 N_t, \quad V_t = P_t - P_t^2 N_{t-1}, \quad t = n, \dots, 1 \quad (2.17)$$

Example: Nile River Flow at Aswan Dam

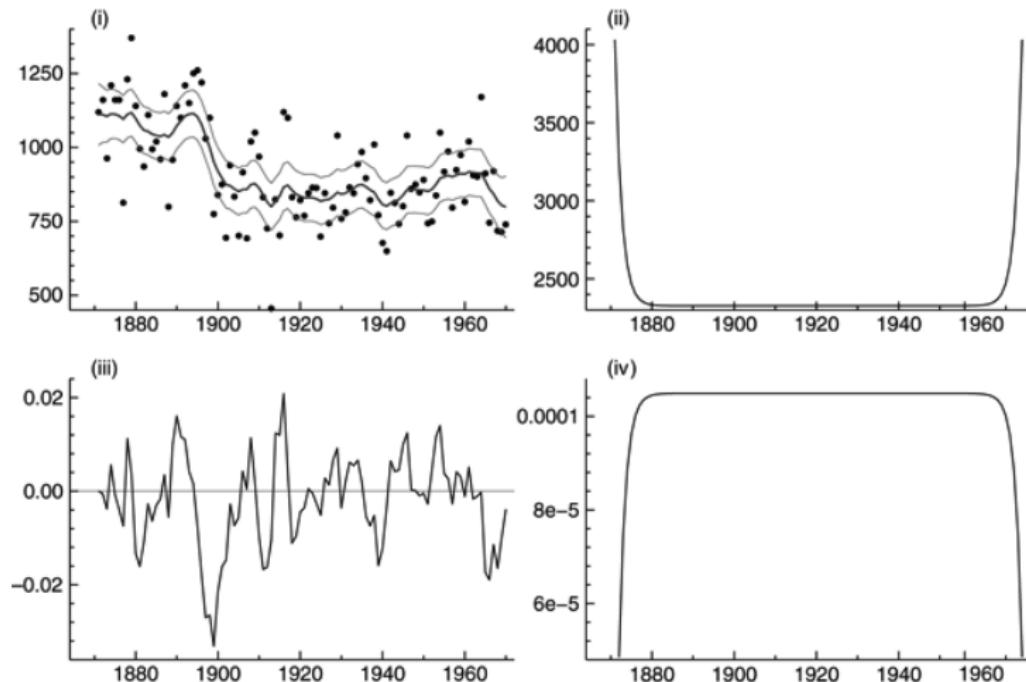


Figure : Fig. 2.2 Nile data and output of state smoothing recursion: (i) data (dots), smoothed state \hat{s}_t and its 90% confidence intervals;(ii) smoothed state variance V_t ;(iii) smoothing cumulant r_t ; (iv) smoothing variance cumulant N_t .

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Smoothed observation disturbances

Smoothed state disturbances

Example (continued)

Cholesky decomposition and smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Disturbance Smoothing - Overview

Smoothed observation disturbances:

$$\hat{\epsilon}_t = E(\epsilon_t | Y_n) = y_t - \hat{\alpha}_t \quad (2.18)$$

Useful for detecting additive outliers.

Smoothed state disturbances:

$$\hat{\eta}_t = E(\eta_t | Y_n) = \hat{\alpha}_{t+1} - \hat{\alpha}_t \quad (2.19)$$

Useful for detecting structural breaks in the state equation.

The formulae for these given below have computational advantages over direct calculation from quantities derived above. They will be derived more generally for the full linear state space model in subsequent lectures - see D&K Chapter 4 for details.

Smoothed Observation Disturbances and Variances

Recall $v_t = y_t - a_t$ and the backward recursion (2.15) for r_t . In terms of these we have the smoothed observation disturbances

$$\hat{\epsilon}_t = \sigma_\epsilon^2 u_t, \quad u_t = F_t^{-1} v_t - K_t r_t, \quad t = n, \dots, 1 \quad (2.20)$$

u_t is called the smoothing error.

We also have the smoothed variances

$$\text{Var}(\epsilon_t | Y_n) = \sigma_\epsilon^2 - \sigma_\epsilon^4 D_t, \quad D_t = F_t^{-1} + K_t^2 N_t, \quad t = n, \dots, 1 \quad (2.21)$$

Since v_t and r_t are independent, $\text{Var}(r_t) = N_t$, $\text{Var}(v_t) = F_t$

$$\text{Var}(u_t) = \text{Var}(F_t^{-1} v_t - K_t r_t) = F_t^{-1} + K_t^2 \text{Var}(r_t) = D_t \quad (2.22)$$

and hence $\text{Var}(\hat{\epsilon}_t) = \sigma_\epsilon^4 D_t$.

Smoothed State Disturbances and Variances

The smoothed mean of the disturbances is

$$\hat{\eta}_t = E(\eta_t | Y_n) = \sigma_\eta^2 r_t, \quad t = n, \dots, 1 \quad (2.23)$$

where r_t is calculated using the backward recursion (2.17) and the unconditional variance is:

$$\text{Var}(\hat{\eta}_t) = \sigma_\eta^4 \text{Var}(r_t) = \sigma_\eta^4 N_t.$$

We also have the smoothed variance

$$\text{Var}(\eta_t | Y_n) = \sigma_\eta^2 - \sigma_\eta^4 N_t, \quad t = n, \dots, 1 \quad (2.24)$$

where N_t are defined in (2.17)

Example: Nile River Flow at Aswan Dam

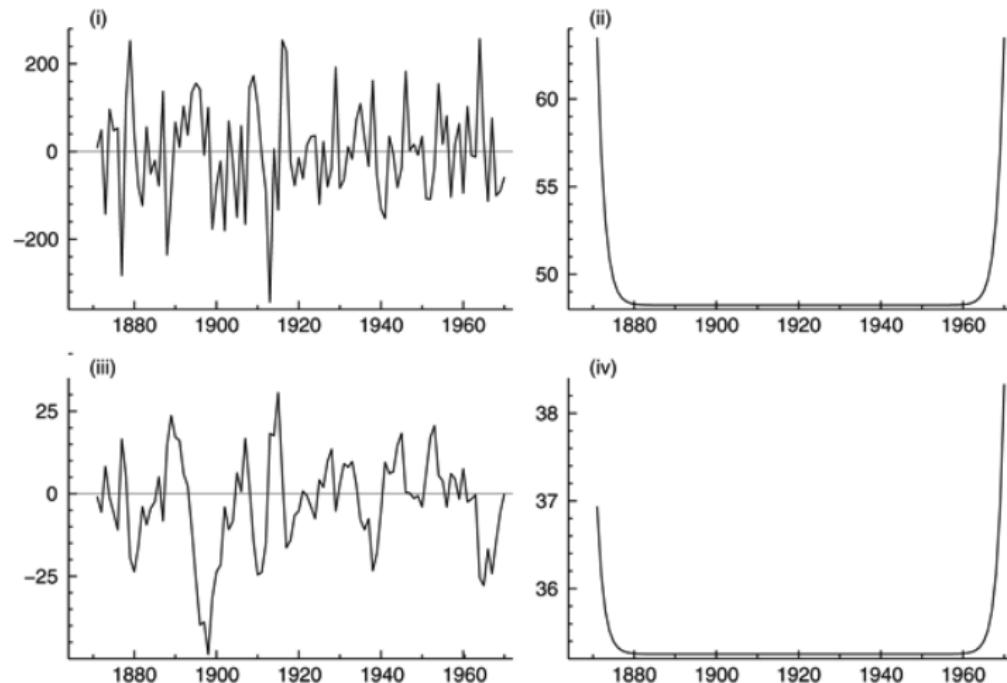


Figure : Fig. 2.3 Output and disturbance smoothing recursion: (i) observation error $\hat{\epsilon}_t$; ii observation error variance $\text{Var}(\epsilon_t | Y_n)$; (iii) state error $\hat{\eta}_t$; (iv) state error variance $\text{Var}(\eta_t | Y_n)$

Cholesky decomposition and smoothing – 1

An alternative, direct derivation to find $\hat{\epsilon} = E(\epsilon|Y_n)$, where $\hat{\epsilon} = (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)'$, is directly through general regression result for multivariate normal distributions. Note that

$$\begin{bmatrix} Y_n \\ \epsilon \end{bmatrix} \sim N \left(\begin{bmatrix} a_1 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega & \sigma_\epsilon^2 I_n \\ \sigma_\epsilon^2 I_n & \sigma_\epsilon^2 I_n \end{bmatrix} \right) \quad (2.25)$$

so that

$$\epsilon|Y_n \sim N \left(\sigma_\epsilon^2 \Omega^{-1} [y_n - a_1 1], \sigma_\epsilon^2 (I_n - \sigma_\epsilon^2 \Omega^{-1}) \right)$$

from which we get

$$\hat{\epsilon} = \sigma_\epsilon^2 \Omega^{-1} [y_n - a_1 1]$$

Exercise: prove (2.25)

Cholesky decomposition and smoothing – 2

Recall that $\Omega^{-1} = C'F^{-1}C$ and $v = C(Y - a_1\mathbf{1})$ giving

$$\hat{\epsilon} = \sigma_\epsilon^2 C' F^{-1} v$$

Recall from (2.20) that $\hat{\epsilon} = \sigma_\epsilon^2 u$ and hence an alternative expression for u in (2.20) is

$$u = C' F^{-1} C (Y_n - a_1 \mathbf{1}) = \Omega^{-1} (Y_n - a_1 \mathbf{1}).$$

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models



The Simulation Smoother

A simulation from the local level model (2.1) and (2.2) can be obtained as follows:

$$\epsilon_t^+ \stackrel{\text{indep}}{\sim} N(0, \sigma_\epsilon^2), \quad \eta_t^+ \stackrel{\text{indep}}{\sim} N(0, \sigma_\eta^2), \quad t = 1, \dots, n, \quad (2.26)$$

$$y_t^+ = \alpha_t^+ + \epsilon_t^+, \quad \alpha_t^+ + \eta_t^+, \quad t = 1, \dots, n. \quad (2.27)$$

However this does not give us samples conditional on the observations Y_n . To do this we use the following scheme. A **conditional simulation – the simulation smoother**. Using the disturbance smoothing equations (2.20) compute $\hat{\epsilon}_t = E(\epsilon_t | Y_n)$ and $\hat{\epsilon}_t^+ = E(\epsilon_t^+ | Y_n^+)$ and perform the **mean corrections**:

$$\tilde{\epsilon}_t = \epsilon_t^+ - \hat{\epsilon}_t^+ + \hat{\epsilon}_t \quad (2.28)$$

then based on these we get conditional (on Y_n) samples of

$$\tilde{\alpha}_t = y_t - \tilde{\epsilon}_t, \quad \tilde{\eta}_t = \tilde{\alpha}_{t+1} - \tilde{\alpha}_t, \quad t = 1, \dots, n \quad (2.29)$$

Simulating Sample Paths: Local Linear Model

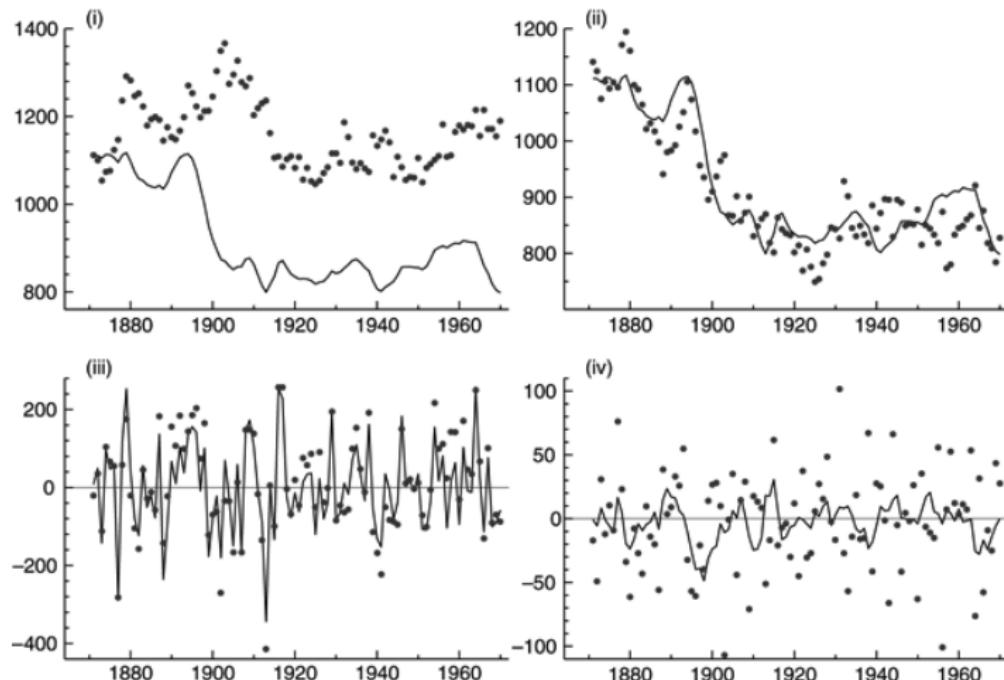


Figure : Fig. 2.4 Simulation:(i) smoothed state $\hat{\alpha}_t$ (solid line) and sample α_t^+ (ii) smoothed state $\hat{\alpha}_t$ (solid line) and sample $\tilde{\alpha}_t$ (dots); (iii) smoothed observation error $\hat{\epsilon}_t$ (solid line) and sample $\tilde{\epsilon}_t$ (dots); (iv) smoothed state error $\hat{\eta}_t$ (solid line) and sample $\tilde{\eta}_t$ (dots).

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Example (continued)

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models

Missing Observations – Background

- ▶ Many existing time series methods require that a complete record of observed data y_1, y_2, \dots, y_n be available for model fitting, smoothing, filtering and forecasting.
- ▶ Use of state space methods make handling of missing data (gaps in time where the time series is not observed) particularly easy.
- ▶ For most of the purposes of this course we don't really need this material but it is interesting and worth knowing about.
- ▶ However, since the concepts also underpin forecasting out of sample, we will consider this topic for the local level model of Chapter 2. Interested students can learn more in Chapter 4 for the general linear state space model.

Missing Observations – Filtering (1)

We start assuming there is **only one missing gap** with times $j = \tau, \dots, \tau^* - 1$ where the observations are not available, $1 < \tau < \tau^* \leq n$.

Filtering: Recall (see Basic Notation for Filtering and forecasting)

Filtering: $\alpha_t | Y_t \sim N(a_{t|t}, P_{t|t})$.

1-step Prediction: $\alpha_{t+1} | Y_t \sim N(a_{t+1}, P_{t+1})$

With missing data we continue to use this notation but with the understanding that Y_t means the information available up to time t . Hence, if $\tau \leq t < \tau^*$, Y_t is equivalent to $Y_{\tau-1}$.

Missing Observations – Filtering (2)

Hence, for $t = \tau, \dots, \tau^* - 1$

$$a_{t|t} = E(\alpha_t | Y_t) = E(\alpha_t | Y_{\tau-1}) = E(\alpha_\tau + \sum_{j=\tau}^{t-1} \eta_j | Y_{\tau-1}) = a_\tau$$

$$a_{t+1} = E(\alpha_{t+1} | Y_t) = E(\alpha_t | Y_{\tau-1}) = E(\alpha_\tau + \sum_{j=\tau}^t \eta_j | Y_{\tau-1}) = a_\tau$$

$$P_t | t = \text{Var}(\alpha_t | Y_t) = \text{Var}(\alpha_t | Y_{\tau-1})$$

$$= \text{Var}(\alpha_\tau + \sum_{j=\tau}^{t-1} \eta_j | Y_{\tau-1}) = P_\tau + (t - \tau) \sigma_\eta^2$$

$$P_{t+1} = \text{Var}(\alpha_{t+1} | Y_t) = \text{Var}(\alpha_{t+1} | Y_{\tau-1})$$

$$= \text{Var}(\alpha_\tau + \sum_{j=\tau}^t \eta_j | Y_{\tau-1}) = P_\tau + (t - \tau + 1) \sigma_\eta^2$$

Missing Observations – Filtering (3)

These last equations can be done recursively for $t = \tau, \dots, \tau^* - 1$ as follows

$$a_{t|t} = a_t, \quad P_{t|t} = P_t, \quad (2.30)$$

$$a_{t+1} = a_t, \quad P_{t+1} = P_t + \sigma_\eta^2. \quad (2.31)$$

Hence the filtering and forecasting recursions in (2.7), (2.8) and (2.9) can be applied but with $K_t = 0$ for the times of missing observations $t = \tau, \dots, \tau^* - 1$.

Note that when $K_t = 0$ the variance P_{t+1} grows linearly in the innovation variance σ_η^2 because factor $1 - K_t = 1$ the term P_t is not discounted during updates across these missing times.

Missing Observations – Smoothing (1)

State smoothing is derived similarly – D&Kpp29-30 for details.
The state smoothing backward recursions (2.15) are used in
periods of observed Y_t and becomes

$$r_{t-1} = r_t, \quad \hat{a}_t = a_t + P_t r_{t-1}, \quad t = \tau, \dots, \tau^* - 1 \quad (2.32)$$

through the missing time points.

Hence (2.15) can be used for all t again by taking $K_t = 0$ and
hence L_t for the missing time points.

Example continued

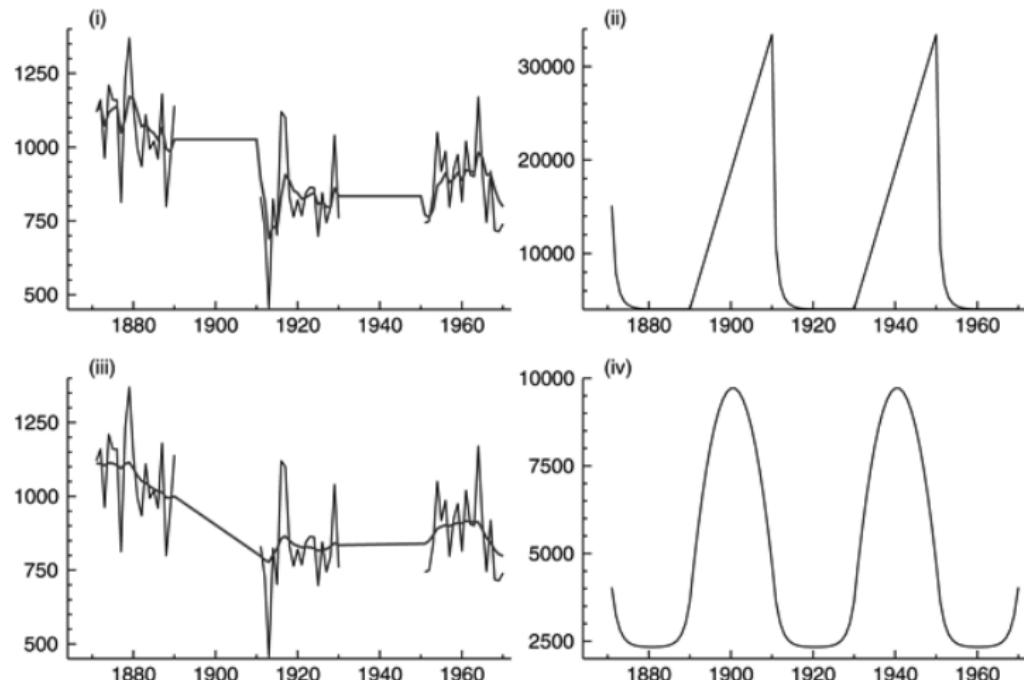


Figure : Fig. 2.5 Filtering and smoothing output when observations are missing: (i) data and filtered state a_t (extrapolation); (ii) filtered state variance P_t ; (iii) data and smoothed state \hat{a}_t (interpolation); (iv) smoothed state variance V_t

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Example (continued)

Example (continued)

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models

Forecasting future values –1

Forecast **future** y_t for $t = n + 1, \dots, J$ for a specified time horizon J using **minimum mean square error forecasts** to minimise

$$E([y_{n+j} - \bar{y}_{n+j}]^2 | Y_n).$$

This give the conditional expectation

$$\bar{y}_{n+j} = E(y_{n+j} | Y_n) = E(\alpha_{n+j} | Y_n) = \bar{a}_{n+j}.$$

since $E(\epsilon_{n+j} | Y_n) = 0$ and the notation $\bar{a}_{n+j} = E(\alpha_{n+j} | Y_n)$ is used.

The forecast error, $y_{n+j} - \bar{y}_{n+j}$ has zero mean and variance

$$\bar{F}_{n+j} = \text{Var}(Y_{n+j}) = \text{Var}(\alpha_{n+j} | Y_n) + \text{Var}(\epsilon_{n+j} | Y_n) = \bar{P}_{n+j} + \sigma_\epsilon^2$$

where we have used the notation $\bar{P}_{n+j} = \text{Var}(\alpha_{n+j} | Y_n)$.

Forecasting future values –2

Using the Kalman Filter to Forecast

By conceiving of the future values y_{n+1}, \dots, y_{n+J} as missing values the filtering equations for missing data (2.30) and (2.31) can be applied in a routine way by setting $K_t = 0$ for $t = n + 1, \dots, n + J$

Note that for the local linear model:

$$\bar{a}_{n+j+1} = \bar{a}_{n+j}, \quad \bar{P}_{n+j+1} = \bar{P}_{n+j} + \sigma_\eta^2, \quad j = 1, \dots, J - 1$$

and over the horizon $t + 1, \dots, t + J$

- ▶ the forecast stays horizontal, and
- ▶ the variance increases linearly in lead time j with coefficient σ_η^2 .

Example continued

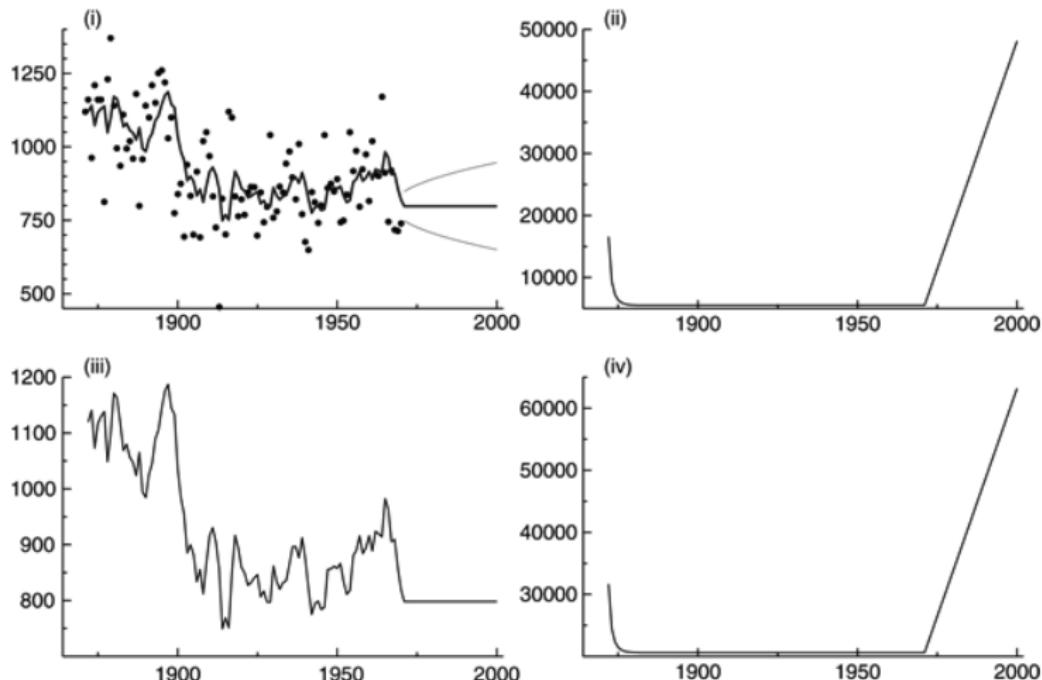


Figure : Fig. 2.6 Nile data and output of forecasting: (i) data (dots), state forecast a_t and 50% confidence intervals; (ii) state variance P_t ; (iii) observation forecast $E(y_t | Y_{t-1})$; (iv) observation forecast variance F_t .

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Linear State Space Models



Brief Notes on Initializing the Kalman Filter

So far we assumed the initial state $\alpha_1 \sim N(a_1, P_1)$ where a_1 and P_1 are known. In practice this is unlikely. Two approaches:

Diffuse Initial Prior Fix a_1 at an arbitrary value and let $P_1 \rightarrow \infty$.

Estimate Initial Values Assume α_1 is a constant (parameter) that is to be estimated along with other parameters by maximum likelihood from the first observation y_1 .

Both approaches (perhaps surprisingly) lead to the same initialization of the Kalman filter by putting $a_{1|1} = y_1$ and $P_{1|1} = \sigma_\epsilon^2$.

This equivalence extends to the general linear Gaussian model – see later. Refer to D&K§2.9 for full details.

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Logliklihood

Concentrated logliklihood

Example (continued)

Diagnostic Checking

The Loglikelihood – Fixed Initial Conditions

Assume that a_1 and P_1 specifying $y_1 \sim N(a_1, P_1)$ are known and fixed.

Using the joint density of $Y_n \sim N(a_1 1, \Omega)$ in (2.5) we get the loglikelihood for the unknown parameters $\phi = (\sigma_\epsilon^2, \sigma_\eta^2)$ as

$$\log L(\phi | Y_n) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Omega| - \frac{1}{2} (Y_n - a_1 1)' \Omega^{-1} (Y_n - a_1 1). \quad (2.33)$$

Using the Cholesky decomposition results $\Omega = CFC'$,
 $\log |\Omega| = \log |F| = \sum_{t=1}^n F_t$, $v = C(Y_n - a_1 1)$ (2.33) becomes

$$\log L(\phi | Y_n) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \left(\log F_t + \frac{v_t^2}{F_t} \right). \quad (2.34)$$

In the form (2.34) the likelihood can be calculated easily using the Kalman filter.

The Loglikelihood – Diffuse Initial Conditions

In practice a_1 and P_1 are unlikely to be known. One way to remove its influence from the likelihood $\log L$ defined in (2.34) by adding $\frac{1}{2} \log P_1$ to it and letting $P_1 \rightarrow \infty$. Thus:

$$\begin{aligned}\log L_d(\psi | Y_n) &= -\frac{1}{2} \lim_{P_1 \rightarrow \infty} \left(\log \frac{F_1}{P_1} + \frac{v_1^2}{F_1} \right) \\ &\quad - \frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{t=2}^n \left(\log F_t + \frac{v_t^2}{F_t} \right) \\ &= -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{t=2}^n \left(\log F_t + \frac{v_t^2}{F_t} \right) \quad (2.35)\end{aligned}$$

and using the facts: $F_1/P_1 \rightarrow 1$, $v_1^2/F_1 \rightarrow 0$ and v_t , F_t remain finite as $P_1 \rightarrow \infty$.

Since P_1 does not depend on the parameters ψ , $\hat{\psi}$ maximizing (2.34) converge to the maximizer of (2.35) as $P_1 \rightarrow \infty$.

The Concentrated Loglikelihood -1

An alternative parameterization of the local level model is $\sigma_\eta^2 = q\sigma_\epsilon^2$ and the basic parameters are now (q, σ_ϵ^2) .

Let $P_t^* = P_t/\sigma_\epsilon^2$ and $F_t^* = F_t/\sigma_\epsilon^2$.

Modifications to recursions (2.7) and (2.9) give $F_t^* = P_t^* + 1$ and $P_{t+1}^* = P_t^*(1 - K_t) + q$ initialised with $a_2 = y_1$ and $P_t^* = 1 + q$.

The diffuse loglikilihood is then

$$\log L_d = \frac{n}{2} \log(2\pi) - \frac{n-1}{2} \log \sigma_\epsilon^2 - \frac{1}{2} \sum_{t=2}^n \left(\log F_t^* + \frac{v_t^2}{\sigma_\epsilon^2 F_t^*} \right) \quad (2.36)$$

The Concentrated Loglikelihood -2

Noting that F_t^* do not depend on σ_ϵ^2 maximising (2.36) w.r.t σ_ϵ^2 gives

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-1} \sum_{t=2}^n \frac{v_t^2}{F_t^*} \quad (2.37)$$

Back substituting this in (2.36) gives the **concentrated** diffuse loglikelihood

$$\log L_{dc} = \frac{n}{2} \log(2\pi) - \frac{n-1}{2} - \frac{n-1}{2} \log \hat{\sigma}_\epsilon^2 - \frac{1}{2} \sum_{t=2}^n \log F_t^*. \quad (2.38)$$

This is maximised over the parameter q , which is sometimes loosely called the “signal-to-noise”.

Nile data: Maximum likelihood estimation

Nonlinear non-Gaussian State Space Models

Local Level Model

Local Level Model: Definition

Filtering

Forecast errors

State Smoothing

Disturbance Smoothing

Simulation

Missing Observations

Forecasting

Initialisation

Parameter Estimation

Diagnostic Checking

Outliers and structural breaks

Example (continued)

Example (continued)

Sample Moments of Forecast Errors

Under the normal distribution assumptions used in the local level model the standardised 1-step ahead forecast errors

$$e_t = \frac{v_t}{\sqrt{F_t}} \stackrel{\text{indep}}{\sim} N(0, 1).$$

Define central sample moments

$$m_1 = \frac{1}{n} \sum_{t=1}^n e_t, \quad m_q = \frac{1}{n} \sum_{t=1}^n (e_t - m_1)^q, \quad q = 2, 3, 4.$$

Skewness and kurtosis:

$$S = \frac{m_3}{\sqrt{m_2^3}} \stackrel{\text{approx}}{\sim} N(0, \frac{6}{n}), \quad K = \frac{m_4}{m_2^2} \stackrel{\text{approx}}{\sim} N(3, \frac{24}{n})$$

Diagnostic Tests Based on Forecast Errors –1

In practice the estimated forecast errors $\hat{e}_t = \hat{v}_t / \sqrt{\hat{F}_t}$ are used.
The above distributions continue to hold approximately for large n .

Normality assumptions can be checked using:

- ▶ The normal quantile-quantile plot with the Shapiro-Wilks test.
- ▶ Comparison of the combined

$$\hat{N} = n \left\{ \frac{\hat{S}^2}{6} + \frac{(\hat{K} - 3)^2}{24} \right\}.$$

with the $\chi^2_{(2)}$ upper tail quantiles

Diagnostic Tests Based on Forecast Errors –2

Heteroscedasticity (constant variance) can be checked using:

$$\hat{H}(h) = \frac{\sum_{t=n-h+1}^n \hat{e}_t^2}{\sum_{t=1}^h \hat{e}_t^2} \stackrel{\text{approx}}{\sim} F_{h,h}$$

(both tails compared).

Serial correlation can be checked using the Box-Ljung portmanteau statistic:

$$\hat{Q}(k) = n(n+2) \sum_{j=1}^k \frac{\hat{c}_j^2}{n-j} \stackrel{\text{approx}}{\sim} \chi_k^2$$

where the estimated serial correlations are

$$\hat{c}_j = \frac{1}{n\hat{m}_2} \sum_{t=j+1}^n (\hat{e}_t - \hat{m}_1)(\hat{e}_{t-j} - \hat{m}_1)$$

Example continued

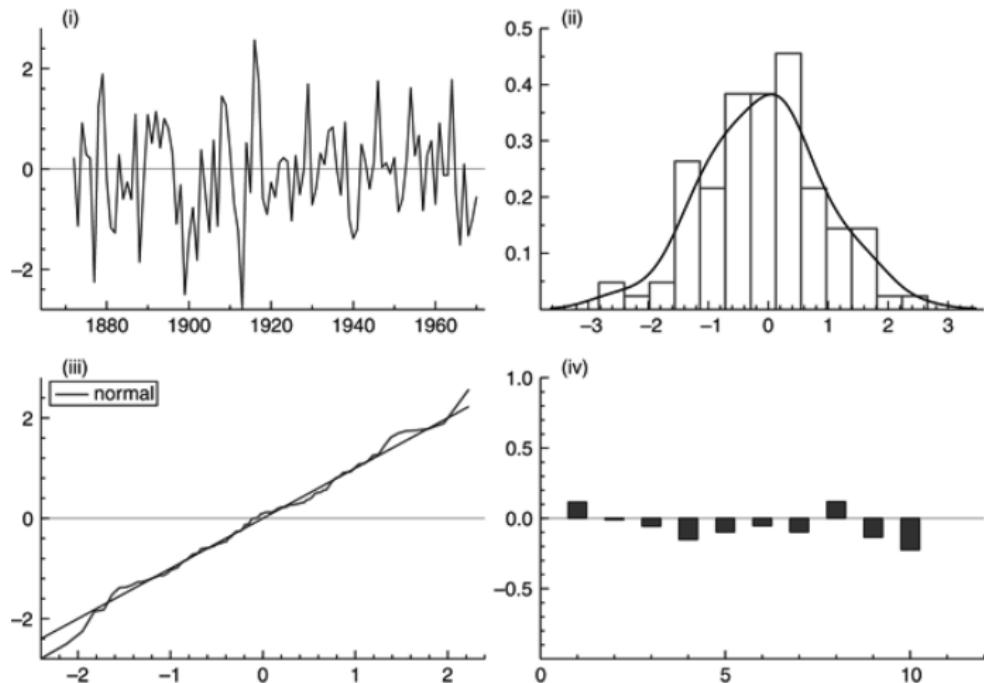


Figure : Fig. 2.7 Diagnostic plots for standardised prediction errors: (i) standardised residuals; (ii) histogram plus estimated density; (iii) ordered residuals; (iv) correlogram.

Detection of Outliers, Structural Breaks

Can be based on auxiliary residual (standardised smoothed residuals) defined as:

$$u_t^* = \frac{\hat{\epsilon}_t}{\sqrt{\text{Var}(\hat{\epsilon}_t)}} = \hat{D}_t^{-\frac{1}{2}} \hat{u}_t, \quad r_t^* = \frac{\hat{\eta}_t}{\sqrt{\text{Var}(\hat{\eta}_t)}} = \hat{N}_t^{-\frac{1}{2}} \hat{r}_t$$

Although these are not uncorrelated, large positive or negative values, can be used to indicate:

- an outlier using u_t^* .
- break in level using r_t^2 .

Example continued

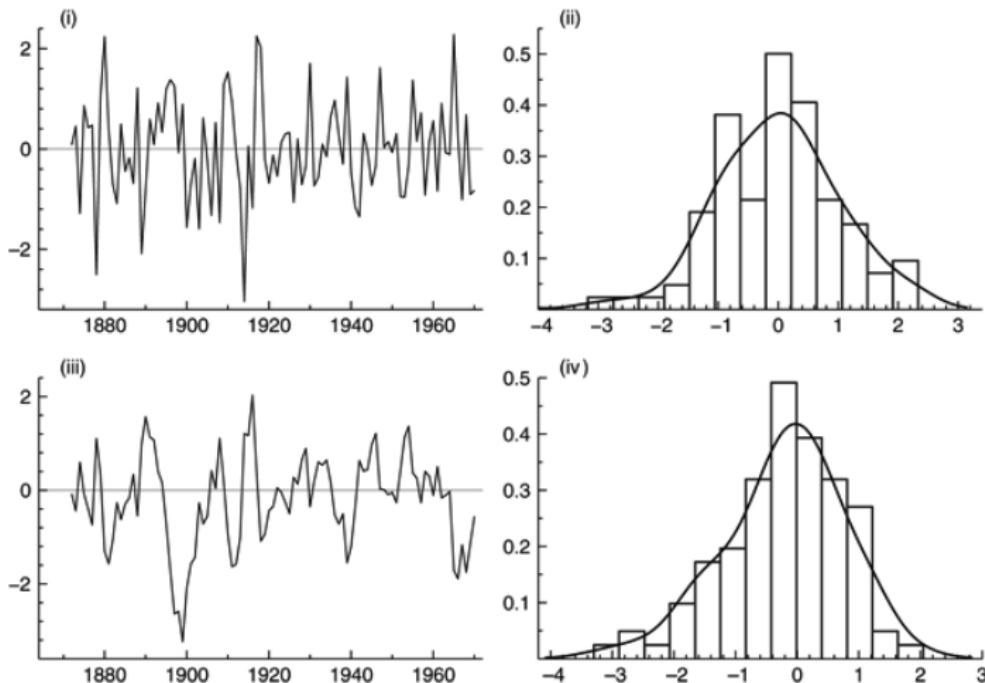


Figure : Fig. 2.8 Diagnostic plots for auxiliary residuals: (i) observation residual u_t^* ; (ii) histogram and estimated density for u_t^* ; (iii) state residual r_t^* ; (iv) histogram and estimated density for r_t^* .

Topic: 3: Linear State Space Models

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear SSS Model: General Form

Example 1: Structural Time Series Model

Including explanatory variables

ARIMA models

Regression Models

Other examples and comments

Nile River Flow: Local linear model with intervention term

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Linear SS Model: General Form

Observation Vector: The $p \times 1$ vector

$$y_t = Z_t \alpha_t + \epsilon_t, \quad \epsilon_t \stackrel{\text{indep}}{\sim} N(0, H_t), \quad t = 1, \dots, n \quad (3.1)$$

State Vector: The $m \times 1$ vector

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \stackrel{\text{indep}}{\sim} N(0, Q_t), \quad t = 1, \dots, n \quad (3.2)$$

- ▶ $\alpha_1 \sim N(a_1, P_1)$ independently to ϵ_t and η_t which are also assumed to be independent sequences.
- ▶ Matrices Z_t , T_t , R_t , H_t and Q_t depend on an unknown parameter vector ψ to be estimated.
- ▶ The matrices Z_t and T_{t-1} can depend on $y_{1:(t-1)}$.
- ▶ Dropping the normal distribution assumption, $\epsilon_t \sim (0, H_t)$ and $\eta_t \sim (0, Q_t)$ gives the general linear state space model.

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear SSS Model: General Form

Example 1: Structural Time Series Model

Including explanatory variables

ARIMA models

Regression Models

Other examples and comments

Nile River Flow: Local linear model with intervention term

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Example 1: Structural Time Series Model

Local Linear Trend Model – State Equation:

$$\begin{aligned}\mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\stackrel{\text{indep}}{\sim} N(0, \sigma_\xi^2) \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\stackrel{\text{indep}}{\sim} N(0, \sigma_\zeta^2)\end{aligned}\tag{3.3}$$

In matrix form:

$$\begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix}$$

Local Linear Trend Model – Observation Equation:

$$\begin{aligned}Y_t &= \mu_t + \epsilon_t, & \epsilon_t &\stackrel{\text{indep}}{\sim} N(0, \sigma_\epsilon^2) \\ &= (1 \quad 0) \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \epsilon_t\end{aligned}$$

Structural TS Model: Seasonal Component

Time Varying Seasonal Component of State (Assume s 'seasons' in period):

$$\gamma_{t+1} = - \sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \quad \omega_t \stackrel{\text{indep}}{\sim} N(0, \sigma_\omega^2) \quad (3.4)$$

(initialization required – see general approach to this later on).

Alternative forms (see D&K§3.2.2):

- ▶ Quasi random walk for each of s seasonal components adjusted to sum to zero.
- ▶ Trigonometric functions with random walk evolution of coefficients.
- ▶ Quasi random trinometric form.

Basic Structural TS Model – 1

The combined trend and seasonal components structural model can be put in the form (3.1), (3.2) by defining the state vector:

$$\alpha_t = (\mu_t, \nu_t, \gamma_t, \gamma_{t-1}, \dots, \gamma_{t-s+2})', \quad (3.5)$$

and system matrices:

$$\begin{aligned} Z_t &= (Z_{[\mu]}, Z_{[\gamma]}), & T_t &= \text{diag}(T_{[\mu]}, T_{[\gamma]}) \\ R_t &= \text{diag}(R_{[\mu]}, R_{[\gamma]}), & Q_t &= \text{diag}(Q_{[\mu]}, Q_{[\gamma]}) \end{aligned} \quad (3.6)$$

where

$$Z_{[\mu]} = (1, 0), \quad Z_{[\gamma]} = (1, 0, \dots, 0)$$

and

Basic Structural TS Model – 1

$$T_{[\mu]} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad T_{[\gamma]} = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & 0 \end{bmatrix},$$

$$R_{[\mu]} = I_2, \quad R_{[\gamma]} = (1, 0, \dots, 0)'$$

$$Q_{[\mu]} = \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}, \quad Q_{[\gamma]} = \sigma_\omega^2.$$

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear SSS Model: General Form

Example 1: Structural Time Series Model

Local linear trend model

Seasonal components

Local linear model combined with seasonal

Including explanatory variables

ARIMA models

Regression Models

Other examples and comments

Nile River Flow: Local linear model with intervention term

Linear State Space Model - Filtering, Smoothing and Forecasting

Explanatory variables and intervention effects

Regressors: x_{1t}, \dots, x_{kt} , coefficients β_1, \dots, β_k .

Intervention terms are a special case of regressors. Examples:

Step $w_t = 0$ if $t < \tau$, $w_t = 1$ if $t \geq \tau$.

Slope $w_t = 0$ if $t < \tau$, $w_t = 1 + t - \tau$ if $t \geq \tau$.

Pulse $w_t = 0$ if $t \neq \tau$, $w_t = 1$ if $t = \tau$.

Put in linear state space form by:

- ▶ augmenting the state vector in (3.5) with $\beta_t = (\beta_{1t}, \dots, \beta_{kt})'$;
- ▶ to keep $\beta_t = \beta$ for all t add a block diagonal I_k matrix to T_t and a zero block to R_t (no noise in the β updates); and,
- ▶ include the regressors in $Z_t = (Z_{[\mu]}, Z_{[\gamma]}, x_{1t}, \dots, x_{kt})$.

Exercise: Put the local level, linear trend with seasonal for quarterly data and a single covariate x_t in form of general state space.

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear SSS Model: General Form

Example 1: Structural Time Series Model

Including explanatory variables

ARIMA models

Regression Models

Other examples and comments

Nile River Flow: Local linear model with intervention term

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

ARIMA models – 1

Popularized by George Box and Gwillam Jenkins around 1970 with their influential book. Trends and seasonal components in series are not modelled (as in the structural model above) but are removed by differencing at lag 1 and at the seasonal lag s . The result is a stationary time series with serial dependence modelled using the autoregressive moving average class of models.

Differencing Operators: Let $\Delta y_t = y_t - y_{t-1}$ be lag 1 differencing and $\Delta_s y_t = y_t - y_{t-s}$ be seasonal differencing. If d repeated applications of Δ and D repeated applications of Δ_s are required to eliminate trend and seasonal non-stationarity (any order of application will result in the same outcome) we let

$$y_t^* = \Delta^d \Delta_s^D y_t$$

be the stationary series. **Autoregressive Moving Average Process:**

$$y_t^* = \sum_{j=1}^p \phi_j y_{t-j}^* + \zeta_t + \sum_{j=1}^q \theta_j \zeta_{t-j}, \quad \zeta_t \stackrel{\text{indep}}{\sim} N(0, \sigma_\zeta^2) \quad (3.7)$$

ARIMA models – 2

The above is denoted ARIMA(p, d, q) \times (0, D , 0) _{s} because there are no seasonal lag AR or MA terms. Note $p, q \geq 0$. Pure AR has $q = 0$, pure MA has $p = 0$. White noise has $p = q = 0$.

Rewrite for state space form using $r = \max(p, q + 1)$ and extend as needed the AR and MA coefficients using 0 coefficients:

$$y_t^* = \sum_{j=1}^r \phi_j y_{t-j}^* + \zeta_t + \sum_{j=1}^{r-1} \theta_j \zeta_{t-j}, \quad \zeta_t \stackrel{\text{indep}}{\sim} N(0, \sigma_\zeta^2) \quad (3.8)$$

ARIMA models – 3

Example 1: No differencing $d = D = 0$, general r .

$$\alpha_t = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \cdots + \phi_r y_{t-r+1} + \theta_1 \zeta_t + \cdots + \theta_{r-1} \zeta_{t-r+2} \\ \phi_3 y_{t-1} + \cdots + \phi_r y_{t-r+2} + \theta_1 \zeta_t + \cdots + \theta_{r-1} \zeta_{t-r+3} \\ \vdots \\ \phi_r y_{t-1} + \theta_{r-1} \zeta_t \end{pmatrix} \quad (3.9)$$

$$T_t = \begin{bmatrix} \phi_1 & 1 & & 0 \\ \vdots & & \ddots & \\ \phi_{r-1} & 0 & & 1 \\ \phi_r & 0 & \cdots & 0 \end{bmatrix}, \quad R_t = \begin{pmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{r-1} \end{pmatrix} \quad \eta_t = \zeta_{\textcolor{red}{t+1}}. \quad (3.10)$$

Finally, select $H_t = 0$ for the observation noise variance so that
 $\epsilon_t = 0$.

ARIMA models – 4

Example 2: No differencing $d = D = 0$, general $r = 2$.

State equation:

$$\begin{pmatrix} y_{t+1} \\ \phi_2 y_t + \theta_1 \zeta_{t+1} \end{pmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \theta_1 \zeta_t \end{pmatrix} + \begin{pmatrix} 1 \\ \theta_1 \end{pmatrix} \zeta_{t+1}$$

Observation equation:

$$y_t = Z_t \alpha_t, \quad H_t = 0, \quad \epsilon_t = 0$$

where $Z_t = (1, 0)$.

Exercise: Extend this to include $d = 1$ differencing.

ARIMA models – 5

Example 3: double differencing $d = 2$, $D = 0$, $p = 2$, $q = 1$

$$y_t = (1 \ 1 \ 1 \ 0) \alpha_t$$

$$\alpha_{t+1} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & \phi_1 & 1 \\ 0 & 0 & \phi_2 & 0 \end{bmatrix} \alpha_t + \begin{pmatrix} 0 \\ 0 \\ 1 \\ \theta_1 \end{pmatrix} \zeta t + 1$$

$$\alpha_t = \begin{pmatrix} y_{t-1} \\ \Delta y_{t-1} \\ y_t^* \\ \phi_2 y_{t-1}^* + \theta_1 \zeta_t \end{pmatrix}$$

Note that:

$$\Delta y_t = \Delta^2 y_t + \Delta y_{t-1}, \quad y_t = \Delta y_t + y_{t-1} = \Delta^2 y_t + \Delta y_{t-1} + y_{t-1}.$$

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear SSS Model: General Form

Example 1: Structural Time Series Model

Including explanatory variables

ARIMA models

Regression Models

Other examples and comments

Nile River Flow: Local linear model with intervention term

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Regression Models – 1

Basic regression model

$$y_t = X_t \beta + \epsilon_t, \quad \epsilon \stackrel{\text{indep}}{\sim} N(0, H_t)$$

Put $Z_t = X_t$, $T_t = I_k$, $R_t = Q_t = 0$ to specify $\alpha_t = \alpha_1 = \beta$

The Kalman filter applied to the state space model defined this way gives a recursive (in $t = 1, \dots, n$) method to arrive at the **generalized least squares estimator**:

$$\hat{\beta} = \left(\sum_{t=1}^n X_t' H_t^{-1} X_t \right)^{-1} \sum_{t=1}^n X_t' H_t^{-1} y_t$$

Regression Models – 2

Time varying coefficient regression model

To let the regression coefficient vector β vary through time the components are sometimes specified to follow a random walk

$\beta_{t+1} = \beta_t + \eta_t$ where $\eta_t \stackrel{\text{indep}}{\sim} N(0, Q)$ and Q is diagonal.

As a linear state space model choose: $\alpha_t = \beta_t$, $Z_t = X_t$,
 $T_t = R_t = I_k$.

Regression Models – 3

Regression with ARMA errors

$$y_t = X_t \beta + \xi_t.$$

where ξ_t follows an ARMA model (3.8). Define α_t as in (3.9) and let

$$\alpha_t^* = \begin{pmatrix} \beta_t \\ \alpha_t \end{pmatrix}$$

where $\beta_t = \beta$. The state equation corresponding to (3.10) is $\alpha_{t+1} = T\alpha_t + R\eta_t$ so let

$$T^* = \begin{bmatrix} I_k & 0 \\ 0 & T \end{bmatrix}, \quad R^* = \begin{bmatrix} 0 \\ R \end{bmatrix}, \quad Z_t^* = (X_t \ 1 \ 0 \ \cdots \ 0).$$

Hence

$$y_t = Z_t^* \alpha_t^*, \quad \alpha_{t+1}^* = T^* \alpha_t^* + R^* \eta_t$$

is the state space form (3.1), (3.2).

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear SSS Model: General Form

Example 1: Structural Time Series Model

Including explanatory variables

ARIMA models

Regression Models

Other examples and comments

Nile River Flow: Local linear model with intervention term

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

More Examples of Models put in Linear State Space Form

See D&K Chapter 3 for details on:

- ▶ Including (business) cycle components.
- ▶ Multivariate structural time series, seemingly unrelated time series models, latent risk models.
- ▶ Exponential smoothing.
- ▶ Dynamic factor models.
- ▶ State space in continuous time.
- ▶ Spline smoothing for time series.

Further Notes on Linear State Space Form

- ▶ Estimating parameters, smoothing and forecasting when there are missing observations is straightforward when the above models are written in state space form.
- ▶ The representation of the ARMA model as a linear state space model used above is only one such. There are many other ways to do this.
- ▶ The above examples show that using the linear state space form is a flexible approach to building up models with a variety of features.

Example: Local linear model with intervention term

To allow for the change in flow associated with construction of the Aswan dam in 1899, we introduce the intervention variable x_t which is zero until 1898 and unity thereafter.

$$y_t = Z_t \alpha_t + \epsilon_t, \quad Z_t = (1, \textcolor{red}{x_t}), \quad \alpha_t = (\mu_t, \lambda_t)'$$

$$\mu_t = \mu_{t-1} + \xi_t, \quad \lambda_t = \lambda_{t-1} + \zeta_t$$

$$(\xi_t, \zeta_t)' \sim N(0, \text{diag}(\sigma_\xi^2, \sigma_\zeta^2)).$$

If $\sigma_\zeta^2 = 0$ the regression coefficient λ_t is constant which we assume.

Local Level Model	Intervention Model
$\hat{\sigma}_\epsilon^2 = 15098.7$	$\hat{\sigma}_\epsilon^2 = 16925.6$
$\hat{\sigma}_\eta^2 = 1469.16$	$\hat{\sigma}_\xi^2 = 0.2131$
	$\hat{\lambda} = -244.33$

Example: Intervention Model for Nile River Flow at Aswan Dam

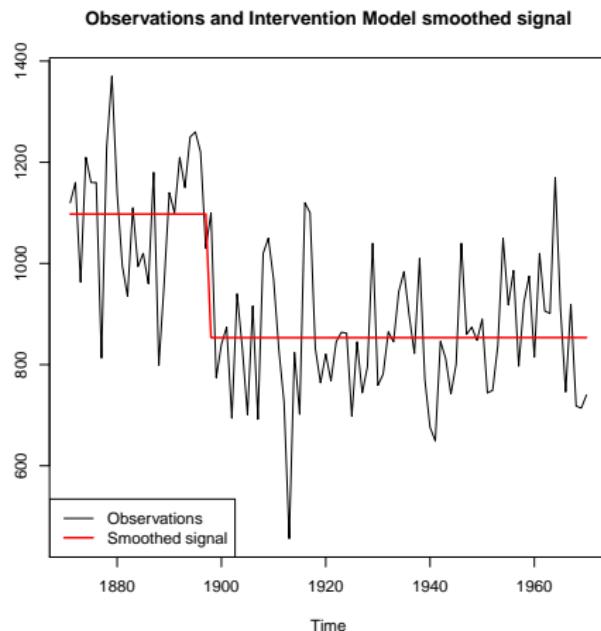


Figure : Observed flow and smoothed signal estimate from intervention model.

Example: Comparison of Intervention and Local Linear Models

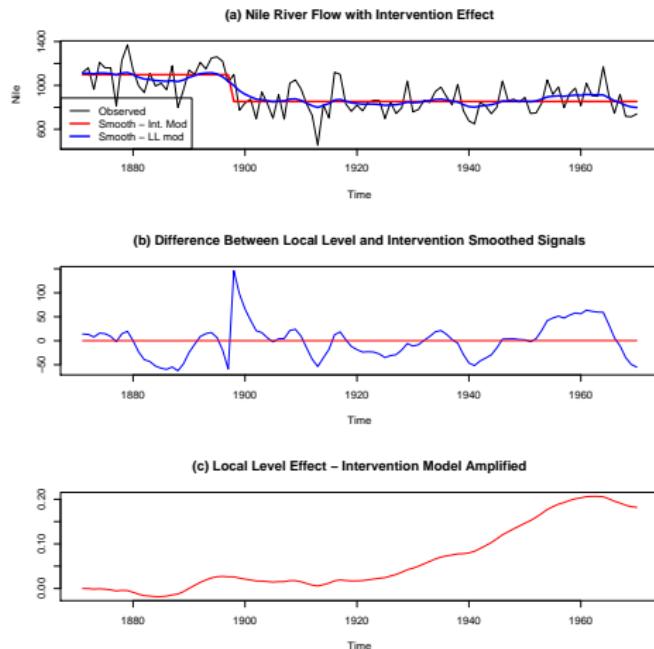


Figure : (a) Observed flow and smoothed signal estimate from intervention model and local linear model; (b) Difference between signal estimates from intervention model and local linear model; (c) Amplified local level component $\hat{\mu}_t$ of state vector in intervention model

Topic: 4: Filtering Smoothing and Forecasting

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Review: Regression lemma, MVLUE, Bayesian

Linear SS Model: Recap General Form

Filtering

Kalman Filter Recursions D&K§4.3.2

State Estimation Errors, Forecast Errors D&K§4.3.5

State Smoothing

Disturbance Smoothing

Simulation Smoothing

Matrix Formulations

Lemma 1: Conditional Normal Distribution

Lemma

Let x and y be jointly normally distributed random vectors with

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma'_{xy} & \Sigma_{yy} \end{bmatrix}\right)$$

then the conditional distribution of x given y is

$$x|y \sim N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy})$$

Proof: D&K§4.2 – Blackboard

Lemma 2: MVLUE

Linear Estimates $\bar{x} = \beta + \gamma y$ where β and γ are conformable vectors and matrices.

Unbiased $E(x - \bar{x}) = 0$.

Minimum Variance A linear unbiased x^* such that $\text{Cov}(x - \bar{x}) \geq \text{Cov}(x - x^*)$ for all linear unbiased estimators \bar{x} .

Lemma

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

is MVLUE of x given y with error variance matrix

$$\text{Var}(x - \hat{x}) = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma'_{xy}$$

Proof: D&K §4.2 – Blackboard.

Linear SS Model: General Form D&K§4.3

$$\begin{aligned}y_t &= Z_t \alpha_t + \epsilon_t, & \epsilon_t &\stackrel{\text{indep}}{\sim} N(0, H_t) \\ \alpha_{t+1} &= T_t \alpha_t + R_t \eta_t, & \eta_t &\stackrel{\text{indep}}{\sim} N(0, Q_T) \\ a_1 &\sim N(a_1, P_1)\end{aligned}\tag{4.1}$$

Vector		Matrix	
y_t	$p \times 1$	Z_t	$p \times m$
α_t	$m \times 1$	T_t	$m \times m$
ϵ_t	$p \times 1$	H_t	$p \times p$
η_t	$r \times 1$	R_t	$m \times r$
		Q_t	$r \times r$
a_1	$m \times 1$	P_1	$m \times m$

Conditional Distributions for Kalman Filtering D&K§4.3

Recall from before that since all distributions are normal any conditional and marginal distributions are normal. Hence

Filtering: $\alpha_t | Y_t \sim N(a_{t|t}, P_{t|t})$.

1-step Prediction: $\alpha_{t+1} | Y_t \sim N(a_{t+1}, P_{t+1})$

Also, as before, define the one step ahead forecast error

$$v_t = y_t - E(y_t | Y_{t-1}) = y_t - E(Z_t \alpha_t + \epsilon_t | Y_{t-1}) = y_t - Z_t a_t.$$

The Kalman Filter is a recursive method for updating the conditional means and variances for filtering and prediction.

Kalman Filter Recursions

For $t = 1, \dots, n$ do:

$$\begin{aligned} v_t &= y_t - Z_t a_t, & F_t &= Z_t P_t Z_t' + H_t, \\ a_{t|t} &= a_t + P_t Z_t' F_t^{-1} v_t, & P_{t|t} &= P_t - P_t Z_t' F_t^{-1} Z_t P_t, \\ a_{t+1} &= T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t (T_t - K_t Z_t)' + R_t Q_t R_t', \end{aligned} \tag{4.2}$$

where $K_t = T_t P_t Z_t' F_t^{-1}$ is the Kalman gain matrix and the initial state distribution is assumed to be $\alpha_1 \sim N(a_1, P_1)$.

After $a_{t|t}$ and $P_{t|t}$ are computed, prediction of α_{t+1} requires only

$$a_{t+1} = T_t a_{t|t}, \quad P_{t+1} = T_t P_{t|t} T_t' + R_t Q_t R_t'.$$

Vector		Matrix	
v_t	$p \times 1$	F_t	$p \times p$
a_t	$m \times 1$	K_t	$m \times p$
$a_{t t}$	$m \times 1$	P_t	$m \times m$
		$P_{t t}$	$m \times m$

State Estimation Errors, Forecast Errors

State estimation errors:

$$x_t = \alpha_t - a_t, \quad \text{Var}(x_t) = P_t$$

Forecast errors or innovations:

$$v_t = y_t - E(y_t | Y_{t-1}) = y_t - Z_t a_t = Z_t x_t + \epsilon_t.$$

where

$$x_{t+1} = L_t x_t + R_t \eta_t - K_t \epsilon_t, \quad K_t = T_t P_t Z_t' F_t^{-1}, \quad L_t = T_t - K_t Z_t.$$

Exercise. Prove the above.

Innovations Analogue of SS Model

$$v_t = Z_t x_t + \epsilon_t, \quad x_{t+1} = L_t x_t + R_t \eta_t - K_t \epsilon_t, \quad x_1 = \alpha_1 - a_1.$$

Alternative derivation of $P_{t+1} = \text{Var}(x_{t+1})$:

$$\begin{aligned} P_{t+1} &= \text{Var}(x_{t+1}) = E[(\alpha_{t+1} - a_{t+1}) x'_{t+1}] \\ &= E(\alpha_{t+1} x'_{t+1}) \\ &= E[(T_t \alpha_t + R_t \eta_t)(L_t x_t + R_t \eta_t - K_t \epsilon_t)'] \\ &= T_t P_t L'_t + R_t Q_t R'_t. \end{aligned}$$

Independence of Innovations

The one-step ahead forecasts errors or innovations v_t are independent, shown as follows:

Recall

$$p(y_1, \dots, y_n) = p(y_1) \prod_{t=2}^n p(y_t | Y_{t-1}) = \prod_{t=1}^n p(v_t)$$

where $p(y_1) = p(v_1)$. Since $E(y_t | Y_{t-1})$ is a linear function of y_t, \dots, y_{t-1} , the Jacobian of the transformation from $Y_n = (y_1, \dots, y_n)'$ to $v = (v_1, \dots, v_n)'$ is unity so that the left hand side is also $p(v_1, \dots, v_n)$ and hence

$$p(v_1, \dots, v_n) = \prod_{t=1}^n p(v_t)$$

proving independence.

Notation for State Smoothing D&K§4.4

Objective: For each $t = 1, \dots, n$ find the conditional density of $\alpha_t | Y_n$, that is the conditional density of the state given the complete set of observations.

Assume: normality and that $\alpha_1 \sim N(a_1, P_1)$ where a_1 and P_1 are known.

State smoothing: Calculation of $\hat{\alpha}_t = E(\alpha_t | Y_n)$

Other forms of smoothing:

- ▶ Fixed interval smoothing: $E(\alpha_t | y_t, \dots, y_s)$.
- ▶ Fixed point smoothing: $\hat{\alpha}_{t|n} = E(\alpha_t | Y_n)$ for t fixed.
- ▶ Fixed lag smoothing: $\hat{\alpha}_{n-j|n} = E(\alpha_{n-j} | Y_n)$ for a fixed $j > 0$ and $n = j+1, j+2, \dots$

We will not pursue these other forms of smoothing here and will concentrate on state smoothing.

State Smoothing Recursions D&K§4.4.4

Let $\hat{\alpha}_t = E(\alpha_t | Y_n)$ and $V_t = \text{Var}(\alpha_t | Y_n)$. Initialise at $r_n = 0$ and $N_n = 0$ and for $t = n, \dots, 1$ (backward recursion) do:

$$\begin{aligned} r_{t-1} &= Z_t' F_t^{-1} v_t + L_t' r_t, & N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \\ \hat{\alpha}_t &= a_t + P_t r_{t-1}, & V_t &= P_t - P_t N_{t-1} P_t. \end{aligned} \tag{4.3}$$

Vector		Matrix	
r_t	$m \times 1$	N_t	$m \times m$
$\hat{\alpha}_t$	$m \times 1$	V_t	$m \times m$
u_t	$p \times 1$	D_t	$p \times p$
$\hat{\epsilon}_t$	$p \times 1$		
$\hat{\eta}_t$	$r \times 1$		

Blackboard: Derive (4.3)

Smoothed Disturbances D&K§4.5

The smoothed estimate of ϵ_t is $\hat{\epsilon}_t = E(\epsilon_t | Y_n)$.

Recall that information in Y_n is equivalent to that in Y_{t-1} and v_t, \dots, v_n and ϵ_t and v_t are jointly independent of Y_{t-1} so that

$$\hat{\epsilon}_t = E(\epsilon_t | Y_{t-1}, v_t, \dots, v_n) = \sum_{j=t}^n E(\epsilon_t v_j') F_j^{-1} v_j, \quad t = 1, \dots, n.$$

using the multivariate normal regression lemma (Lemma 4.1).

The smoothed estimate of η_t is $\hat{\eta}_t = E(\eta_t | Y_n)$.

Similarly we have

$$\hat{\eta}_t = E(\eta_t | Y_{t-1}, v_t, \dots, v_n) = \sum_{j=t}^n E(\eta_t v_j') F_j^{-1} v_j, \quad t = 1, \dots, n.$$

D&Kpp.93-94 derive appropriate recursive expressions for $\hat{\epsilon}_t$ and $\hat{\eta}_t$ which we summarize shortly.

Variance Matrices for Smoothed Disturbances

Using Lemma 4.1 we have

$$\begin{aligned}\text{Var}(\epsilon_t | Y_n) &= \text{Var}(\epsilon_t | Y_{t-1}, v_t, \dots, v_n) \\ &= H_t - \sum_{j=t}^n \text{Cov}(\epsilon_t, v_j) F_j^{-1} \text{Cov}(\epsilon_t, v_j)' \\ &= H_t - H_t D_t H_t, \quad D_t = F_t^{-1} + K_t' N_t K_t,\end{aligned}$$

and

$$\begin{aligned}\text{Var}(\eta_t | Y_n) &= \text{Var}(\eta_t | Y_{t-1}, v_t, \dots, v_n) \\ &= \text{Var}(\eta_t) - \sum_{j=t}^n \text{Cov}(\eta_t, v_j) F_j^{-1} \text{Cov}(\eta_t, v_j)' \\ &= Q_t - Q_t R_t' N_t R_t Q_t, \quad N_t = Z_t' F_t^{-1} Z_t + L_t' N_t L_t,\end{aligned}$$

Summary: Smoothed Disturbance Backward Recursions D&K§4.5.3

Start with $r_n = 0$ and $N_n = 0$ and for $t = n, \dots, 1$ do:

$$\begin{aligned}\hat{\epsilon}_t &= H_t(F_t^{-1}v_t - K'_t r_t), \text{Var}(\epsilon_t | Y_n) = H_t - H_t(F_t^{-1} + K'_t N_t K_t)H_t, \\ \hat{\eta}_t &= Q_t R'_t r_t, \quad \text{Var}(\eta_t | Y_n) = Q_t - Q_t R'_t N_t R_t Q_t, \\ r_{t-1} &= Z'_t F_t^{-1} v_t + L'_t r_t, \quad N_{t-1} = Z'_t F_t^{-1} Z_t + L'_t N_t L_t.\end{aligned}\tag{4.4}$$

Alternative Computational Form for $\hat{\epsilon}_t$, $\text{Var}(\epsilon_t | Y_n)$

The following forms rely directly on the system matrices Z_t and T_t :

$$\begin{aligned}\hat{\epsilon}_t &= H_t u_t, & \text{Var}(\epsilon_t | Y_n) &= H_t - H_t D_t H_t, \\ u_t &= F_t^{-1} v_t - K_t' r_t, & D_t &= F_t^{-1} + K_t' N_t K_t, \\ r_{t-1} &= Z_t' u_t + T_t' r_t, & N_{t-1} &= Z_t' D_t Z_t + T_t' N_t T_t \\ &&&\quad - Z_t' K_t' N_t T_t - T_t' N_t K_t Z_t.\end{aligned}\tag{4.5}$$

They are computationally more efficient because Z_t and T_t usually contain many 0 or 1 elements.

Computations for State and Disturbance Smoothing

- ▶ State and disturbance smoothing both require carrying out the (forward recursion) Kalman filter (4.2).
- ▶ State smoothing then carries out the backward recursions of (4.3). These involve use of a_t and P_t which are not sparse, hence increasing computational time.
- ▶ Disturbance smoothing carries out the backward recursions of (4.4) which only require v_t , F_t and K_t thereby reducing the storage required.

Fast State Smoothing D&K§4.6.2

For the simulation smoother (to be considered next) applied to obtaining samples of the state given the observations we do not need the variances $V_t = \text{Var}(\alpha_t | Y_n)$ in (4.3). Note that by taking conditional expectations given Y_n in the state equation of (4.1) we get

$$\hat{\alpha}_{t+1} = T_t \hat{\alpha}_t + R_t \hat{\eta}_t = T_t \hat{\alpha}_t + R_t Q_t R_t' r_t. \quad (4.6)$$

initialized by $\hat{\alpha}_1 = a_1 + P_1 r_0$ with r_0 obtained by the backward recursions in (4.3).

Note: this method does not require storage of a_t and P_t nor multiplication of the full matrices P_t . Also, there are storage savings. Finally, T_t and $R_t Q_t R_t'$ are sparse and contain many zeros and ones so that use of (4.6) is fast.

Simulation Smoothing D&K§4.9

Simulation Smoothing is concerned with drawing samples of state or disturbance vectors **conditional on the observations being held fixed.**

It will provide the basis for simulation needed to approximate likelihoods or perform Bayesian analysis in the non-Gaussian, nonlinear models studied in this course.

The method described here is that proposed by Durbin and Koopman (2002) and is based on mean correcting unconditional samples of states or disturbances. As a result it is simple and computationally efficient to implement.

Other methods are available.

Simulation Smoothing of Disturbances: Notation

Let $w = (\epsilon'_1, \eta'_1, \dots, \epsilon'_n, \eta'_n)'$ and denote

$$\hat{w} = E(w|Y_n), \quad W = \text{Var}(w|Y_n)$$

For the linear Gaussian model (4.1) it follows that
 $w|Y \sim N(\hat{w}, W)$.

The conditional mean vector \hat{w} is easily calculated using the recursions for $\hat{\epsilon}_t$ and $\hat{\eta}_t$ in (4.4) and (4.5) and for the mean correction method W is not required.

The unconditional distribution of w is

$$p(w) = N(0, \Phi), \quad \text{where } \Phi = \text{diag}(H_1, Q_1, \dots, H_n, Q_n).$$

Taking draws of w^+ is often simple because H_t and Q_t are scalar or diagonal particularly in cases we will consider later.

Simulation Smoothing of Disturbances: Method

To take draws of $w^+ \sim p(w)$:

1. Draw $\alpha_1^+ \sim p(\alpha_1)$ and y_t^+ recursively from the linear Gaussian model (4.1) using the w_+ . Call the resulting stacked vector y^+ .
2. Compute $\hat{w}^+ = E(w|y^+)$ using (4.4) and (4.5).
3. Put

$$\tilde{w} = w^+ - \hat{w}^+ + \hat{w} \quad (4.7)$$

Then $\tilde{w} \sim N(0, W)$.

Proof of this result relies on Lemma 4.1 which shows that the conditional variance (here W) is *not* a function of the conditioning vector. Details D&Kp.108 – see blackboard.

Simulation Smoothing of Disturbances: Notes

- ▶ Both the ease of taking draws of w^+ from $N(0, W)$ and the simplicity of the adjustment using (4.7) makes the simulation smoother computationally efficient in comparison with earlier methods.
- ▶ The idea had been suggested at least as early as 1974 by the French geostatistician Journel in the context of sampling from Gaussian field for mineral exploration.

Simulation Smoothing: States D&K§4.9.2

1. Initialise $\alpha_1^+ \sim p(\alpha_1)$, generate w^+ as above and then use the resulting ϵ_t^+, η_t^+ in (4.1) to generate a state vector α^+ and observation vector y^+ .
2. Using the Kalman filter and the smoother recursions (4.5) along with the forward (fast) recursion (4.6) to compute $\hat{\alpha} = E(\alpha|Y_n)$ and $\hat{\alpha}^+ = E(\alpha|Y^+)$.
3. Put $\tilde{\alpha} = \alpha^+ - \hat{\alpha}^+ + \hat{\alpha}$.

Then $\tilde{\alpha} \sim p(\alpha|Y_n)$.

Matrix Form: State Space Model – 1 D&K§4.13.1

Observation equation:

$$Y_n = Z\alpha + \epsilon, \quad \epsilon \sim N(0, H) \quad (4.8)$$

where

$$Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad Z = \begin{bmatrix} Z_1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & Z_n & 0 \end{bmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \alpha_{n+1} \end{pmatrix},$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad H = \begin{bmatrix} H_1 & & 0 \\ & \ddots & \\ 0 & & H_n \end{bmatrix}. \quad (4.9)$$

Matrix Form: State Space Model – 2

State equation:

$$\alpha = T(\alpha_1^* + R\eta), \quad \eta \sim N(0, Q) \quad (4.10)$$

where

$$T = \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 \\ T_1 & I & 0 & 0 & 0 & 0 \\ T_{2 \cdot 1} & T_2 & I & 0 & 0 & 0 \\ T_{3 \cdot 1} & T_{3 \cdot 2} & T_3 & \ddots & & \vdots \\ & & & & I & 0 \\ T_{n \cdot 1} & T_{n \cdot 2} & T_{n \cdot 3} & \cdots & T_n & I \end{bmatrix}, \quad (4.11)$$

where

$$T_{n \cdot j} = T_n T_{n-1} \cdots T_j$$

and (next slide ...)

Matrix Form: State Space Model – 3

$$\alpha_1^* = \begin{pmatrix} \alpha_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad R = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ R_1 & 0 & & \\ 0 & R_2 & & 0_n \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & R_n \end{bmatrix}, \quad (4.12)$$

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \quad Q = \begin{bmatrix} Q_1 & & 0 \\ & \ddots & \\ 0 & & Q_n \end{bmatrix}. \quad (4.13)$$

Let a_1^* be the column vector with a_1 in the first block and zeros in the remaining n blocks and let P^* be the matrix with $(n+1) \times (n+1)$ blocks of zero matrices except for the $(1, 1)$ th block which contains P_1 . Put

$$Q^* = P_1^* + RQR'.$$

Matrix Form: Expectations and Variances for State Space Model

Using the above notation we can write $E(\alpha_1^*) = a_1^*$, $\text{Var}(\alpha_1^*) = P_1^*$, $E(\alpha) = Ta_1^*$ and

$$\text{Var}(\alpha) = \text{Var}(T[(\alpha_1^* - a_1^*) + R\eta]) \quad (4.14)$$

$$= T(P_1^* + RQR')T' \quad (4.15)$$

$$= TQ^*T'. \quad (4.16)$$

Also, since

$$Y_n = ZT\alpha_1^* + ZTR\eta + \epsilon \quad (4.17)$$

we get $E(Y_n) = \mu = ZTa_1^*$ and

$$\Omega = \text{Var}(Y_n) \quad (4.18)$$

$$= ZT(P_1^* + RQR')T'Z' + H \quad (4.19)$$

$$= ZTQ^*T'Z' + H. \quad (4.20)$$

Densities in terms of Matrix Expressions D&K§4.13.2

Using (4.18) we get

$$\log p(Y_n) = \text{const.} - \frac{1}{2} \log |\Omega| - \frac{1}{2} ((Y_n - \mu)' \Omega^{-1} (Y_n - \mu)). \quad (4.21)$$

From (4.14)

$$\log p(\alpha) = \text{const.} - \frac{1}{2} \log |V^*| - \frac{1}{2} (\alpha - a^*)' V^{*-1} (\alpha - a^*) \quad (4.22)$$

where $a^* = E(\alpha) = Ta_1^*$ and $V^* = \text{Var}(\alpha) = TQ^*T'$.

Using the observation equation (4.8) gives

$$\log p(Y_n|\alpha) = \text{const.} - \frac{1}{2} \log |H| - \frac{1}{2} (Y_n - \theta)' H^{-1} (Y_n - \theta) \quad (4.23)$$

where $\theta = Z\alpha$ is referred to as the signal. Hence

$$p(Y_n|\alpha) = p(y_n|\theta).$$

Matrix Cholesky Form of Filtering D&K§4.13.3

D&K§4.13.3 provide details of the form of a non-singular lower triangular matrix C with $n \times n$ blocks for which

$$v = C(Y_n - \mu), \quad F = \text{Var}(v) = C\Omega C' \quad (4.24)$$

and $F = \text{diag}(F_1, \dots, F_n)$ is block diagonal. Hence $|\Omega^{-1}| = |F|^{-1}$, so we can rewrite the joint density (4.21) (which defines the likelihood) as

$$\log p(Y_n) = \text{const.} - \frac{1}{2} \log |F| - \frac{1}{2} v' F^{-1} v. \quad (4.25)$$

in which v and K are efficiently calculated using the Kalman filter.

Smoothing Using Matrix Form D&K§4.13.4

Using Lemma 4.1 we have

$$\hat{\epsilon} = E(\epsilon|Y_n) = \text{Cov}(\epsilon, Y_n)\Omega^{-1}(Y_n - \mu)$$

But $\text{Cov}(\epsilon, Y_n) = H$ so that using (4.24) gives

$$\hat{\epsilon} = Hu, \quad u = \Omega^{-1}(Y_n - \mu) \dots = F^{-1}v - K'r$$

– for details see D&Kp.118. and the definitions of u and r are consistent with those in (4.5), (4.3) respectively.

Similarly the smoothed state disturbance vector $\hat{\eta} = E(\eta|Y_n)$ and the smoothed estimator of α are given by completely analogous matrix expressions to their recursive counterparts:

$$\hat{\eta} = QR'r, \quad \hat{\alpha} = Ta_1^* + TQ^*r$$

– for details see DK pp.118-9.

Matrix Expressions for Signal D&K§4.13.5

Recall the signal $\theta = Z\alpha$. Obviously this has mean

$$E(\theta) = \mu = E(Z\alpha) = Za^* = ZTa_1^*$$

and variance denoted by

$$\Psi = \text{Var}(\theta) = ZV^*Z' = ZTQ^*T'Z$$

Note that in this notation, $\text{Cov}(\theta, Y_n) = \text{Var}(\theta) = \Psi$ and $\text{Var}(Y_n) = \Omega = \Psi + H$. Applying Lemma 4.1 again gives

$$\hat{\theta} = E(\theta|Y_n) = \mu + \Psi\Omega^{-1}(Y_n - \mu), \quad \text{Var}(\theta|Y_n) = \Psi - \Psi\Omega^{-1}\Psi. \quad (4.26)$$

The conditional mean can alternatively be rewritten as

$$\hat{\theta} = (\Psi^{-1} + H^{-1})^{-1}(\Psi^{-1}\mu + H^{-1}Y_n)$$

Matrix form of simulation smoothing D&K§4.13.6

Simulation smoothing of signal In line with previous results we can express samples $\tilde{\theta} \sim p(\theta|Y_n)$ as

$$\tilde{\theta} = \theta^+ - \hat{\theta}^+ + \hat{\theta}.$$

Note that

$$\begin{aligned}\tilde{\theta} - \theta^+ &= \hat{\theta} - \hat{\theta}^+ \\ &= [\mu + \Psi\Omega^{-1}(Y_n - \mu)] - [\mu + \Psi\Omega^{-1}(y^+ - \mu)] \\ &= \Psi\Omega^{-1}(Y_n - y^+)\end{aligned}$$

Hence

$$\tilde{\theta} = \theta^+ + \Psi\Omega^{-1}(Y_n - y^+).$$

This leads to the steps:

- ▶ Simulate α^+ , $\theta^+ = Z\alpha^+$ and $y^+ = \theta^+ + \epsilon^+$ from the linear Gaussian state space model (4.1).
- ▶ Apply the Kalman filter and smoother to (4.1) with $a_1 = 0$ and 'observations' $y_t - y_t^+$.

Topic: 5: Approximate Filtering and Smoothing for Nonlinear non-Gaussian State Space Models

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Approaches to Approximate Filtering and Smoothing

Approximation Using Mode Estimation

Mode estimation for linear Gaussian signal and conditionally independent observations

Mode estimation for exponential family response distributions

Mode estimation for the stochastic volatility model

Optimality and Multimodality

Mode estimation by linearization

Mode estimation by matching only first derivatives

Various Approaches to Approximate Filtering and Smoothing

For various nonlinear non-Gaussian models several approaches have been suggested:

- ▶ Extended Kalman Filter
- ▶ Unscented Kalman Filter
- ▶ Approximation by data transformation
- ▶ Modal approximation

We concentrate on approximation by mode estimation, but briefly summarize the other three methods.

Extended Kalman Filter

This is applied to models of the form

$$y_t = Z_t(\alpha_t) + \epsilon_t, \quad \alpha_{t+1} = T_t(\alpha_t) + R_t(\alpha_t)\eta_t \quad (5.1)$$

where $Z_t(\alpha_t)$, $T_t(\alpha_t)$ and $R_t(\alpha_t)$ are differentiable functions of α_t and ϵ_t and η_t are serially and mutually uncorrelated random sequences with mean zero and covariance matrices $H_t(\alpha_t)$ and $Q_t(\alpha_t)$ respectively.

The equations (5.1) are linearised by Taylor series expansion around $a_t = E(\alpha_t | Y_{t-1})$ or $a_{t|t} = E(\alpha_t | Y_t)$ as appropriate leading to a linear state space form for which Kalman filtering can be applied.

Details in D&K§10.2.

Examples include multiplicative trend-cycle decomposition and power growth models.

Unscented Kalman Filter

Based on the unscented transformation which finds an approximation to the density of $y = f(x)$ where $x \sim N(\bar{x}, P_{xx})$ based on a discrete set of *sigma points* x_0, \dots, x_{2m+1} and associated *sigma weights* w_0, \dots, w_{2m+1} summing to unity selected so that $\bar{x} = \sum w_i x_i$ and $P_{xx} = \sum (w_i (x_i - \bar{x})(x_i - \bar{x})')$.

This gives an approximation to the continuous density of $y = f(x)$ by a discrete density $y_i = f(x_i)$ with mean vector and covariance matrix matching those of $f(x)$.

– see D&K §10.3 for details.

Approximation using data transformations – 1

Transform into a form for which the linear Gaussian model is used as an approximation.

Example 1 Multiplicative model for trend μ_t and cycle or seasonal γ_t with additive errors:

$$y_t = \mu_t \times \gamma_t + \epsilon_t$$

This is approximated by taking logs

$$\log(y_t) \approx \log(\mu_t) + \log(\gamma_t) + u_t.$$

Approximation using data transformations – 2

Example 2 Stochastic volatility model (1.22):

$$y_t = \mu + \sigma \exp\left(\frac{1}{2}\theta_t\right)\epsilon_t, \quad \epsilon_t \stackrel{\text{indep}}{\sim} N(0, 1)$$

where θ_t is a Gaussian autoregressive process.

Transform using logs of squares, $\log(y_t - \mu)^2 = \kappa + \theta_t + \xi_t$, where $\kappa = \log \sigma^2 + E(\log \epsilon_t^2)$ and $\xi_t = \log \epsilon_t^2 - E(\log \epsilon_t^2)$. Although the noise terms ξ_t is not normally distributed the linearity of $\log(y_t - \mu)^2$ allows use of the Kalman filter and smoothing methods. Estimation is based on assuming a normal distribution for the ξ_t leading to ‘quasi-maximum likelihood’.

Mode estimation - Motivation

The EKF and UKF were proposed to deal with non-linearities in $Z_t(\alpha_t)$ in the model

$$y_t = Z_t(\alpha_t) + \epsilon_t \quad (5.2)$$

where

$$\alpha_{t+1} = T_t(\alpha_t) + R_t(\alpha_t)\eta_t \quad (5.3)$$

where ϵ_t and η_t are mean zero, uncorrelated with covariances $H_t(\alpha_t)$ and $Q_t(\alpha_t)$ respectively. Note that the observation errors, while not necessarily Gaussian, are additive in (5.2).

For non-Gaussian observation densities additivity of error structures is not always appropriate so we consider instead of (5.2)

$$y_t \sim p(y_t | \alpha_t) \quad (5.4)$$

When $p(y_t | \alpha_t)$ has mean $Z_t(\alpha_t)$ and covariance matrix $H_t(\alpha_t)$ then the model consisting of (5.4) and (5.3) can be approximated using the EKF or the UKF. But, these do not account adequately for the non-Gaussian nature of $p(y_t | \alpha_t)$.

Mode Estimation: linear Gaussian state with non-Gaussian observation density

We initially focus on the case where:

- ▶ $p(y_t|\alpha_t)$ is any density.
- ▶ α_t evolves linearly over time with Gaussian errors η_t
- ▶ θ_t is a linear function of the state vector and it is sufficient for specifying the conditional distribution of $y_t|\alpha_t$; that is $p(y_t|\alpha_t) = p(y_t|\theta_t)$.

This leads to the model

$$y_t \sim p(y_t|\theta_t), \quad \theta_t = Z_t \alpha_t, \quad \alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \sim N(0, Q_t) \quad (5.5)$$

Computation of the mode of $\theta = (\theta'_1, \dots, \theta'_n)'$ conditional on all observations Y_n leads to a linear Gaussian state space model.

Mode Estimation: linear Gaussian state space model

This is the model with a Gaussian state, Gaussian observation density. In vector form

$$Y_n = \theta + \epsilon, \quad \theta = Z\alpha, \quad \epsilon \sim N(0, H) \quad (5.6)$$

where (see (4.10)) $\alpha = T(\alpha_1^* + R\eta)$ and H is block-diagonal (see (4.11), (4.12), (4.13) etc.) and we define

$$\Psi = ZT(P_1^* + RQR')T'Z'$$

Recall (slide 167) that the conditional distribution of the signal given the observations is

$$\theta | Y_n \sim N(\hat{\theta}, \text{Var}(\theta | Y_n)) \quad (5.7)$$

where

$$\hat{\theta} = E(\theta | Y_n) = (\Psi^{-1} + H^{-1})^{-1}(\Psi^{-1}\mu + H^{-1}Y_n) \quad (5.8)$$

and $\mu = E(\theta) = Za_1^*$.

Mode Estimation: linear Gaussian state space model

Since the mode and mean vectors coincide for the multivariate normal distribution it follows that $\hat{\theta}$ given in (5.8) is also the mode of the smoothed density $p(\theta|Y_n)$. Recall that the mode of a density is the point (vector) at which the density is maximised.

Mode Estimation: linear Gaussian signal, general conditionally independent observation density – 1

The model is summarized as follows:

$$Y_n \sim p(Y_n|\theta) = \prod_{t=1}^n p(y_t|\theta_t), \quad \theta \sim N(\mu, \Psi) \quad (5.9)$$

The mode of the smoothed density $p(\theta|Y_n)$ cannot be obtained in closed form (i.e. as an explicit expression). However we can find it quite efficiently using numerical methods.

Note that

$$\begin{aligned} \log p(\theta|Y_n) &= \log p(Y_n, \theta) - \log p(Y_n) \\ &= \log p(Y_n|\theta) + \log p(\theta) - \log p(Y_n) \end{aligned} \quad (5.10)$$

$p(Y_n)$ does NOT depend on θ . Why?

Mode Estimation: linear Gaussian signal, general conditionally independent observation density – 2

We can maximise $\log p(\theta|Y_n)$ with respect to θ to find the mode of $p(\theta|Y_n)$ numerically. For many of our models this can be done very efficiently and reliably using the Newton-Raphson method.

Some important useful notation: Let $p(\cdot|\cdot)$ represent $p(\theta|Y_n)$ or $p(Y_n|\theta)$ depending on the context. Denote

$$\dot{p}(\cdot|\cdot) = \frac{\partial \log p(\cdot|\cdot)}{\partial \theta}, \quad \ddot{p}(\cdot|\cdot) = \frac{\partial^2 \log p(\cdot|\cdot)}{\partial \theta \partial \theta'},$$

NOTE: \dot{p} and \ddot{p} refer to differentiation of the **log density**

We seek the modal value $\hat{\theta}$ at which $\log p(\theta|Y_n)$ is maximised which, assuming some regularity of the densities involved, is the point at which $\dot{p}(\theta|Y_n)|_{\theta=\hat{\theta}} = 0$.

Newton Raphson method for mode estimation

Let $\tilde{\theta}$ be an initial guess at the mode. Using a Taylor series expansion of $\dot{p}(\theta|Y_n)$ around $\tilde{\theta}$ we get

$$\dot{p}(\theta|Y_n) \approx \dot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}} + \ddot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta}).$$

Assume there is a value $\tilde{\theta}^+$ which makes the left hand side zero and rearranging we get the **Newton Raphson update**

$$\tilde{\theta}^+ = \tilde{\theta} - [\ddot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}}]^{-1} \dot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}} \quad (5.11)$$

Of course, unless $\dot{p}(\theta|Y_n)$ is linear in θ , this single update will not produce the mode, but it should be closer than the previous guess.

The updates in (5.11) are repeated (putting $\tilde{\theta} = \tilde{\theta}^+$) until convergence, as measured by relative change in $\tilde{\theta}^+ - \tilde{\theta}$ or in how close $\dot{p}(\theta|Y_n)|_{\theta=\tilde{\theta}}$ is to the zero vector. **The converged value is assumed to be the mode $\hat{\theta}$**

Mode Estimation: linear Gaussian signal, general conditionally independent observation density – 3

We now apply this to finding the mode of $p(\theta|Y_n)$. We differentiate (5.10) with respect to θ to get

$$\dot{p}(\theta|Y_n) = \dot{p}(Y_n|\theta) + \dot{p}(\theta), \quad \ddot{p}(\theta|Y_n) = \ddot{p}(Y_n|\theta) + \ddot{p}(\theta) \quad (5.12)$$

where for the Gaussian signal density we have

$$\dot{p}(\theta) = \frac{\partial \log p(\theta)}{\partial \theta} = -\Psi^{-1}(\theta - \mu), \quad \ddot{p}(\theta) = -\Psi^{-1}. \quad (5.13)$$

and for the conditionally independent observation density (for which $\log p(Y_n|\theta) = \sum_{t=1}^n \log p(y_t|\theta_t)$)

$$\dot{p}(Y_n|\theta) = [\dot{p}(y_1|\theta_1), \dots, \dot{p}(y_n|\theta_n)]', \quad (5.14)$$

$$\ddot{p}(Y_n|\theta) = \text{diag}[\ddot{p}(y_1|\theta_1), \dots, \ddot{p}(y_n|\theta_n)]. \quad (5.15)$$

Mode Estimation: linear Gaussian signal, general conditionally independent observation density – 4

Let

$$A = -[\ddot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}}]^{-1}, \quad \tilde{Y}_n = \tilde{\theta} + A\dot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} \quad (5.16)$$

then the Newton-Raphson update (5.11) applied to the current model is **EXERCISE**:

$$\begin{aligned}\tilde{\theta}^+ &= \tilde{\theta} - \{\ddot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} - \Psi^{-1}\}^{-1}\{\dot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}} - \Psi^{-1}(\tilde{\theta} - \mu)\} \\ &= (\Psi^{-1} + A^{-1})^{-1}(A^{-1}\tilde{Y}_n + \Psi^{-1}\mu).\end{aligned} \quad (5.17)$$

Note

- ▶ (5.17) is similar to (5.8) which is the mode of the smoothed signal density for the linear Gaussian state space model.
- ▶ Therefore, if we consider a linear Gaussian state space model with observations \tilde{Y}_n (denoted x in D&K) and observation noise covariance H replaced by A we could use the Kalman filter and smoother for the Gaussian linear state space model to perform the Newton-Raphson update steps in (5.17).
- ▶ Details follow.

Mode Estimation: linear Gaussian signal, general conditionally independent observation density – 5

To implement (5.17) we set up a linear Gaussian state space model as follows. Note the state evolution and disturbance variances are as before. Thus T , R , Q , Z and $\mu = ZTa_1^*$ are as in (5.5).

$$\tilde{Y}_n = \theta + \tilde{\epsilon}, \quad \tilde{\epsilon} \sim N(0, A) \quad (5.18)$$

where, as before, $\alpha = T(\alpha_1^* + R\eta)$ and $\Psi = ZT(P_1^* + RQR')T'Z'$. Application of the Kalman filter and smoother can be used to find the mode $\hat{\theta}$ for this approximating model in the form (5.8) which, calling the mode $\tilde{\theta}^+$, gives

$$\tilde{\theta}^+ = E(\theta | \tilde{Y}_n) = (\Psi^{-1} + A^{-1})^{-1}(\Psi^{-1}\mu + A^{-1}\tilde{Y}_n)$$

matching the Newton-Raphson update in (5.17).

Mode Estimation: linear Gaussian signal, general conditionally independent observation density – 6

Notes:

1. Convergence is rapid (around 10 iterations for sufficiently good accuracy).
2. The method (unadorned) is not valid if $\ddot{p}(y|\theta)$ is not negative definite because the resulting $H = A$ would not be positive definite. Thus $p(y|\theta)$ must be logconcave in θ which is satisfied for many examples – see below.
3. If $p(y|\theta)$ is not logconcave the method can be modified using alternative derivations.
4. The above finds the mode of the smoothed signal density. If the mode of the smoothed state is needed the above derivation can be carried out using $p(\alpha|Y_n)$ in place of $p(\theta|Y_n)$ leading to significant increase in computational effort for the Newton-Raphson iterates.
5. But, once we have $\hat{\theta}$ (see slide 167 or D&K§4.13.5) for a linear Gaussian state space model we can readily calculate $\hat{\alpha}$ using

$$\hat{\alpha} = Ta_1^* + TQ^*T'Z'\Psi^{-1}(\hat{\theta} - \mu).$$

Mode estimation for the exponential family response distributions – 1

Recall the exponential family density in (1.6), taking logs

$$\log p(y_t|\theta_t) = y_t' \theta_t - b_t(\theta_t) + c_t(y_t).$$

Hence

$$\dot{p}(y_t|\theta_t) = y_t - \dot{b}_t(\theta_t), \quad \ddot{p}(y_t|\theta_t) = -\ddot{b}_t(\theta_t)$$

where

$$\dot{b}_t(\theta_t) = \frac{\partial b_t(\theta_t)}{\partial \theta_t}, \quad \ddot{b}_t(\theta_t) = \frac{\partial^2 b_t(\theta_t)}{\partial \theta_t \partial \theta_t'}.$$

For short, let $\dot{b}_t = \dot{b}_t(\tilde{\theta}_t)$ and $\ddot{b}_t = \ddot{b}_t(\tilde{\theta}_t)$ where $\tilde{\theta}_t$ is the current value of θ_t in the iterative method for calculating the mode.

Mode estimation for the exponential family response distributions – 2

Substitution of the derivative expressions in (5.16) gives

$$A_t = \ddot{b}_t^{-1}, \quad \tilde{y}_t = \tilde{\theta}_t + \ddot{b}_t^{-1} y_t - \ddot{b}_t^{-1} \dot{b}_t \quad (5.19)$$

for the t th element A which is diagonal matrix and \tilde{y}_t is the t th element of the new observation vector \tilde{Y}_n .

Since $\ddot{b}_t(\theta_t) = \text{Var}(Y_t|\theta_t)$, A is positive definite for the exponential family. Hence the iterative method of computing the mode given above can be used and will converge correctly from *any starting value for θ !*

The exponential family is logconcave.

Mode estimation for the exponential family response distributions – 3

Examples from the Exponential Family

Table 10.2 in D&K contain the essential quantities required to construct A_t and \tilde{y}_t needed for the iterative search for the mode of $p(\theta_t | Y_n)$. These are $\ddot{b}_t(\theta_t)$ and $\ddot{b}_t(\theta_t)^{-1} \ddot{b}_t(\theta_t)$

Exercise in class: Reproduce the elements of this table.

Exercise in class: Find the corresponding elements for the normal response distribution. Discuss.

Mode estimation for the stochastic volatility model

– Normal Distribution Case

Recall this model from (1.22), (1.5) with conditional log density (1.24) which, letting $z_t = (y_t - \mu)/\sigma$ becomes

$$\log p(y_t|\theta_t) = -\frac{1}{2} [\log 2\pi\sigma^2 + \theta_t + z_t^2 \exp(-\theta_t)].$$

Hence

$$\dot{p}_t(y_t|\theta_t) = -\frac{1}{2} [1 - z_t^2 \exp(-\theta_t)], \quad \ddot{p}_t(y_t|\theta_t) = -\frac{1}{2} z_t^2 \exp(-\theta_t).$$

Substituting these into (5.19) give

$$A_t = \exp(\tilde{\theta}_t)/z_t^2, \quad \tilde{y}_t = \tilde{\theta}_t + 1 - \exp(\tilde{\theta}_t)/z_t^2$$

and using the autoregressive state equation (or more generally any ARMA state representation) to define Ψ the Newton-Raphson iterations in (5.17) can be applied to find the required mode.

Mode estimation for the stochastic volatility model

- *t*-Distribution Case

Here

$$\log p(y_t | \theta_t) = \text{constant} - \frac{1}{2} [\theta_t + (\nu + 1) \log q_t], \quad q_t = 1 + \exp(-\theta_t) \frac{z_t^2}{\nu - 2}$$

hence

$$\dot{p}_t(y_t | \theta_t) = -\frac{1}{2} [1 - (\nu + 1)(q_t^{-1} - 1)], \quad \ddot{p}_t(y_t | \theta_t) = -\frac{1}{2}(\nu + 1)(q_t^{-1} - 1)q_t^{-1}.$$

giving

$$A_t = 2(\nu + 1)^{-1}(\tilde{q}_t - 1)^{-1}\tilde{q}_t^2, \quad \tilde{y}_t = \tilde{\theta}_t + \tilde{q}_t^2 - \frac{1}{2}A_t$$

where \tilde{q}_t is q_t evaluated at $\tilde{\theta}_t$

Note that q_t is positive for all θ_t so that $A = \text{diag}(A_1, \dots, A_n)$ is positive definite and the method based on the Kalman filter smoother can be applied and will converge to the require mode.

Optimality Property of the mode $\hat{\alpha}$ of $p(\alpha|Y_n)$

Two ways in which the mode is used:

- ▶ To obtain, as above, a linear Gaussian approximating model which can be used for simulation in importance sampling for integral evaluation.
- ▶ As an estimate of the state vector (or signal).

Optimality of the mode:

- ▶ Obviously $\hat{\alpha}$, being the mode of $p(\alpha|Y_n)$, is the **most probable value** of the state vector given the observations.
- ▶ It can also be shown to be the **solution of an optimal estimation equation** and hence is analogous to the optimality of maximum likelihood estimation for a fixed dimension parameter. For details see D&K §10.7.5.

Multimodality?

- ▶ In the above it is implicit that there is a single mode for $p(\alpha|Y_n)$.
- ▶ In most examples we consider the existence of several modes does not arise.
- ▶ However, if it does occur, D&K §10.7.5 provide some guidance as to how to proceed.

Mode Estimation by linearization –1

An alternative computation of the mode of the smoothed signal density matches the first and second derivatives of the $p(\theta|Y_n)$ and the smoothed signal density of the linear Gaussian model. Now

$$\log g(\theta|Y_n) = \log g(Y_n|\theta) + \log g(\theta) - \log g(Y_n)$$

where $g(\theta) = p(\theta)$ for a linear Gaussian signal and

$$\log g(Y_n|\theta) = \log g(Y_n|\alpha) = \text{constant} - \frac{1}{2} \log |H| - \frac{1}{2}(Y_n - \theta)' H^{-1} (Y_n - \theta).$$

Using the 'dot' notation for the derivatives of the log densities g with respect to θ we get

$$\dot{g}(Y_n|\theta) = H^{-1}(Y_n - \theta), \quad \ddot{g}(Y_n|\theta) = -H^{-1}.$$

Mode Estimation by linearization –2

Since $g(\theta) = p(\theta)$ we can match the first two derivatives of $p(\theta|Y_n)$ with those of $g(\theta|Y_n)$ by matching them for $p(Y_n|\theta) = g(Y_n|\theta)$, that is

$$H^{-1}(Y_n - \theta) = \dot{p}(Y_n|\theta), \quad -H^{-1} = \ddot{p}(Y_n|\theta). \quad (5.20)$$

However, $\dot{p}(Y_n|\theta)$ and $\ddot{p}(Y_n|\theta)$ are (typically nonlinear) functions of θ so that (5.20) need to be solved iteratively.

As before, for a given value $\tilde{\theta}$ we define an observation disturbance variance A and observation vector \tilde{Y}_n . Let

$$A = -[\ddot{p}(Y_n|\theta)_{\theta=\tilde{\theta}}]^{-1}, \quad \tilde{Y}_n = \tilde{\theta} + A\dot{p}(Y_n|\theta)|_{\theta=\tilde{\theta}}$$

and use the Kalman filter and smoother assuming a linear Gaussian model to get the smoothed state estimate; call this $\tilde{\theta}^+$. Replace $\tilde{\theta}^+ \rightarrow \tilde{\theta}$ and linearize as in (5.20).

At convergence the final linearised model is the linear Gaussian model with the same conditional mode of $\theta|Y_n$ as in the original non-Gaussian nonlinear model.

Linear Gaussian Signal plus noise

In this special case linearisation based on first derivatives alone can be useful. We focus on the case where y_t is a univariate, $y_t = \theta_t + \epsilon_t$ where $\epsilon_t \sim p(\epsilon_t)$ and hence $p(y_t|\theta_t) = p(\epsilon_t)$. Matching $\dot{g}(Y_n|\theta) = \dot{p}(Y_n|\theta)$ gives

$$H^{-1}(Y_n - \theta) = \dot{p}(\epsilon) = \dot{p}(Y_n|\theta)$$

Hence we transform the linear Gaussian signal plus non-Gaussian noise into the linear Gaussian signal plus Gaussian noise by setting $\tilde{Y}_n = Y_n$ (no transformation needed) and $A = \text{diag}(A_1, \dots, A_n)$ where

$$A_t = (y_t - \tilde{\theta}_t) \left[\dot{p}(\epsilon_t) \Big|_{\epsilon_t=y_t-\tilde{\theta}_t} \right]^{-1}$$

Application: linear Gaussian signal, heavy-tailed noise – 1

Let $\epsilon_t \sim p(\epsilon_t)$ where $p(\epsilon_t)$ is given by the t -distribution (1.20).

We first try mode estimation by matching the first two derivatives.

Here

$$\dot{p}(\epsilon_t) = -(\nu + 1)s_t^{-1}\epsilon_t^2, \quad \ddot{p}(\epsilon_t) = (\nu + 1)s_t^{-1}[2s_t^{-1}\epsilon_t - 1],$$

where $s_t = (\nu - 2)\sigma_\epsilon^2 + \epsilon_t^2$.

But $2s_t^{-1}\epsilon_t - 1$ is not positive for all ϵ_t and hence when the iterative method for finding the mode based on two derivatives is implemented there is no guarantee that $\ddot{p}(\epsilon_t)$ will always be positive, something that is required for the use of the approximating linear Gaussian model to be employed.

Jungbacker and Koopman (2007) have a method of overcoming this. An alternative is to do the matching on the basis of first derivatives only as on the previous slide.

Application: linear Gaussian signal, heavy-tailed noise – 2

Given an initial guess at the mode $\tilde{\theta}$, set $\tilde{s}_t = (\nu - 2)\sigma_\epsilon^2 + \tilde{\epsilon}_t^2$ and let the observation variance at time t be

$$A_t = (\nu + 1)^{-1} s_t$$

in the linear Gaussian signal plus noise model with $\tilde{Y}_n = Y_n$ and all other terms as for the linear Gaussian signal component.

Use the Kalman filter and smoother as before to get an updated estimate of the mode $\tilde{\epsilon}^+$ and so on.

Application: linear Gaussian signal, heavy-tailed noise – 3

Mixtures of normals and more general heavy tailed distributions in the linear Gaussian signal plus noise model can also be handled by matching first derivatives only.

Note: this method will give an estimate of the mode of the signal θ given the observations Y_n which may all that is required in some applications. However, for deriving an importance sampling density it does not match the approximating Gaussian model in terms of its covariance structure as is the case for matching based on the first and second derivatives.

General linearization method for mode estimation

The above method of mode estimation was tailored to the situation where the non-Gaussian nonlinear aspects of the model are in the conditional distribution of the observations given the signal (or the state).

When this is not the case the second order expansion method can sometimes be used but a method based only on linearization can be applied to more general models.

– D&K §10.7.

Topic: 6: Importance Sampling for Smoothing and Estimation

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Importance sampling for smoothing and estimation

Importance Sampling - Overview

Basics of Importance Sampling

Implementation of Importance Sampling

Applications of Importance Sampling

Importance Sampling for Likelihoods

Objectives and Background

Estimate a conditional mean: Let $x(\alpha)$ be some function of α .

$$\bar{x} = E[x(\alpha) | Y_n] = \int x(\alpha) p(\alpha | Y_n) d\alpha. \quad (6.1)$$

Examples: $E(\alpha_t | Y_n)$, $\text{Var}(\alpha_t | Y_n)$

Other applications include estimating conditional densities and likelihoods which we discuss below.

Use of importance sampling methods dates to the 1950's and have been used extensively since then.

If random samples could be drawn of $\alpha | Y_n$ can be taken then a sample average over a sufficiently large number of these could be used to approximate (6.1).

Basic concepts of IS –1

Since, for our applications an explicit expression for $p(\alpha|Y_n)$ is not available, random samples of $\alpha|Y_n$ can be difficult to obtain.

Importance sampling seeks an approximating density $g(\alpha|Y_n)$ close to $p(\alpha|Y_n)$ (in some sense) from which samples are easily taken.

Sample averages over random draws from a random variable with density $g(\alpha|Y_n)$ are used to estimate a suitably adjusted integral.

To focus development for our objectives we consider the model

$$y_t \sim p(y_t|\alpha_t), \quad \alpha_{t+1} = T_t(\alpha_t) + R_t\eta_t, \quad \eta_t \sim p(\eta_t), \quad (6.2)$$

where $p(y_t|\alpha_t)$ and $p(\eta_t)$ may not be Gaussian densities.

We usually let $g(Y_n, \alpha)$, $g(Y_n)$, $g(Y_n|\alpha)$ and $g(\alpha|Y_n)$ be the Gaussian densities in the approximating linear Gaussian model to (6.2).

Basic concepts of IS –2

We re-express (6.1) as

$$\bar{x} = \int x(\alpha) \frac{p(\alpha|Y_n)}{g(\alpha|Y_n)} g(\alpha|Y_n) d\alpha = E_g \left[x(\alpha) \frac{p(\alpha|Y_n)}{g(\alpha|Y_n)} \right], \quad (6.3)$$

where E_g is expectation based on $g(\alpha|Y_n)$.

For most models we are interested in $p(\alpha|Y_n)$ and $g(\alpha|Y_n)$ are complicated. An alternative to (6.3) is

$$\bar{x} = \frac{g(Y_n)}{p(Y_n)} E_g \left[x(\alpha) \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)} \right] = \frac{g(Y_n)}{p(Y_n)} E_g [x(\alpha) w(\alpha, Y_n)] \quad (6.4)$$

where the **importance weights** are

$$w(\alpha, Y_n) = \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)}.$$

Basic concepts of IS –3

Note that putting $x(\alpha) = 1$ in (6.1) gives $\bar{x} = 1$ and equating this to (6.4) with $x(\alpha) = 1$ gives

$$1 = \frac{g(Y_n)}{p(Y_n)} E_g [w(\alpha, Y_n)]$$

so that marginal (observational) density of Y_n is obtained as

$$p(Y_n) = g(Y_n) E_g [w(\alpha, Y_n)]$$

and the conditional **mean of $x(\alpha)$** in (6.4) can be re-expressed as:

$$\bar{x} = \frac{E_g [x(\alpha)w(\alpha, Y_n)]}{E_g [w(\alpha, Y_n)]}, \quad \text{where} \quad w(\alpha, Y_n) = \frac{p(\alpha, Y_n)}{g(\alpha, Y_n)}. \quad (6.5)$$

Basic concepts of IS –4

The **importance sampling recipe**:

- ▶ Choose N independent draws $\alpha^{(1)}, \dots, \alpha^{(N)}$ from the importance density $g(\alpha|Y_n)$; and,
- ▶ Calculate the importance weighted sample average

$$\hat{x} = \frac{\frac{1}{N} \sum_{i=1}^N x_i w_i}{\frac{1}{N} \sum_{i=1}^N w_i}, \quad \text{where } x_i = x(\alpha^{(i)}) \quad \text{and} \quad w_i = w(\alpha^{(i)}, Y_n). \quad (6.6)$$

\hat{x} is consistent: As $N \rightarrow \infty$, $\hat{x} \xrightarrow{a.s.} \bar{x}$.

Follows from the strong law of large numbers which holds provided the importance weights w_i are not ‘degenerate’.

Special Case: Gaussian signal, non-Gaussian Observations.

Consider the following special case of (6.2):

$$y_t \sim p(y_t | \theta_t), \quad \theta_t = Z_t \alpha_t, \quad \alpha_{t+1} = T_t \alpha_t + R_t \eta_t, \quad \eta_t \sim N(0, Q_t), \quad (6.7)$$

and $R'_t R_t = I_r$ where r is the number of disturbances in η_t .

Here $p(\alpha) = g(\alpha)$ and hence

$$\frac{p(Y_n, \alpha)}{g(Y_n, \alpha)} = \frac{p(\alpha)}{g(\alpha)} \frac{p(Y_n | \alpha)}{g(Y_n | \alpha)} = \frac{p(Y_n | \alpha)}{g(Y_n | \alpha)} = \frac{p(Y_n | \theta)}{g(Y_n | \theta)}.$$

As a result, (6.5) becomes

$$\bar{x} = \frac{E_g [x(\alpha) w^*(\theta, Y_n)]}{E_g [w^*(\theta, Y_n)]}, \quad w^*(\theta, Y_n) = \frac{p(Y_n | \theta)}{g(Y_n | \theta)} \quad (6.8)$$

with the advantage that θ_t has dimension typically much smaller than that of α_t and, in particular, θ_t is scalar if the observations y_t are scalar.

Implementation of Importance Sampling

Two aspects covered:

- ▶ Implementation in terms of state noise η_t process. Often it is simpler to implement the importance sampling methodology using state disturbances $\eta_t = R'_t(\alpha_{t+1} - \alpha_t)$ and to do this reformulation of the above importance sampling formulae are required.
- ▶ Using antithetic variables. These provide a means of “doubling” up the benefit of a single draw and saving on simulation time.

Implementation using η

For an initial value α_1 and a sequence of state disturbances $\eta_1, \dots, \eta_t, \dots$ use the recursion $\alpha_{t+1} = T_t \alpha_t + R_t \eta_t$ to re-express $x(\alpha_t)$ as

$$x(\alpha_t) = x^*(\eta_1, \dots, \eta_{t-1}, \alpha_1)$$

or, for short and in vector form, $x(\alpha) = x^*(\eta)$. Note that (6.1) can be rewritten as

$$\bar{x} = \int x(\alpha) p(\alpha | Y_n) d\alpha = \int x^*(\eta) p(\eta | Y_n) d\eta \quad (6.9)$$

where $p(\eta | Y_n)$ is the conditional density of the state disturbances given the observations.

The analogue of (6.8) is

$$\bar{x} = \frac{E_g [x^*(\eta) w^*(\eta, Y_n)]}{E_g [w^*(\eta, Y_n)]}, \quad w^*(\eta, Y_n) = \frac{p(\eta, Y_n)}{g(\eta, Y_n)} \quad (6.10)$$

Implementation using η : Examples –1

In (6.10) E_g is w.r.t the importance density $g(\eta|Y_n)$, the conditional density of η given Y_n in the approximating model and

$$p(\eta, Y_n) = \prod_{t=1}^n p(\eta_t)p(y_t|\theta_t), \quad \theta_t = Z_t\alpha_t.$$

Some special cases:

- ▶ **Signal plus noise:** $y_t = \theta_t + \epsilon_t$, $p(y_t|\theta_t) = p(\epsilon_t)$ and similarly for the approximating model so that

$$p(\eta, Y_n) = \prod_{t=1}^n p(\eta_t)p(\epsilon_t), \quad g(\eta, Y_n) = \prod_{t=1}^n g(\eta_t)g(\epsilon_t).$$

Implementation using η : Examples –2

- ▶ **Linear, Gaussian state:** $p(\eta_t) = g(\eta_t)$ and the importance weights in (6.10) simplify to

$$w^*(\theta, Y_n) = \prod_{t=1}^n \frac{p(y_t | \theta_t)}{g(\epsilon_t)}$$

with further simplification in the Gaussian signal plus noise case to

$$w^*(\epsilon) = \prod_{t=1}^n \frac{p(\epsilon_t)}{g(\epsilon_t)}.$$

Antithetic variables – 1

For simulations based on random draws of η from the approximating Gaussian density $g(\eta|Y_n)$ the methods of D&K §4.9.1 (or see slides 155-158) can be used.

Efficiency of sampling is improved by getting
'two-for-the-price-of-one' via **antithetic variables**

An antithetic variable is a function of the random draw of the vector η **which** is equiprobable with η and when used in conjunction increases the efficiency of estimation.

Two types are typically recommended by Durbin and Koopman:

- ▶ Balanced for location.
- ▶ Balanced for scale.

We assume that the components of η are to be i.i.d draws from a normal distribution with known variance.

We will only (and briefly) discuss the first type. See D&K §11.4.3 and the references cited therein for more details of the second type.

Antithetic variables – 2

Antithetics balanced for location.

Let η be the first draw and compute $\hat{\eta} = E_g(\eta|Y_n)$ using the disturbance smoother (slides 155-158). Put $\check{\eta} = 2\hat{\eta} - \eta$. Then $\check{\eta} - \hat{\eta} = -(\eta - \hat{\eta}) \sim N$ with mean 0 and the same variance.

$\check{\eta}$ and η are equi-probable and conditional means based on them are negatively correlated.

Note, if both are used in the Monte-Carlo average of the type (6.6) then the negative correlation between the two terms in the sums will lead to a reduction in variance of the sample mean giving increased precision for no additional sampling effort.

Applications of Importance Sampling

- ▶ Estimating functions of the state vector:
 - ▶ mean functions D&K §11.5.1)
 - ▶ variance functions D&K §11.5.2
 - ▶ conditional densities D&K §11.5.3
 - ▶ conditional distribution functions D&K §11.5.4
 - ▶ forecasting and estimating with missing observations
D&K §11.5.5
- ▶ Maximum likelihood estimation: approximating (estimating) the likelihood, optimising it and getting standard errors
D&K §11.6

We will cover the use of importance sampling for likelihood estimation in detail next.

IS for likelihood estimation – 1

For the general non-linear non-Gaussian model with joint density of observations and states given by $p(Y_n, \alpha; \psi)$ which depends on unknown parameters ψ the likelihood can be written as

$$\begin{aligned} L(\psi) &= \int p(Y_n, \alpha; \psi) d\alpha \\ &= \int \frac{p(Y_n, \alpha; \psi)}{g(Y_n, \alpha; \psi)} g(Y_n, \alpha; \psi) d\alpha \\ &= g(Y_n; \psi) \int \frac{p(Y_n, \alpha; \psi)}{g(Y_n, \alpha; \psi)} g(\alpha|Y_n; \psi) d\alpha \\ &= L_g(\psi) E_g[w(\alpha, Y_n; \psi)], \end{aligned} \tag{6.11}$$

in which $L_g(\psi) = g(Y_n; \psi)$, $g(\alpha, Y_n; \psi)$ is an approximating Gaussian joint density, $g(\alpha|Y_n; \psi)$ is the Gaussian importance density (simply the conditional density of $\alpha|Y_n$ in the approximating Gaussian model for any ψ), E_g is expectation with respect to this.

IS for likelihood estimation – 2

importance weights: $w(\alpha, Y_n; \psi) = p(Y_n, \alpha; \psi)/g(Y_n, \alpha; \psi)$.

Notes:

- ▶ (6.11) expresses the likelihood for the original problem in terms of the *easily calculated Gaussian likelihood* for an approximating Linear Gaussian state space model times an adjustment that can be approximated using importance sampling.
- ▶ Note, unlike the treatment in D&K §11.6.1, we have kept ψ in the notation to stress that all quantities used to obtain the approximating Gaussian linear model and to perform the importance sampling do depend on ψ . Hence, as ψ is varied to obtain the maximum of the likelihood the approximating model will need to be recomputed and the importance sampling estimate recalculated.

IS for likelihood estimation – 3

Notes (continued):

- ▶ Any method of approximating the joint density $p(Y_n, \alpha)$ by $g(Y_n, \alpha)$ can be used but keep in mind the two key requirements: accuracy of approximation and ease of simulation in the approximating model. The modal approximations often perform well.
- ▶ Obviously if $p(Y_n, \alpha)$ and $g(Y_n, \alpha)$ are 'close' (up to a possible constant that may depend on ψ) for all α then the importance weights $w(\alpha, Y_n; \psi)$ will be close to a constant and the adjustment integral can be estimated with a Monte-Carlo sum based on a 'small' number of random samples.
- ▶ For computations, Durbin and Koopman recommend working with the signal $\theta = Z_t \alpha_t$ and the state disturbances η_t as discussed on slide 210. Additionally they recommend using antithetic variables. See D&K §11.6.1 for details.

Maximising the Likelihood – 1

Let $w_i(\psi) = w(\alpha^{(i)}, Y_n; \psi)$ be the importance weights in (6.11) evaluated at the i th draw from the importance density $g(\alpha | Y_n; \psi)$. Replacing the expectation in (6.11) by the sample average

$$\bar{w}(\psi) = \frac{1}{N} \sum_{i=1}^N w_i(\psi)$$

and defining estimated likelihood as $\hat{L}_{\text{DK}}(\psi) = L_g(\psi)\bar{w}(\psi)$ and taking logs gives

$$\log \hat{L}_{\text{DK}}(\psi) = \log L_g(\psi) + \log \bar{w}(\psi)$$

which is typically maximised by iterative numerical optimisation techniques.

In order to compare the methods suggested by Durbin and Koopman with others in the literature we introduce some notation for likelihood and approximate likelihood estimates. This notation is not the same as in D&K §11.6.

Maximising the Likelihood – 2

Denote by:

- ▶ $\hat{\psi}$ the parameter vector that maximises the theoretical likelihood (6.11). This true likelihood and its maximiser are typically not obtainable exactly for nonlinear non-Gaussian models.
- ▶ $\hat{\psi}_{\text{DK}}$ the maximiser of the simulation based approximate likelihood $\hat{L}_{\text{DK}}(\psi)$.

$\hat{L}_{\text{DK}}(\psi)$ is typically maximised using a numerical optimizer to iterate through ψ values from some initial value (more on how to start the iterations later).

Maximising the Likelihood – 3

Summary of procedure to optimize the approximate likelihood $L_{DK}(\psi)$.

1. Choose a 'suitable' starting value ψ .
2. Determine a linear Gaussian approximating model (typically with modal approximation). This approximating model depends on the current value of ψ .
3. Calculate the approximate likelihood $L_{DK}(\psi)$ by calculating $L_g(\psi)$ and, using the simulation smoother, calculate $\bar{w}(\psi)$.
4. Use an iterative numerical optimizer to provide an updated estimate - this may entail calculating $L_{DK}(\psi)$ at a number of points to create numerical first and possible second derivatives depending on the optimising method used.
5. Iterate to convergence.

It is important to ensure that at each iteration of the optimization process additional randomness is **not** introduced. This can be achieved by reusing the same random numbers for the simulation smoother.

Maximising the Likelihood – 4

At this stage we have an approximate estimate $\hat{\psi}_{\text{DK}}$. Other aspects that we need to consider are:

- ▶ Bias and consistency.
 - ▶ How close is $\hat{\psi}_{\text{DK}}$ to $\hat{\psi}$ (the MLE) as the simulation effort increases ($N \rightarrow \infty$)?
 - ▶ Is the correct MLE $\hat{\psi}$ consistent for the true ψ as the observed sample size $n \rightarrow \infty$?
 - ▶ How do we estimate finite sample bias?
- ▶ How do we estimate variability due to simulation as well as the usual variance of MLE estimation?
- ▶ Can we prove a CLT for the estimates?

We return to these topics later after we have introduced an alternative method of approximation.

Topic: 7: Likelihood Estimation using Laplace Approximation and Importance Sampling

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Importance sampling for smoothing and estimation

Likelihood Estimation using Laplace Approximation and
Importance Sampling

Laplace Approximation for nonlinear non-Gaussian models

Augmenting the Laplace Approximation by Importance
Sampling

Comparing Laplace and Gaussian approximate likelihoods

Laplace Approximation: nonlinear non-Gaussian models – 1

We revisit the model (5.9). Also recall the notation in (5.10). We state that again here for ease of reference:

$$Y_n \sim p(Y_n|\theta) = \prod_{t=1}^n p(y_t|\theta_t), \quad \theta \sim N(\mu, \Psi) \quad (7.1)$$

Now, let

$$F(\theta, Y_n) = \log p(\theta, Y_n) = \log p(Y_n|\theta) + \log p(\theta)$$

and write the likelihood (see slide 216, equation 6.11) as

$$L(\psi) = \int e^{F(\theta, Y_n)} d\theta. \quad (7.2)$$

It is obvious from (5.10) that the mode of the smoothed signal density $p(\theta|Y_n)$ is the same as the mode of the joint density $p(\theta, Y_n)$ with respect to θ .

Laplace Approximation: nonlinear non-Gaussian models – 2

Note:

- ▶ This mode can be obtained by maximising $F(\theta, Y_n)$ over θ by the Newton-Raphson iterations as on slide 184. Let $\dot{F}(\theta, Y_n) = \dot{p}(\theta, Y_n)$ and $\ddot{F}(\theta, Y_n) = \ddot{p}(\theta, Y_n)$.
- ▶ At the mode $\tilde{\theta}$ we have $\dot{F}(\tilde{\theta}, Y_n) = 0$ and the second order Taylor series of $F(\theta, Y_n)$ around the mode is

$$F_a(\theta, Y_n) = F(\tilde{\theta}, Y_n) - \frac{1}{2}(\theta - \tilde{\theta})'(\Psi^{-1} + A^{-1})(\theta - \tilde{\theta}) \quad (7.3)$$

- ▶ Hence

$$F(\theta, Y_n) = F_a(\theta, Y_n) + R(\theta, \tilde{\theta}) \quad (7.4)$$

where $R(\theta, \tilde{\theta})$ is the remainder in the Taylor series expansion.

Laplace Approximation: nonlinear non-Gaussian models – 3

Using (7.4) we rewrite the likelihood (7.2) as

$$\begin{aligned} L(\psi) &= \int e^{F(\tilde{\theta}, Y_n) - \frac{1}{2}(\theta - \tilde{\theta})'(\Psi^{-1} + A^{-1})(\theta - \tilde{\theta}) + R(\theta, \tilde{\theta})} d\theta \\ &= \frac{(2\pi)^{n/2}}{|\Psi^{-1} + A^{-1}|^{1/2}} e^{F(\tilde{\theta}, Y_n)} \\ &\quad \times \int \frac{|\Psi^{-1} + A^{-1}|^{1/2}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(\theta - \tilde{\theta})'(\Psi^{-1} + A^{-1})(\theta - \tilde{\theta})} e^{R(\theta, \tilde{\theta})} d\theta \\ &= L_{\text{LA}}(\psi) \int e^{R(\theta, \tilde{\theta})} p_a(\theta | Y_n) d\theta, \end{aligned} \tag{7.5}$$

where $p_a(\theta | Y_n)$ is the $N(\tilde{\theta}, (\Psi^{-1} + A^{-1})^{-1})$ distribution - we interpret this in a moment - and $L_{\text{LA}}(\psi)$ is the Laplace approximation to the likelihood $L(\psi)$ defined as ...

Laplace Approximation: nonlinear non-Gaussian models – 4

$$\begin{aligned}L_{\text{LA}}(\psi) &= \frac{(2\pi)^{n/2}}{|\Psi^{-1} + A^{-1}|} e^{F(\tilde{\theta}, Y_n)} \\&= \frac{(2\pi)^{n/2}}{|\Psi^{-1} + A^{-1}|^{1/2}} p(Y_n | \tilde{\theta}) p(\tilde{\theta}) \\&= \frac{1}{|\Psi|^{1/2} |\Psi^{-1} + A^{-1}|^{1/2}} p(Y_n | \tilde{\theta}) e^{-\frac{1}{2}(\tilde{\theta} - \mu)' \Psi^{-1} (\tilde{\theta} - \mu)} \\&= \frac{|\Psi^{-1}|^{1/2}}{|\Psi^{-1} + A^{-1}|^{1/2}} e^{\log p(Y_n | \tilde{\theta}) - \frac{1}{2}(\tilde{\theta} - \mu)' \Psi^{-1} (\tilde{\theta} - \mu)}. \quad (7.6)\end{aligned}$$

- ▶ $L_{\text{LA}}(\psi)$ is function of the mode $\tilde{\theta}$ and hence, once that is determined by the Newton-Raphson method, at convergence A is also determined.
- ▶ We let $\hat{\psi}_{\text{LA}}$ be the estimate obtained by maximising the Laplace approximate likelihood $L_{\text{LA}}(\psi)$.

Laplace Approximation with Importance Sampling – 1

The Laplace approximating likelihood can be improved:

- ▶ By using higher order Taylor series expansions - this has been done in a number of applications such as spatial modelling in which there is a latent Gaussian random field and observations are from the exponential family conditional on this field. So far little seems to have been done systematically for time series models – however Durbin and Koopman suggest use of higher order expansions to get starting values (discussed later).
- ▶ By using importance sampling to obtain a Monte Carlo estimate of the integral term in (7.5). We discuss this aspect next.

Laplace Approximation with Importance Sampling – 2

Denote the integral term in (7.5) as

$$r_a(\psi) = E(e^{R(\theta, \tilde{\theta}; \psi)}) = \int e^{R(\theta, \tilde{\theta}; \psi)} p_a(\theta | Y_n, \psi) d\theta \quad (7.7)$$

in which we have used ψ in the notation to emphasise dependence of all quantities on the unknown parameter.

If $p_a(\theta | Y_n; \psi)$ is highly concentrated around $\tilde{\theta}$ then the error term $R(\theta, \tilde{\theta}; \psi)$ in the Taylor series expansion will be close to 0 and hence the integrand will be close to 1 and $r_a(\psi)$ will also be close to 1.

Laplace Approximation with Importance Sampling – 3

As suggested in [Davis and Rodriguez-Yam, 2005], the correction term $r_a(\psi)$ in (7.7) can be estimated by Monte Carlo simulation by taking draws $\theta^{(i)}$ from $p_a(\theta|Y_n; \psi)$ as an importance sampler to get

$$\bar{r}_{\text{LA}}(\psi) = \frac{1}{N} \sum_{i=1}^N e^{R(\theta^{(i)}, \tilde{\theta}; \psi)} \quad (7.8)$$

We denote the resulting estimate of the log likelihood

$$\log \hat{L}_{\text{LA-IS}}(\psi) = \log L_{\text{LA}}(\psi) + \log \bar{r}_{\text{LA}}(\psi) \quad (7.9)$$

Laplace Approximation with Approximate Importance Sampling – 4

[Davis and Rodriguez-Yam, 2005] also suggest a refined approximation which is considerably faster to run and has good accuracy.

Recall that $\hat{\psi}_{\text{AL}}$ maximizes the Laplace approximate likelihood $L_{\text{LA}}(\psi)$ and that this is computed without simulation. We expand $\bar{e}(\psi) = \log \bar{r}_{\text{LA}}(\psi)$ around $\hat{\psi}_{\text{AL}}$ in a first order Taylor expansion to get

$$\tilde{\bar{e}}(\psi) = \bar{e}(\hat{\psi}_{\text{AL}}) + \frac{\partial \bar{e}(\psi)}{\partial \psi} \Big|_{\hat{\psi}_{\text{AL}}} (\psi - \hat{\psi}_{\text{AL}})$$

where the derivative is evaluated numerically. The approximate importance sampling method give

$$\log \hat{L}_{\text{LA-AIS}} = \log L_{\text{LA}} + \tilde{\bar{e}}(\psi)$$

and requires importance sampling only to get $\bar{e}(\hat{\psi}_{\text{AL}})$ and its derivatives numerically. This is a linear adjustment to the Laplace approximated likelihood.

Comparing $\hat{L}_{\text{DK}}(\psi)$ and $\hat{L}_{\text{LA-IS}}(\psi) - 1$

Equation (7.9) is

$$\log \hat{L}_{\text{LA-IS}}(\psi) = \log L_{\text{LA}}(\psi) + \log \bar{r}_{\text{LA}}(\psi)$$

while, Durbin and Koopman propose using

$$\log \hat{L}_{\text{DK}}(\psi) = \log L_g(\psi) + \log \bar{w}(\psi) \quad (7.10)$$

given on slide 219.

We next compare these two approximations to the likelihood.

Comparing $\hat{L}_{\text{DK}}(\psi)$ and $\hat{L}_{\text{LA-IS}}(\psi) - 2$

- ▶ Both use the same importance sampler, $p_a(\theta|Y_n; \psi)$, for obtaining their adjustment terms $\log \bar{r}_{\text{LA}}(\psi)$ and $\log \bar{w}(\psi)$ respectively.
- ▶ But L_{LA} and L_g are not the same. We will derive the relationship between these two shortly.
- ▶ As a result and in contrast to the use of $\log L_{\text{LA}}(\psi)$ in (7.9), the term $\log L_g(\psi)$ appearing in (7.10) cannot be used as a non-simulation based approximation to the correct likelihood without the adjustment term $\log \bar{w}(\psi)$ or some alternative approximate adjustment such as that considered next.

A non-simulation approximation to $\hat{L}_{\text{DK}}(\psi)$

[Durbin and Koopman, 1997] suggest using

$$\log \hat{L}_{\text{a,DK}}(\psi) = \log L_g(\psi) + \log \hat{w}(\psi) + \log \left(1 + \frac{1}{8} \sum_{t=1}^n \hat{l}_{\textcolor{red}{t}}^{(4)}(\psi) v_t^2 \right) \quad (7.11)$$

where $\hat{l}(\psi) = \log \hat{w}(\psi) = \log p(Y_n|\tilde{\theta}) - \log p_a(Y_n|\tilde{\theta})$,

$v_t = \text{Var}(\theta|Y_n) = [(\Psi^{-1} + A^{-1})^{-1}]_{tt}$ in the approximating model.

This does not require simulation and the fourth derivatives of $l(\psi)$ are easy to calculate for the examples given before. For example for the exponential family $\hat{l}^{(4)}(\psi) = -b^{(4)}(\tilde{\theta})$ (recalling that $\tilde{\theta}$ is a function of ψ).

Relationship between L_{LA} and $L_g - 1$

We first recall the definitions of the various distributions in the approximating Gaussian model used by Durbin and Koopman.

- ▶ $p(\theta)$ and $g(\theta) \sim N(\mu, \Psi)$ are the same because the state equation is linear and Gaussian.
- ▶ The approximating linear state space model uses ‘pseudo’ observations $\tilde{Y}_n = \tilde{\theta} + A\dot{p}(Y_n|\tilde{\theta})$ and assumes that $\tilde{Y}_n = \theta + \epsilon$ where $\epsilon \sim N(0, A)$ so that

$$g(\tilde{Y}_n|\theta) \sim N(\theta, A).$$

- ▶ The importance sampler is the approximating conditional distribution $g(\theta|\tilde{Y}_n) \sim N(\tilde{\theta}, (A^{-1} + \Psi^{-1})^{-1})$. This is the same as the importance sampler defined under (7.5). Thus Durbin and Koopman and Davis and Rodriguez-Yam use the same importance sampler.

Relationship between L_{LA} and $L_g - 2$

We now derive the likelihood for the approximating linear Gaussian state space model. Note that

$$\begin{aligned} g(\tilde{Y}_n) &= \frac{g(\tilde{Y}_n|\theta)g(\theta)}{g(\theta|\tilde{Y}_n)} \\ &= \frac{|A^{-1}|^{1/2}|\Psi^{-1}|^{1/2}}{(2\pi)^{n/2}|A^{-1} + \Psi^{-1}|^{1/2}} \exp\left\{-\frac{1}{2}[Q(\tilde{Y}_n|\theta) + Q(\theta) - Q(\theta|\tilde{Y}_n)]\right\}, \end{aligned} \tag{7.12}$$

where

$$\begin{aligned} Q(\tilde{Y}_n|\theta) &= (\tilde{\theta} + A\dot{p}(Y_n|\tilde{\theta}) - \theta)'A^{-1}(\tilde{\theta} + A\dot{p}(Y_n|\tilde{\theta}) - \theta) \\ &= (\theta - \tilde{\theta} - A\dot{p}(Y_n|\tilde{\theta}))'A^{-1}(\theta - \tilde{\theta} - A\dot{p}(Y_n|\tilde{\theta})) \\ &= (\theta - \tilde{\theta})'A^{-1}(\theta - \tilde{\theta}) - 2(\theta - \tilde{\theta})\Psi^{-1}(\tilde{\theta} - \mu) \\ &\quad + \dot{p}(Y_n|\tilde{\theta})'A\dot{p}(Y_n|\tilde{\theta}) \end{aligned}$$

where we have used the fact that $\dot{p}(Y_n|\tilde{\theta}) = \Psi^{-1}(\tilde{\theta} - \mu)$ at convergence to the mode.

Relationship between L_{LA} and $L_g - 3$

Next

$$\begin{aligned} Q(\theta | \tilde{Y}_n) &= (\theta - \tilde{\theta})' (A^{-1} + \Psi^{-1})(\theta - \tilde{\theta}) \\ &= (\theta - \tilde{\theta})' A^{-1}(\theta - \tilde{\theta}) + (\theta - \tilde{\theta})' \Psi^{-1}(\theta - \tilde{\theta}). \end{aligned}$$

The first term in this expression cancels with the first term in the expression for $Q(\tilde{Y}_n | \theta)$ so that

$$\begin{aligned} Q(\tilde{Y}_n | \theta) + Q(\theta) - Q(\theta | \tilde{Y}_n) &= -2(\theta - \tilde{\theta})\Psi^{-1}(\theta - \mu) + \dot{p}(Y_n | \tilde{\theta})' A \dot{p}(Y_n | \tilde{\theta}) \\ &\quad + (\theta - \mu)\Psi^{-1}(\theta - \mu) - (\theta - \tilde{\theta})' \Psi^{-1}(\theta - \tilde{\theta}) \\ &= (\tilde{\theta} - \mu)' \Psi^{-1}(\tilde{\theta} - \mu) + \dot{p}(Y_n | \tilde{\theta})' A \dot{p}(Y_n | \tilde{\theta}). \end{aligned}$$

Relationship between L_{LA} and $L_g - 4$

Now substituting this quadratic form in (7.12) and noting that the approximating Gaussian likelihood is $L_g(\psi) = g(\tilde{Y}_n)$ we have

$$\begin{aligned} L_g(\psi) &= \frac{|A^{-1}|^{1/2} |\Psi^{-1}|^{1/2}}{(2\pi)^{n/2} |A^{-1} + \Psi^{-1}|^{1/2}} \exp\left\{-\frac{1}{2}(\tilde{\theta} - \mu)' \Psi^{-1} (\tilde{\theta} - \mu) - \frac{1}{2}\dot{p}(Y_n|\tilde{\theta})\right\} \\ &= \frac{|\Psi^{-1}|^{1/2}}{|A^{-1} + \Psi^{-1}|^{1/2}} \exp\left\{\log p(Y_n|\tilde{\theta}) - \frac{1}{2}(\tilde{\theta} - \mu)' \Psi^{-1} (\tilde{\theta} - \mu)\right\} \\ &\quad \times \frac{|A^{-1}|^{1/2}}{(2\pi)^{n/2}} \exp\left\{-\log p(Y_n|\tilde{\theta}) - \frac{1}{2}\dot{p}(Y_n|\tilde{\theta})' A \dot{p}(Y_n|\tilde{\theta})\right\} \\ &= L_{\text{LA}}(\psi) \times B(\psi) \end{aligned}$$

where

$$B(\psi) = \frac{|A^{-1}|^{1/2}}{(2\pi)^{n/2}} \exp\left\{-\log p(Y_n|\tilde{\theta}) - \frac{1}{2}\dot{p}(Y_n|\tilde{\theta})' A \dot{p}(Y_n|\tilde{\theta})\right\}$$

Relationship between L_{LA} and $L_g - 5$

Notes:

- ▶ The Laplace approximation to the likelihood for Poisson response autoregressive state process was proposed in [Davis et al., 1999]. It was further refined for this case in [Davis and Rodriguez-Yam, 2005]. The above derivation is more general and covers all conditionally independent, Gaussian signal type models.
- ▶ Computations are done with the innovations algorithm (see [Brockwell and Davis, 2009] for example) which is particularly fast for the autoregressive state equation.
- ▶ This connection between L_{LA} and L_g is shown in [Davis and Rodriguez-Yam, 2005].
- ▶ The fact that $L_a(\psi) \neq L_g(\psi)$ demonstrates that $L_g(\psi)$ cannot be used as an approximate likelihood without adjustment based on importance sampling.

Application to the Polio Data

Table : Estimates and standard errors for key parameters in various methods applied to the polio series. Note: $\hat{\sigma}_\alpha^2 = \hat{\sigma}_\epsilon^2 / (1 - \hat{\phi}^2)$.

Method (source)	$\hat{\beta}_2$	se($\hat{\beta}_2$)	$\hat{\phi}$	se($\hat{\phi}$)	$\hat{\sigma}^2$	$\hat{\sigma}_\alpha^2$
MCEM [Chan and Ledolter, 1995]	-4.62	1.38	0.89	0.04	0.09	0.41
MCEM[NL] [McCulloch, 1997]	-4.35	1.96	0.10	0.36	0.50	0.51
Bayes [Oh and Lim, 2001]	-4.24	1.72	0.66	0.16	0.32	0.56
PQL[NL] [Breslow and Clayton, 1993]	-3.46	3.04	0.70	0.13	0.26	0.51
AL [Davis and Rodriguez-Yam, 2005]	-3.81	2.77	0.63	0.23	0.29	0.48
AL-BC [Davis and Rodriguez-Yam, 2005]	-3.96	2.77	0.73	0.23	0.30	0.65
AIS [Davis and Rodriguez-Yam, 2005]	-3.75	2.87	0.66	0.21	0.27	0.48
AIS-BC [Davis and Rodriguez-Yam, 2005]	-3.76	2.87	0.73	0.21	0.30	0.64
MCNR [Kuk and Cheng, 1999]	-3.82	2.77	0.67	0.18	0.27	0.48
EIS [Jung and Liesenfeld, 2001]	-3.61	2.57	0.68	0.15	0.26	0.48
GLM [Davis et al., 2000]	-4.80	4.11	—	—	—	—
GEE [Zeger, 1988]	-4.35	2.68	0.82	—	0.19	0.57
CPL ₂ [Davis and Yau, 2011]	-4.74	2.54	0.49	0.21	0.37	0.49
IBC[NL] [Kuk, 1995]	-5.01	3.20	0.54	0.28	0.35	0.49

Notes:

- ▶ From [Davis and Dunsmuir, 2014]
- ▶ The use of automatic differentiation as in [Skaug, 2002] gives identical results to the AL method and, because of this are not recorded in Table 1. Further, use of $M = 100$, $M = 1,000$ or $M = 5,000$ importance samples as reported independently by [Skaug, 2002] have very little impact on point estimates.

Additional Notes and Exercises on Approximate Likelihoods – 1

Exercise 7.1 Derive the approximation given in (7.11).

D&K also suggest the simpler approximation

$$\log \hat{L}_{\text{a,DK-Simplified}}(\psi) = \log L_g(\psi) + \log \hat{w}(\psi) \quad (7.13)$$

Exercise 7.2 Show that $\hat{L}_{\text{a,DK-Simplified}}(\psi) = L_{\text{LA}}(\psi)$, the Laplace approximate likelihood derived by [Davis and Rodriguez-Yam, 2005].

Exercise 7.3 On slide 239 it was shown that

$L_{\text{LA}}(\psi) = L_g(\psi)/B(\psi)$. Show that

$$1/B(\psi) = \frac{p(Y_n|\tilde{\theta})}{g(\tilde{Y}_n|\tilde{\theta})}$$

where $g(\tilde{Y}_n|\tilde{\theta})$ is the normal density with mean $\tilde{\theta}$ and diagonal covariance A evaluated at \tilde{Y}_n .

Additional Notes and Exercises on Approximate Likelihoods

- 2

Hence:

- ▶ The Laplace approximation can be calculated efficiently using the Kalman filter and smoothing methods applied to the approximating state space model. which gives L_g . Calculation of the \hat{w} is easily done for the conditionally independent case because it is the ratio of products of marginal densities. See later for applications of this using R.
- ▶ Importance sampling to improve the Laplace approximation can also be easily implemented - we discuss this in more detail later for the Poisson regression model.

Notes on large sample properties of MLE's –1

Refer to slide 222.

There is very little research on the large sample properties of MLE's for these models. Here is what we anticipate will be true:

- ▶ The unrealisable but true maximum likelihood estimate $\hat{\psi}$ will be consistent ($\hat{\psi} \xrightarrow{a.s.} \psi_0$) and asymptotically normally distributed ($\hat{\psi} - \psi_0 \xrightarrow{\text{approx}} N(0, \Omega/n)$) as $n \rightarrow \infty$ where Ω is estimated by

$$\hat{\Omega} = \left[-\frac{\partial^2 \log L(\psi)}{\partial \psi \partial \psi'} \right]^{-1} \Big|_{\psi=\hat{\psi}}$$

These derivatives are not available in closed form and need to be estimated numerically.

Notes on large sample properties of MLE's –2

- ▶ The Laplace, $\hat{\psi}_{LA}$ and the Durbin-Koopman, $\hat{\psi}_{DK}$ estimates will not be consistent but will have small bias in applications considered here. Their large sample distribution may not be the exact normal distribution but that this may be reasonably accurate for practical inference. The asymptotic variance would need to be estimated, as above, using numerical second derivatives.
- ▶ The importance sample augmented Laplace or Durbin-Koopman approximate estimates should be consistent and asymptotically normal as both $n \rightarrow \infty$ and $N \rightarrow \infty$. However estimation of covariance matrices will need to be done carefully by using the same random numbers for evaluation of the numerical second derivatives.
- ▶ D&K §11.6 have more to say on the effect of simulation variability on overall mean squared error.

References I

-  Breslow, N. E. and Clayton, D. G. (1993).
Approximate inference in generalized linear mixed models.
Journal of the American Statistical Association, 88(421):9–25.
-  Brockwell, P. J. and Davis, R. A. (2009).
Time series: theory and methods.
Springer.
-  Chan, K. and Ledolter, J. (1995).
Monte carlo em estimation for time series models involving counts.
Journal of the American Statistical Association, 90(429):242–252.
-  Davis, R. A. and Dunsmuir, W. (2014).
State space models for count time series.
Handbook of Discrete Valued Time Series, page xx.

References II

-  Davis, R. A., Dunsmuir, W., and Wang, Y. (1999).
Modeling time series of count data.
STATISTICS TEXTBOOKS AND MONOGRAPHS,
158:63–114.
-  Davis, R. A., Dunsmuir, W. T., and Wang, Y. (2000).
On autocorrelation in a poisson regression model.
Biometrika, 87(3):491–505.
-  Davis, R. A. and Rodriguez-Yam, G. (2005).
Estimation for state-space models based on a likelihood approximation.
Statistica Sinica, 15(2):381–406.
-  Davis, R. A. and Yau, C. Y. (2011).
Comments on pairwise likelihood in time series models.
Statistica Sinica, 21(1):255.

References III

-  Durbin, J. and Koopman, S. J. (1997). Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84(3):669–684.
-  Jung, R. C. and Liesenfeld, R. (2001). Estimating time series models for count data using efficient importance sampling. *AStA Advances in Statistical Analysis*, 4.
-  Kuk, A. Y. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 395–407.

References IV

-  Kuk, A. Y. and Cheng, Y. W. (1999).
Pointwise and functional approximations in monte carlo maximum likelihood estimation.
Statistics and Computing, 9(2):91–99.
-  McCulloch, C. E. (1997).
Maximum likelihood algorithms for generalized linear mixed models.
Journal of the American statistical Association, 92(437):162–170.
-  Oh, M.-S. and Lim, Y. B. (2001).
Bayesian analysis of time series poisson data.
Journal of Applied Statistics, 28(2):259–271.

References V

-  Skaug, H. J. (2002).
Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models.
Journal of Computational and Graphical Statistics,
11(2):458–470.
-  Zeger, S. L. (1988).
A regression model for time series of counts.
Biometrika, 75(4):621–629.

Topic: 8: Examples and Applications

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Importance sampling for smoothing and estimation

Likelihood Estimation using Laplace Approximation and
Importance Sampling

Examples and Applications

Application to modelling the Polio data

Application to modelling the Polio data

Modelling the Polio data

The regression model is as described in Davis and Rodrigues-Yam (2006).

$$y_t | \theta_t \stackrel{\text{indep}}{\sim} \text{Po}(e^\theta),$$

$$\theta_t = x_t' \beta + \alpha_t,$$

$$\alpha_t = \phi \alpha_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2).$$

This has state space representation with expanded state vector
 $\alpha_t^* = (\beta, \alpha_t)', T_t = \text{diag}(1, \dots, 1, \phi), R_t = (0, \dots, 0, 1)', Q_t = \sigma_\eta^2.$

This is specified in the KFAS package using code in the file 'PolioDataModelling.R'.

You will also need to functions defined in 'functions.R'

Estimates of Polio Regression Model using 'KFS'

	pars_KFS_DRY_AL	pars_KFS_DRY_AIS
Intcpt	0.3497	0.3426
Trend	-3.6316	-3.5685
CosAnnual	0.1557	0.1553
SinAnnual	-0.4611	-0.4601
CosSemiAnnual	0.3974	0.3988
SinSemiAnnual	-0.0088	-0.0089
phi	0.6270	0.6610
σ_η^2	0.2890	0.2720
logL	-260.9102	-261.2267

Note that the regression parameter estimates are NOT maximum likelihood estimates.

Plot of Polio data with mode of $\hat{\theta}$ plus regression

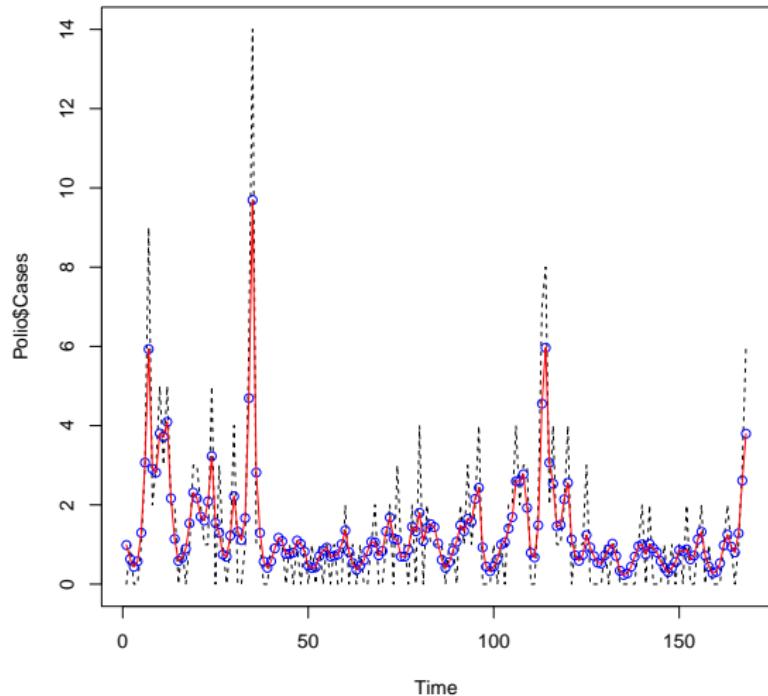
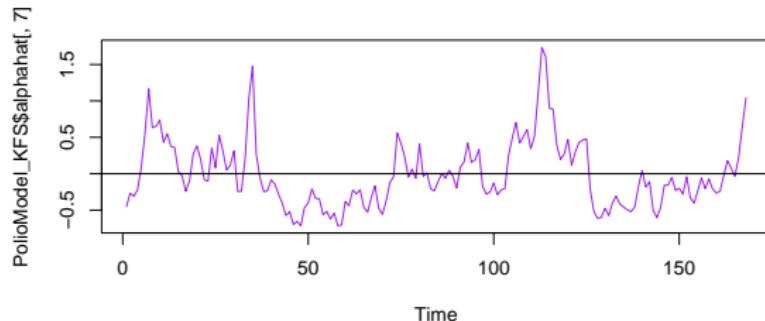


Figure : 8.1.

alphahat and its autocorrelation function



Series `PolioModel_KFS$alphahat[7]`

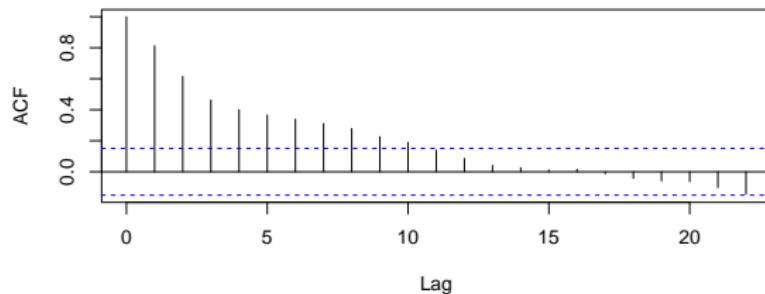


Figure : 8.2.

Convergence of Newton Raphson method for mode estimate -1

Starting iterations at $\tilde{\theta} = 0$ (see red line on next slide)

```
> print(PolioModel_approx_mode$results)
      logL_g      Accuracy
[1,] -296.9465 4.015048e+00
[2,] -246.9921 1.325605e-01
[3,] -250.7655 1.849433e-02
[4,] -251.7817 1.932665e-03
[5,] -251.9634 3.398019e-05
[6,] -251.9671 1.510370e-08
[7,] -251.9671 4.405045e-15
```

Note the direct implementation is fast (7 iterations). Using KFS takes 367 !!!

Convergence of Newton Raphson method for mode estimate -2

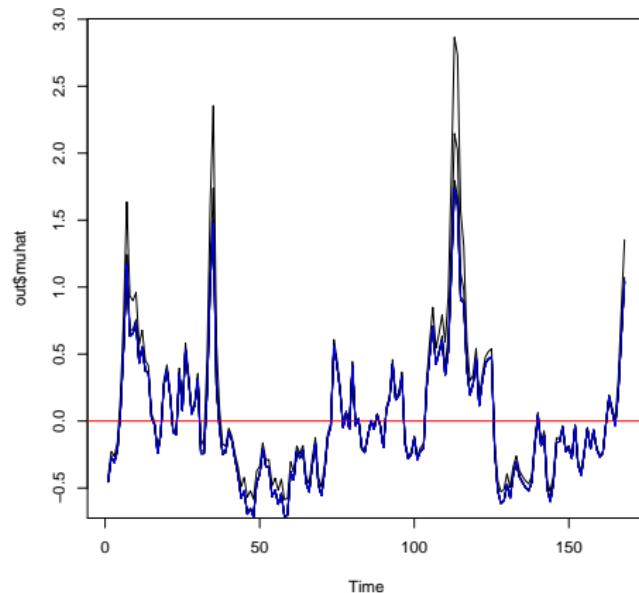


Figure : 8.3.

Polio Data with Posterior Mode

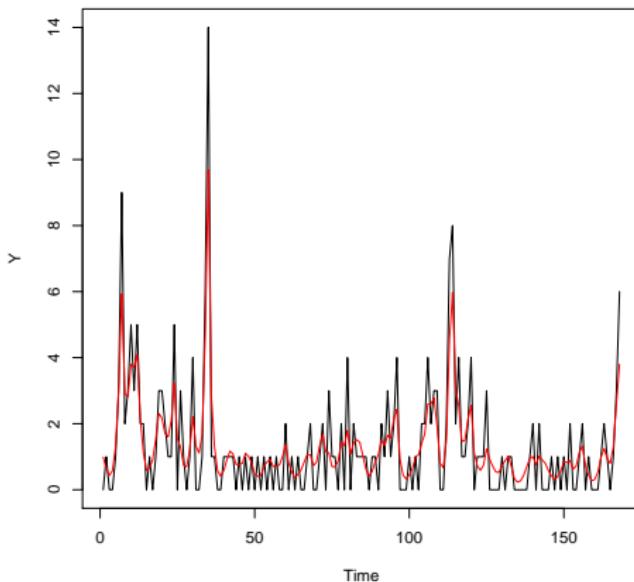


Figure : 8.4.

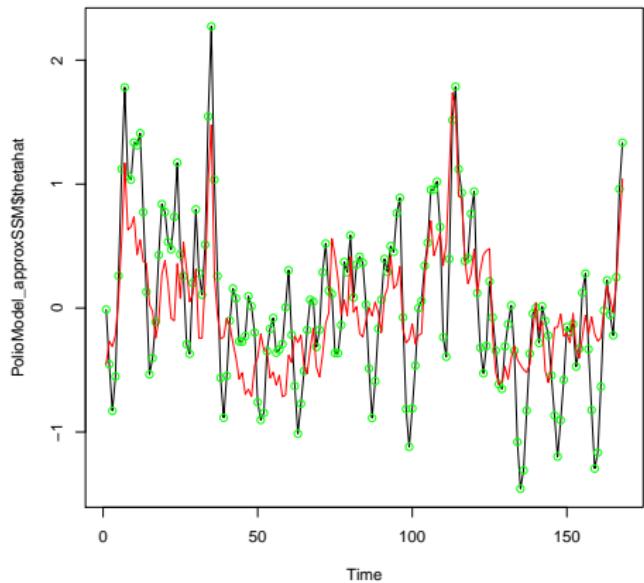


Figure : 8.5.

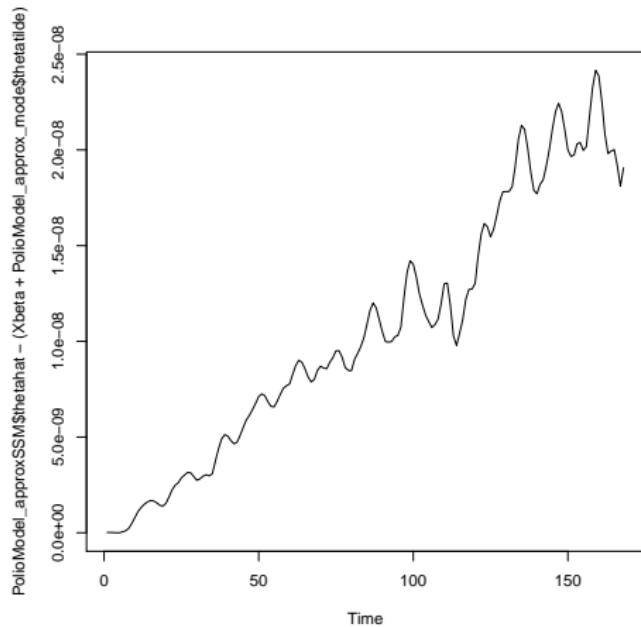


Figure : 8.6.

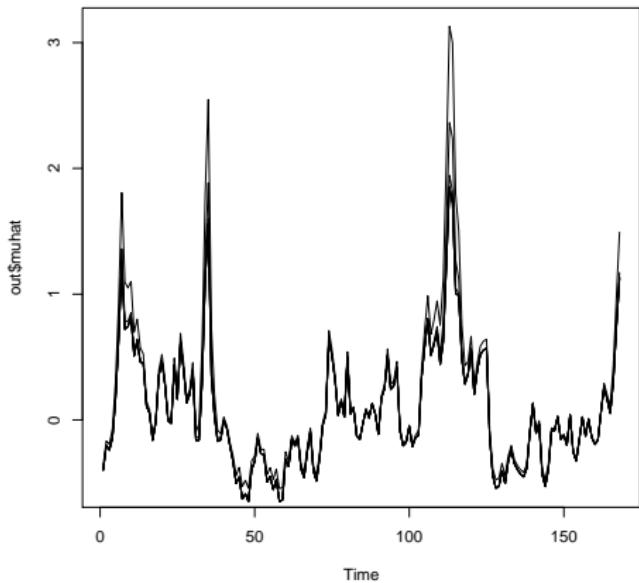


Figure : 8.7.

Polio Data with Posterior Mode

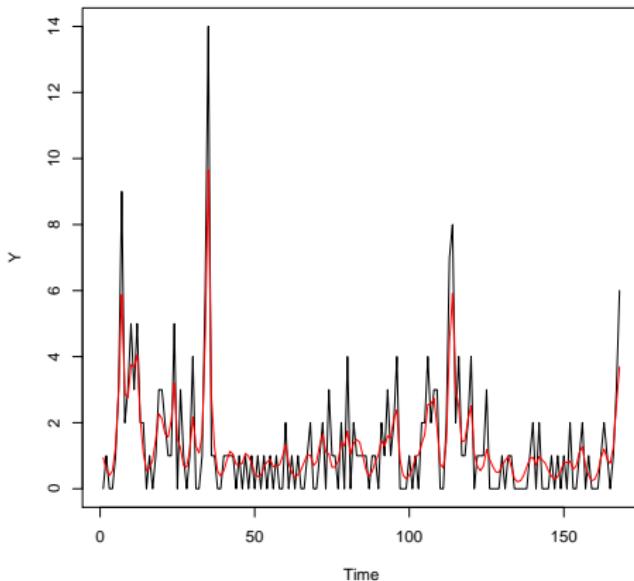


Figure : 8.8.

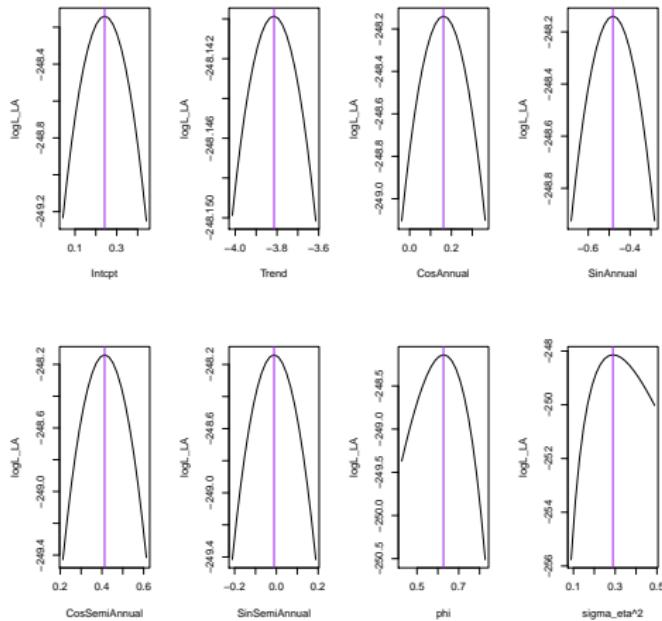
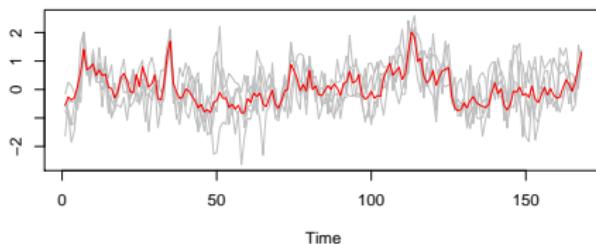


Figure : 8.9.

Simulations $p(\alpha|Y_{\tilde{t}})$ in approx Gaussian model
red=thetatilde



Simulations $p(\alpha|Y_{\tilde{t}})$ in approx Gaussian model
red=thetatilde

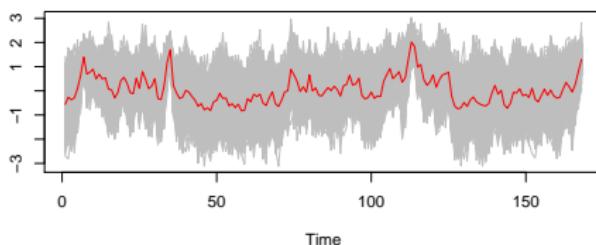


Figure : 8.10.

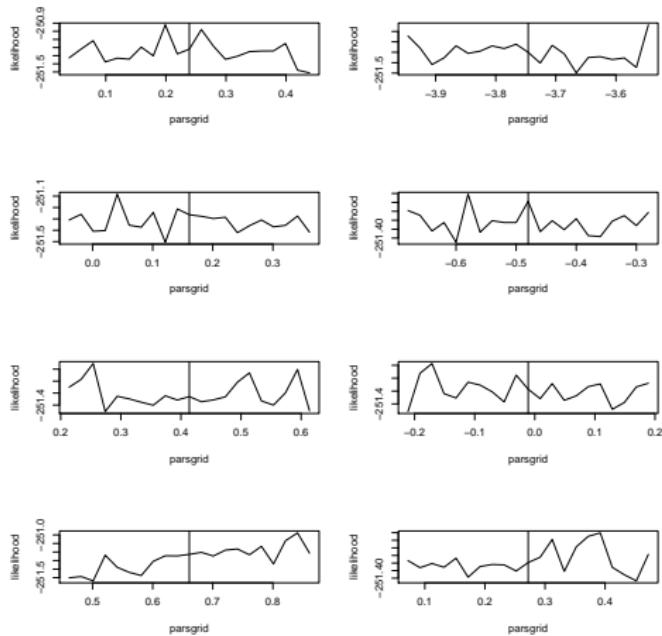


Figure : 8.11.

Simulation of the Polio data

Refer to the R script file 'SimulatePolioData.R' for code to produce the following analysis and graphs.

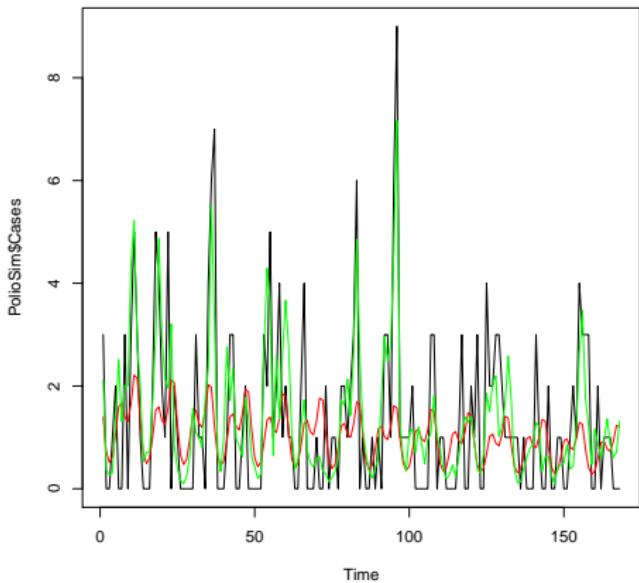


Figure : 8.12.

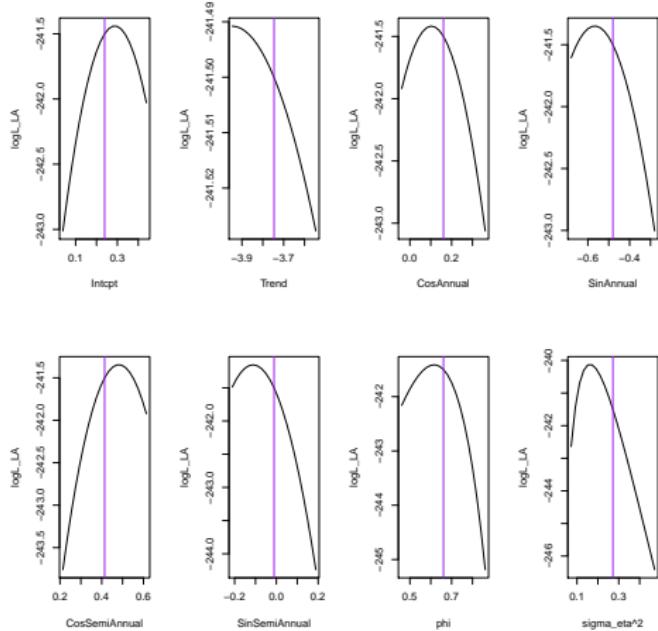


Figure : 8.13.

Convergence of Newton Raphson - 5 iterations

Iterations

	pars	pars	pars_new	L_LA_d1	S.E.s
Intcpt	0.239	0.21076079	0.21076079	4.626140e-07	0.2275289
Trend	-3.746	-2.91307400	-2.91307391	4.883247e-08	2.2785483
CosAnnual	0.161	0.22332647	0.22332647	5.818291e-08	0.1358450
SinAnnual	-0.480	-0.51079189	-0.51079189	-9.258500e-08	0.1551296
CosSemiAnnual	0.414	0.44605479	0.44605479	5.693432e-08	0.1294419
SinSemiAnnual	-0.011	-0.02385042	-0.02385041	-7.421111e-09	0.1259246
phi	0.661	0.46649522	0.46649519	-1.226092e-06	0.1910198
sigma_eta^2	0.272	0.31658023	0.31658023	-5.423795e-07	0.1382880

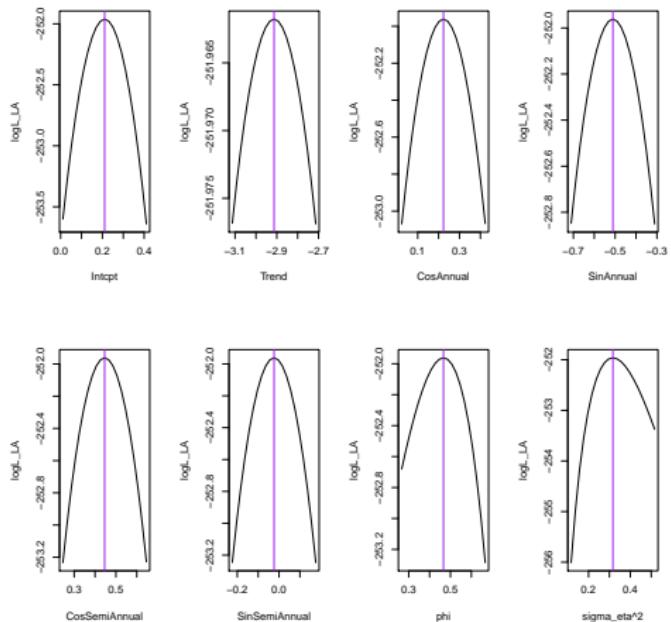
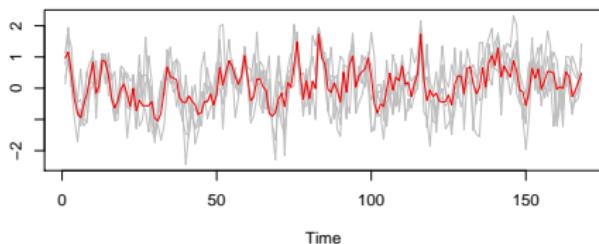


Figure : 8.14.

Simulations $p(\alpha|\tilde{Y_t})$ in approx Gaussian model
red=thetatilde



Simulations $p(\alpha|\tilde{Y_t})$ in approx Gaussian model
red=thetatilde

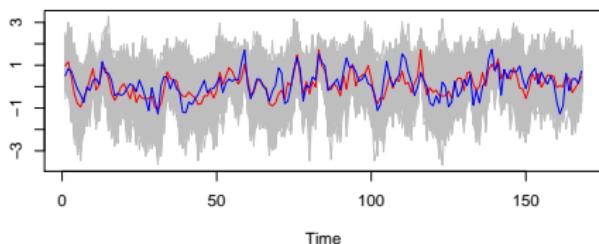


Figure : 8.15.

Simulations $p(\alpha|Y_{\tilde{t}})$ in approx Gaussian model
red=thetatilde

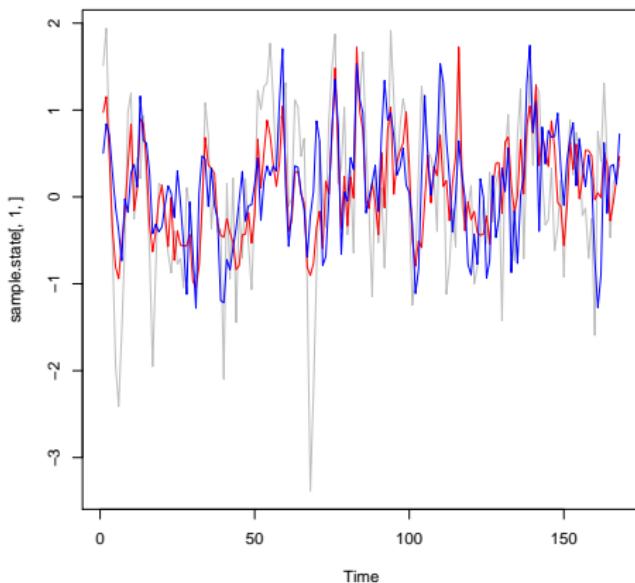


Figure : 8.16.

Topic: 9: Some Theory for GLARMA models

Nonlinear non-Gaussian State Space Models

Local Level Model

Linear State Space Models

Linear State Space Model - Filtering, Smoothing and Forecasting

Approximate Filtering and Smoothing

Importance sampling for smoothing and estimation

Likelihood Estimation using Laplace Approximation and
Importance Sampling

Examples and Applications

Observation Driven Models

Let $\mathcal{H}_t = (Y^{(t-1)}, X^{(t)})$ be the past of the observed count process and the past and present of the regressor variables and assume that the conditional distribution of $Y_t | \mathcal{H}_t$ is a member of the exponential family

$$f(y_t | W_t) = \exp[y_t w_t - m_t b(w_t)) + c(y_t)]$$

for which

$$\mu_t = E(Y_t | W_t) = \dot{b}(W_t), \quad \text{var}(Y_t | W_t) = a(\phi) \ddot{b}(W_t),$$

where \dot{b} and \ddot{b} denote first and second derivatives.

State Process

Typically the state process, W_t is expressed using a link function $g(\cdot)$ as in

$$g(\dot{b}(W_t)) = x_t^T \beta + h(y^{(t-1)}, X^{(t-1)}; \delta)$$

where $h(\cdot)$ is a function to be specified in terms of past observations and their conditional means as well as unknown parameters $\delta^T = (\beta^T, \tau^T)$, τ^T being the additional parameters required to specify the way in which past observations enter into the conditional mean. In the examples discussed here the canonical link will typically be used so that $g(\dot{b}(W_t)) = W_t$.

Some Choices for State Equations

$$h(y^{(t-1)}, x^{(t-1)}; \delta) = \sum_{i=1}^q \gamma_i [\log(\max(y_{t-i}, c)) - x_t^T \beta]$$

$$h(y^{(t-1)}, x^{(t-1)}; \delta) = \sum_{i=1}^q \gamma_i [\log(y_{t-i} + c) - \log(\exp(x_t^T \beta) + c)].$$

$$h(y^{(t-1)}, x^{(t-1)}; \delta) = \sum_{i=1}^q \gamma_i y_{t-i}$$

$$h(y^{(t-1)}, x^{(t-1)}; \delta) = \sum_{i=1}^q \gamma_i [y_{t-i} - \exp(x_t^T \beta)]$$

General form of state equation

Benjamin et al. (2003)'s generalized autoregressive moving average models (GARMA) specify $\mu_t = E(Y_t | \mathcal{H}_t)$ through a link function $g(\mu_t) = \eta_t = x_t^T \beta + \zeta_t$ where

$$\zeta_t = \sum_{j=1}^p \phi_j \mathcal{A}(y_{t-j}, x_{t-j}, \beta) + \sum_{j=1}^q \theta_j \mathcal{M}(y_{t-j}, \mu_{t-j})$$

where $\mathcal{A}(y_{t-j}, x_{t-j}, \beta)$ and $\mathcal{M}(y_{t-j}, \mu_{t-j})$ are functions representing the autoregressive and moving average terms.

Eg. Brumback et al. (2000)'s “transitional GLM” models the health impacts of pollution.

GLARMA models

Infinite distributed lag models:

$$h(y^{(t-1)}, x^{(t-1)}; \tau) = Z_t = \sum_{i=1}^{\infty} \gamma_i e_{t-i}$$

Consider distributed lag structures that are generated by the linear predictor for autoregressive-moving average processes of the form

$$\sum_{i=1}^{\infty} \gamma_i z^i = (1 - \sum_{i=1}^p \phi_i z^i)^{-1} (1 + \sum_{i=1}^q \theta_i z^i) - 1.$$

[Shephard (1995), Davis et al (1999)]

$$e_s = \frac{y_s - \mu_s}{v(\mu_s)}$$

GLARMA models

Infinite distributed lag models (GLARMA models):

$$W_t = g(x_t; \beta) + Z_t, \quad Z_t = \sum_{i=1}^{\infty} \gamma_i(\delta) e_{t-i}$$

where e_t are suitably defined 'residuals' or 'innovations'.

Shephard (1995), Davis et al (1999) consider distributed lag structures that are generated recursively by the linear predictor for autoregressive moving average processes:

$$Z_t = \phi_1(Z_{t-1} + e_{t-1}) + \cdots + \phi_p(Z_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q},$$

for initial values set $e_t = 0$, $Z_t = 0$ for $t \leq 0$.

Choice of residuals

- ▶ Davis et al. (1999, 2003) for Poisson observations in which $\nu(\mu_s) = \mu_s^\lambda$ where $0 < \lambda$. Note when $\lambda = 0.5$ scaling is by the standard deviation $\nu(\mu_s) = \mu_s^{0.5}$.
- ▶ Bernoulli outcomes (Rydberg & Shephard 2003) where $\nu(\mu_s) = \sqrt{\mu_s(1 - \mu_s)}$
- ▶ Binomial outcomes with m_s trials at time s (Lu 2002) where $\nu(\mu_s) = \sqrt{m_s\mu_s(1 - \mu_s)}$
- ▶ Alternative residuals such as the Anscombe and deviance residuals could be considered.

Martingale Property

Under the initial conditions $e_s = 0$ let $\mathcal{F}_{s-1}^e = \{e_t : t \leq s-1\}$ the e_t form a martingale difference sequence with zero mean and unit variance except for the Poisson case.

$$E(e_t^2) = E(\mu_t^{1-2\lambda}), \quad t \geq 1,$$

When $\lambda = 0.5$, $\{e_t\}$ is a weakly stationary white noise process and as t gets large

$$Z_t = \sum_{i=1}^{\infty} \gamma_i e_{t-i}$$

tends to a weakly stationary process with $E(Z_t) = 0$.

General Scaling

For any λ ,

$$E(W_t) = x_t^T \beta,$$

$$\text{var}(W_t) = \text{var}(Z_t) = \sum_{i=1}^t \gamma_i^2 E(\mu_{t-i}^{1-2\lambda}),$$

and for $h > 0$,

$$\text{cov}(W_t, W_{t+h}) = \text{cov}(Z_t, Z_{t+h}) = \sum_{i=1}^t \gamma_i \gamma_{i+h} E(\mu_{t-i}^{1-2\lambda}).$$

Let $\lambda = 0.5$. If $\{Z_t\}$ was a stationary Gaussian process the unconditional mean of Y_t would be exactly

$$E(Y_t) = E(e^{W_t}) = e^{x_t^T \beta + \frac{1}{2} \text{var}(Z_t)} = e^{x_t^T \beta + \frac{1}{2} \sum_{i=1}^{\infty} \gamma_i^2}.$$

giving the usual interpretation of the regression coefficients and an adjustment to the intercept term.

Let $\lambda = 0.5$ then $\{e_t\}$ is a weakly stationary martingale difference sequence and

$$E(Y_t) = E(e^{W_t}) \doteq \exp(x_t^T \beta + \frac{1}{2} \text{var}(Z_t)) \doteq \exp(x_t^T \beta + \frac{1}{2} \sum_{i=1}^{\infty} \gamma_i^2)$$

holds.

- ▶ If the mean process μ_t is large then the approximation is likely to be good (need e_t to behave like a Gaussian process.)
- ▶ Anscombe or deviance residuals may improve the approximation to the normal distribution (Pierce & Schafer, 1986).

Stationarity & ergodicity for GLARMA

Important

- ▶ To ensure that the process does not degenerate to zero nor grow without bound as time progresses.
- ▶ For establishing the large sample distributional properties of parameter estimates.

Simple Model

Davis et al. (2003) study the simple model with $p = 0$, $q = 1$ and $x_t^T \beta = \beta$ and state process

$$W_t = \beta + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}. \quad (9.1)$$

- ▶ The observation process $\{Y_t\}$ is not Markov - $E(Y_t|\mathcal{F}_{t-1}^Y) \neq E(Y_t|T_{t-1})$.
- ▶ BUT the state process $\{W_t\}$ is Markov which is all that is required for asymptotic theory.

Existence of stationary distribution

- ▶ The process $\{W_t\}$ is uniformly ergodic for the case $\lambda = 1$. Hence, there exist unique stationary distributions for both the log-conditional mean and conditional mean processes.
- ▶ For $1/2 \leq \lambda < 1$, there exists a stationary distribution, yet the uniqueness of such a distribution is currently unknown.
- ▶ For other values, $0 < \lambda < 1/2$, the stability properties of the process are not yet understood.
- ▶ for $\lambda = 0$ process diverges.

Concepts in Proof of Stationarity

- ▶ For $1/2 \leq \lambda \leq 1$ show that $\{W_t\}$ is weak Feller, then bounded in probability. Meyn and Tweedie (1993) used to establish existence of a stationary distribution.
- ▶ For $\lambda = 1$ first show that $\{W_t\}$ satisfies Doeblin's condition and is strongly aperiodic and using Meyn and Tweedie (1993) show $\{W_t\}$ has a unique stationary distribution and is uniformly ergodic.

Estimation

Let λ be fixed in

$$e_t(\delta) = (Y_t - \mu_t)/\mu_t^\lambda.$$

The log-likelihood is

$$L(\delta) = \sum_{t=1}^n \{ Y_t W_t(\delta) - e^{W_t(\delta)} \}$$

where

$$W_t(\delta) = x_t^T \beta + \sum_{i=1}^{\infty} \gamma_i(\tau) e_{t-i}(\delta)$$

and $\delta = (\beta^T, \tau^T)^T$, $\tau = (\phi^T, \theta^T)^T$.

CLT for MLE: simple model

For the constant mean regression with MA(1) lag structure and $\lambda = 1$, the asymptotic distribution of the maximum likelihood estimates has been established (Davis et al. 2005) to be $N(0, V^{-1})$, where

$$V = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n e^{W_t(\delta_0)} \frac{\partial W_t(\delta_0)}{\partial \delta} \frac{\partial W_t(\delta_0)}{\partial \delta^T}.$$

- ▶ Argument uses a linearised form of the loglikelihood which is convex in the parameters. Hence can use a functional central limit theorem.
- ▶ Establishing this uses the ergodicity stated above and a standard martingale central limit theorem applied to the martingale differences

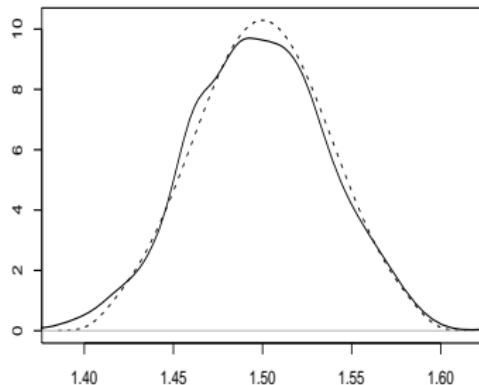
$$\eta_{nt} = \frac{1}{\sqrt{n}} \left(Y_t - e^{W_t(\delta)} \right) \frac{\partial W_t(\delta)}{\partial \delta}$$

CLT for MLE: general model

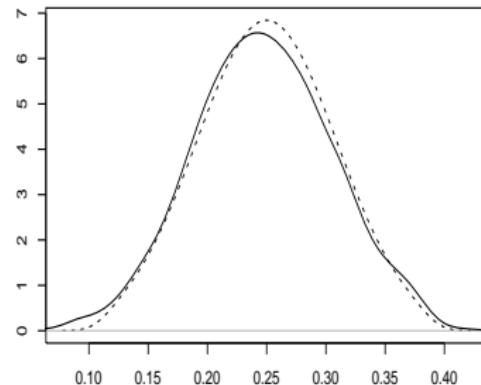
- ▶ A central limit theorem for the maximum likelihood estimators is currently not available for the general model.
- ▶ Simulation results [Davis et al (1999, 2003, 2005), Lu (2002)] for limited models support the supposition that the estimates $\hat{\delta}$ have a multivariate normal distribution for large samples.

Simple Model: CLT simulations

$$W_t = \beta_0 + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}, \quad \lambda = 1, \quad n = 250.$$



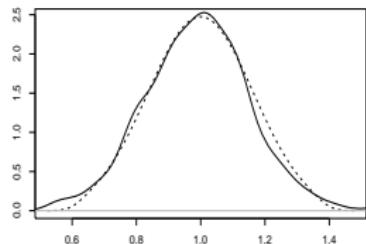
$$\beta_0 = 1.5$$



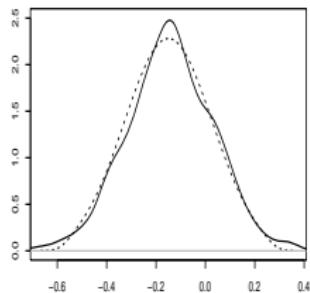
$$\gamma = 0.25$$

Trend Model: CLT Simulation

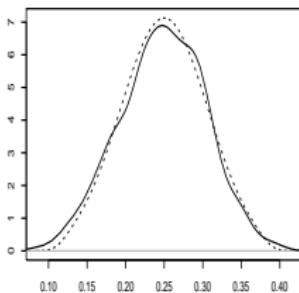
$$W_t = \beta_0 + \beta_1 t/n + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}, \quad \lambda = 1, \quad n = 250.$$



$$\beta_0 = 1$$



$$\beta_1 = -0.15$$



$$\gamma = 0.25$$

Binary case

Street (2000) has established existence of a unique stationary distribution for W_t in the simple model when the observations are binary:

$$W_t = \gamma \frac{n_{t-1} - p_{t-1}}{(p_{t-1}(1 - p_{t-1}))^{0.5}}$$

Future Research Required OD case

- ▶ Conditions for stationarity are needed for more general lag structures.
- ▶ Extend theory to more general regression sequences.

GLARMA Fits to Polio Data - varying λ

Model	$\lambda = 0.5$		Optimal MA(1, 2)	
Covariate:	$\hat{\delta}$	SE	$\hat{\delta}, \hat{\lambda}$	SE
Intercept	0.130	0.114	0.082	0.123
Trend $\times 10^{-3}$	-3.93	2.18	-3.98	2.73
$\cos(2\pi t/12)$	-0.099	0.118	-0.002	0.145
$\sin(4\pi t/12)$	-0.531	0.141	-0.518	0.178
$\cos(2\pi t/12)$	0.211	0.117	0.267	0.110
$\sin(4\pi t/12)$	-0.393	0.116	-0.271	0.112
θ_1	0.218	0.056	0.284	0.058
θ_2	0.127	0.046	0.228	0.051
θ_5	0.087	0.043	-	-
λ	0.5	-	1.141	0.194
$I(\hat{\beta}, \hat{\theta}, \hat{\lambda})$	-118.9		-111.55	

Table : Model estimates for Polio data with different λ values

Comparison: Polio Data Trend

Study	Trend ($\hat{\beta}$)	s.e.($\hat{\beta}$)	Z-statistic
MCNR Kuk & Cheng (1997)	-3.79	2.95	-1.28
MLE by IS Davis & R-Yam (2005)	-3.74	2.82	-1.33
Approx MLE Davis & R-Yam (2005)	-3.81	2.77	-1.38
GLARMA ($\lambda = 0.5$)	-3.93	2.18	-1.80
GLARMA ($\hat{\lambda} = 1.14$)	-3.98	2.73	-1.46
GLARMA Negative Binomial ($\hat{\kappa} = 2.28$)	-4.24	2.71	-1.56

Alternative fits to Polio Data

We consider two more models:

- ▶ Poisson response, score residuals ($\lambda = 1$, fixed).
- ▶ Negative Binomial, Pearson residuals and score residuals.

Polio Data, Poisson Response, Score Residuals ($\lambda = 1$)

```
## Score Type (GAS) Residuals, Fisher Scoring
data(Polio)
y <- Polio[, 2]
X <- as.matrix(Polio[, 3:8])
Polio_Pois_Score_Glarmamod <- glarma(y, X, thetaLags = c(1,2,5), type =
                                     residuals = "Score", maxit = 100, grad = 1e-6)
Polio_Pois_Score_Glarmamod
summary(Polio_Pois_Score_Glarmamod)
```

Polio Data, Poisson Response, Score Residuals ($\lambda = 1$)

Autoregressive Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)	
theta_1	0.30033	0.04429	6.780	1.20e-11	***
theta_2	0.23669	0.04137	5.721	1.06e-08	***
theta_5	0.01824	0.04065	0.449	0.654	

Linear Model Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)	
Intcpt	0.043794	0.119109	0.368	0.71311	
Trend	-3.899761	2.327169	-1.676	0.09379	.
CosAnnual	-0.007278	0.133382	-0.055	0.95649	
SinAnnual	-0.588309	0.147314	-3.994	6.51e-05	***
CosSemiAnnual	0.293552	0.099015	2.965	0.00303	**
SinSemiAnnual	-0.283751	0.110872	-2.559	0.01049	*

Null deviance: 343.0 on 167 degrees of freedom

Residual deviance: 277.4 on 159 degrees of freedom

AIC: 522.6663

Polio Data, Poisson Response, Score Residuals ($\lambda = 1$)

```
## Simplify last model by removing \theta_5:  
  
Polio_Pois_Score_Glarmamod <- glarma(y, X, thetaLags = c(1,2),  
                                         type = "Poi", method = "FS",  
                                         residuals = "Score",  
                                         maxit = 100, grad = 1e-6)  
Polio_Pois_Score_Glarmamod  
summary(Polio_Pois_Score_Glarmamod)
```

Polio Data, Poisson Response, Score Residuals ($\lambda = 1$)

Autoregressive Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)	
theta_1	0.30181	0.04282	7.048	1.81e-12	***
theta_2	0.23476	0.04032	5.822	5.81e-09	***

Linear Model Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)	
Intcpt	0.04766	0.11725	0.406	0.68438	
Trend	-4.03186	2.29823	-1.754	0.07937	.
CosAnnual	-0.02423	0.13356	-0.181	0.85606	
SinAnnual	-0.58966	0.14879	-3.963	7.4e-05	***
CosSemiAnnual	0.30271	0.09827	3.080	0.00207	**
SinSemiAnnual	-0.28516	0.11003	-2.592	0.00955	**

Null deviance: 343.00 on 167 degrees of freedom

Residual deviance: 278.52 on 160 degrees of freedom

AIC: 520.8685

Polio Data, Poisson Response, Score Residuals ($\lambda = 1$)

```
> # loglikelihood adjusted to match that reported in notes  
> # for topic 9 slide 298  
> logLik(Polio_Pois_Score_Glarmamod)+sum(log(factorial(y)))  
[1] -111.9718  
> logLik(Polio_Pois_Pearson_Glarmamod)+sum(log(factorial(y)))  
[1] -118.8901
```

Polio Data, Negative Binomial Response, Pearson Residuals

```
## Negative Binomial Fits
## MA(1,2), Pearson Residuals, Fisher Scoring
data(Polio)
y <- Polio[, 2]
X <- as.matrix(Polio[, 3:8])
Polio_NB_Pearson_Glarmamod <- glarma(y, X, thetaLags = c(1,2),
                                         type = "NegBin", method = "NR",
                                         residuals = "Pearson",
                                         maxit = 100, grad = 1e-6)
summary(Polio_NB_Pearson_Glarmamod )
```

Polio Data, Neg Bin Response, Pearson Residuals

Negative Binomial Parameter:

	Estimate	Std.Error	z-ratio	Pr(> z)
alpha	2.2812	0.7119	3.204	0.00135 **

Autoregressive Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)
theta_1	0.31931	0.10980	2.908	0.00364 **
theta_2	0.21368	0.09873	2.164	0.03045 *

Linear Model Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)
Intcpt	0.14706	0.13762	1.069	0.28525
Trend	-4.23679	2.70825	-1.564	0.11772
CosAnnual	-0.09402	0.16413	-0.573	0.56676
SinAnnual	-0.53681	0.19184	-2.798	0.00514 **
CosSemiAnnual	0.28269	0.14707	1.922	0.05459 .
SinSemiAnnual	-0.31272	0.14659	-2.133	0.03290 *

Null deviance: 201.22 on 167 degrees of freedom

Residual deviance: 147.58 on 159 degrees of freedom

AIC: 509.527

Polio Data, Neg Bin Response, Score Residuals

Negative Binomial Parameter:

	Estimate	Std.Error	z-ratio	Pr(> z)
alpha	2.868	1.062	2.699	0.00695 **

Autoregressive Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)
theta_1	0.27507	0.06140	4.480	7.47e-06 ***
theta_2	0.23494	0.06374	3.686	0.000228 ***

Linear Model Coefficients:

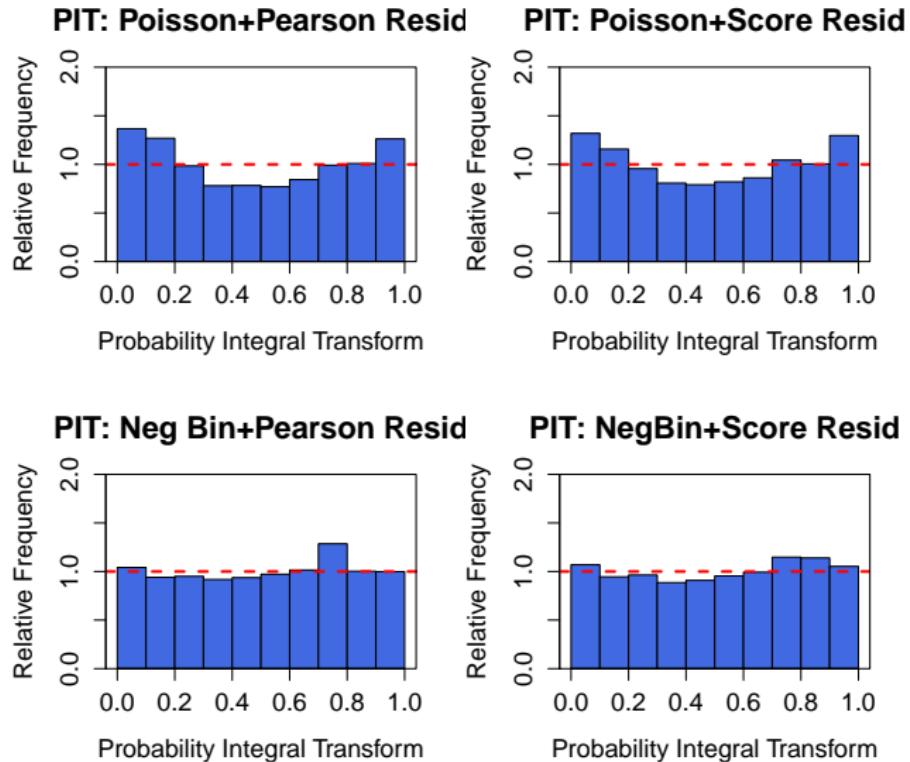
	Estimate	Std.Error	z-ratio	Pr(> z)
Intcpt	0.09312	0.14015	0.664	0.506415
Trend	-4.70140	2.92402	-1.608	0.107867
CosAnnual	-0.04264	0.16396	-0.260	0.794808
SinAnnual	-0.58730	0.17761	-3.307	0.000944 ***
CosSemiAnnual	0.29354	0.12325	2.382	0.017234 *
SinSemiAnnual	-0.25930	0.13546	-1.914	0.055589 .

Null deviance: 201.22 on 167 degrees of freedom

Residual deviance: 255.24 on 159 degrees of freedom

AIC: 504.1576

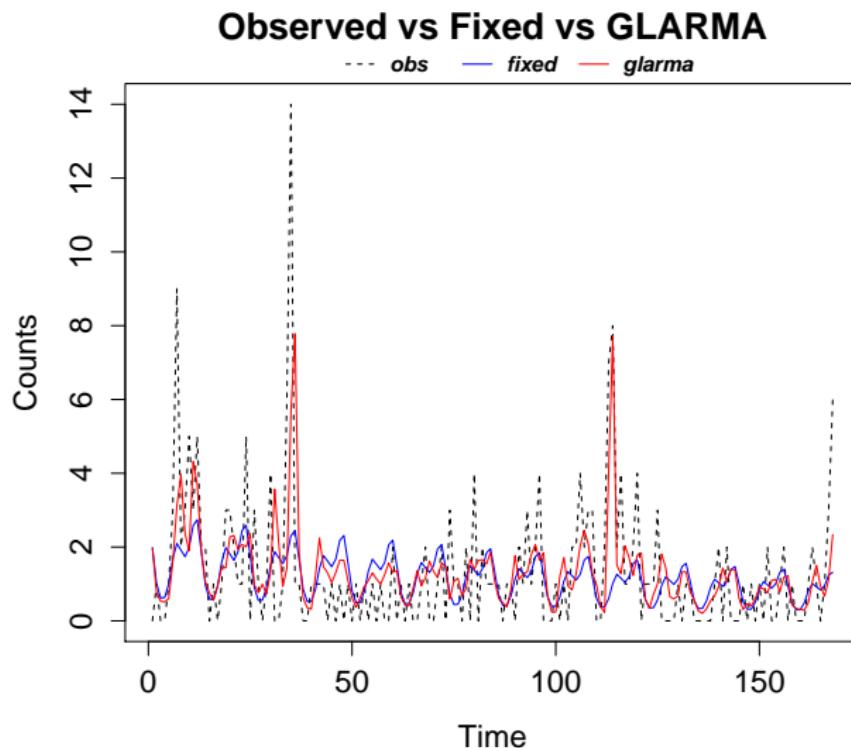
PIT Residuals for 4 models for Polio Data



Comparison of 4 models for Polio Data

Distribution	Residuals	AIC	$\hat{\beta}_{\text{TREND}}$	PIT Plot
Poisson	Pearson	536.7	-3.93 ± 2.14	Bowed
Poisson	Score	520.9	-4.03 ± 2.30	Bowed
Neg Bin	Pearson	509.5	-4.24 ± 2.71	Almost Flat
Neg Bin	Score	504.2	-4.70 ± 2.91	Flat

Polio Data: Fits using Neg Bin with Score Residuals



Using **glarma** for the Boat Race series

Analysis in file `BoatRace_compareKling_with_GLARMA.R`.

Note that the data supplied with the **glarma** package has some errors in the weight differences between winning and losing team.

The file contains a fix up.

We use the data in the file `boatrace.csv`.

For the GLARMA modelling use event time rather than year of event because:

- ▶ Currently the **glarma** package does not handle missing data - tricky to do!
- ▶ An argument could be made that the outcome of previous last races (event) is what matters in influencing the current outcome (weak?).

Boat Race series - finding the best GLARMA model

MA model	AIC	AR model	AIC
MA(1)	198.7	AR(1)	192.3
MA(1,2)	190.6	AR(1,2)	192.7
MA(1,2,3)	192.1	AR(1,2,3)	191.6
AR(1), MA(1)	Failed		
AR(1), MA(2)	189.8		

Boat Race Data, Best GLARMA model

```
Call: glarma(y = Y, X = X, type = "Bin", method = "FS",
    residuals = "Pearson",
    phiLags = c(1), thetaLags = c(2), maxit = 100, grad = 1e-06)
```

Pearson Residuals:

Min	1Q	Median	3Q	Max
-2.3649	-0.7607	0.4468	0.8405	1.8055

Autoregressive Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)
phi_1	0.3349	0.1523	2.199	0.0279 *
theta_2	0.4458	0.1924	2.317	0.0205 *

Linear Model Coefficients:

	Estimate	Std.Error	z-ratio	Pr(> z)
(Intercept)	0.11193	0.26651	0.42	0.67451
Wgt_Diff	0.09351	0.03400	2.75	0.00596 **

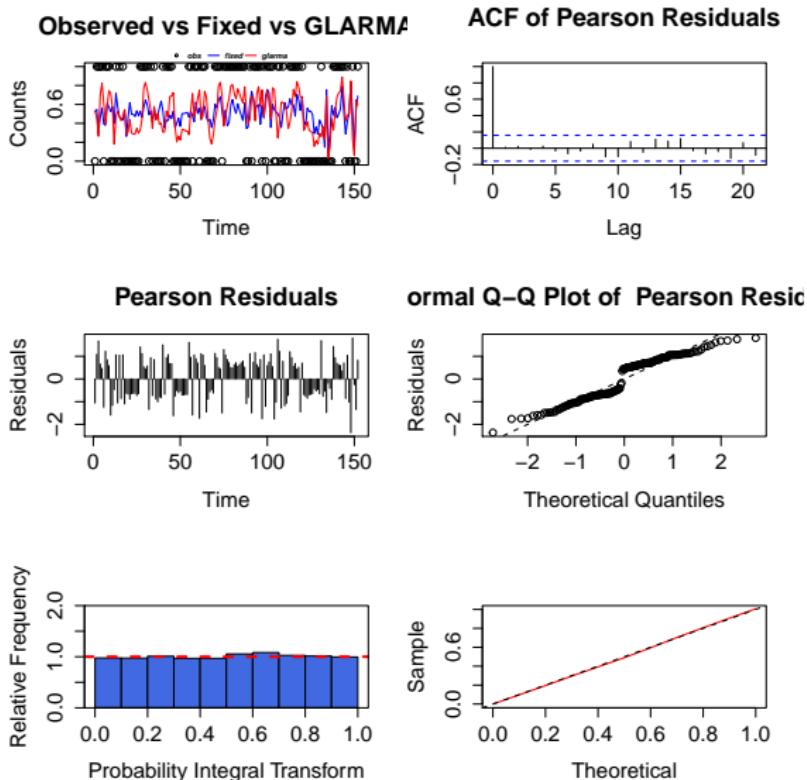
Null deviance: 210.48 on 151 degrees of freedom

Residual deviance: 144.48 on 148 degrees of freedom

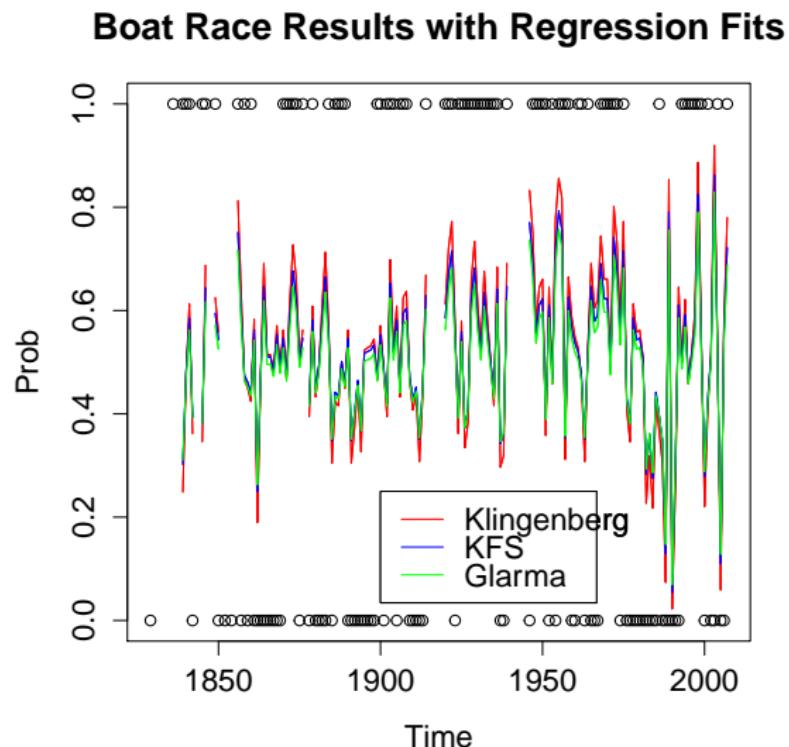
AIC: 189.8285

Number of Fisher Scoring iterations: 18

Boat Race Data: Diagnostics using best GLARMA model

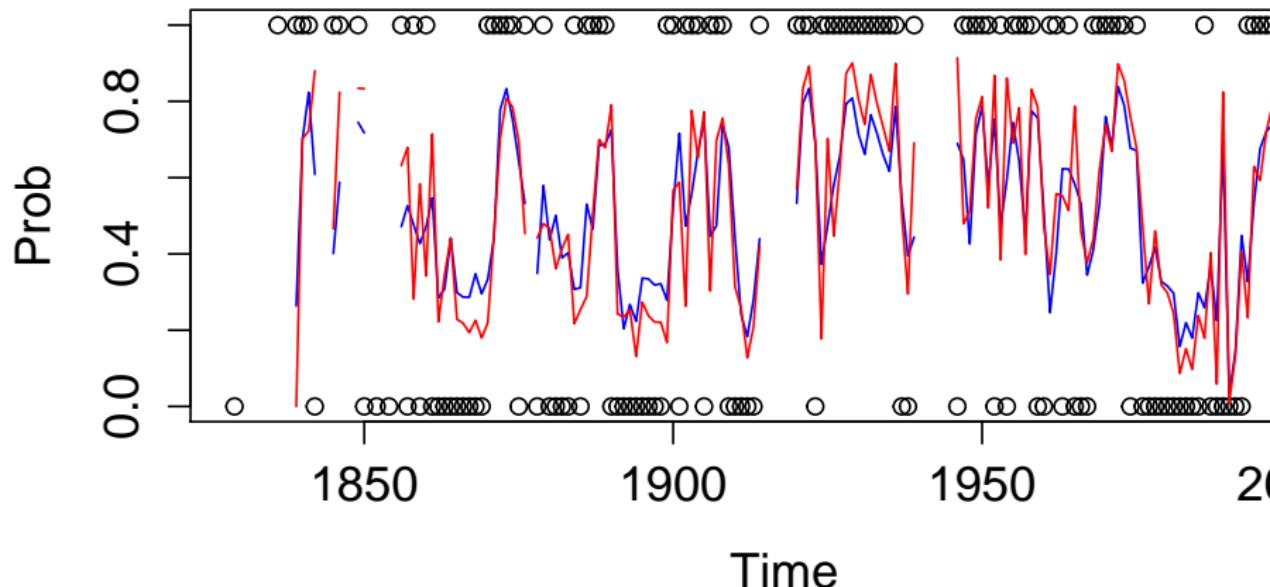


Boat Race Regression fit: Comparing KFS with GLARMA model



Boat Race Data: Filtered Signals: Comparing KFS GLARMA model

Boat Race Results – KFS filtered v glarma fit (blue), KFS filtered signal (red)



References

- DAVIS, R.A., DUNSMUIR, W.T.M., & WANG, Y. (1999). Modelling time series of count data. *Asymptotics, Nonparametrics, and Time Series*, Ed. S. Ghosh, pp. 63–114. New York: Marcel Dekker.
- DAVIS, R.A., DUNSMUIR, W.T.M. & WANG, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika* **87**, 491–506.
- STRETT, S. B. (2000). Some observation driven models for time series. PhD Dissertation, Department of Statistics, Colorado State University.
- LU, H. (2002). Observation driven and parameter driven models in time series of binomial counts, M.Sc. project report, Division of Biostatistics, University of Minnesota.
- DAVIS, R.A., DUNSMUIR, W.T.M. & STREETT, S. (2003). Observation-driven models for Poisson counts. *Biometrika* **90**, 4, 777–790.
- DAVIS, R.A., DUNSMUIR, W.T.M. & STREETT, S. (2005). Maximum likelihood estimation for an observation driven model for Poisson counts. *Methodology, Computing and Applied Probability*, **7**, 149–159.