Nonlinear Time Series Modeling

Part III: Nonlinear and NonGaussian State-Space Models

Richard A. Davis
Colorado State University

(http://www.stat.colostate.edu/~rdavis/lectures)

MaPhySto Workshop

Copenhagen

September 27 — 30, 2004

Part III: Nonlinear and NonGaussian State-Space Models

1. Introduction

1.1 Motivating examples

1.2 Linear state-space models

- Setup
- Random walk + noise
- Local linear trend + seasonality
- Kalman filtering and smoothing
- Other applications (missing values)

1.3 Generalized state-space models

2. Observation-driven models

2.1 GLARMA models for *time series of counts*

- Properties
- Existence and uniqueness of stationary distributions
- Maximum likelihood estimation and asymptotic theory
- Application to asthma data

2.2 GLARMA extensions

- Bernoulli

2.3 Other (BIN)

3. Parameter-driven models

3.1 Estimation

- GLM
- Importance sampling
- Approximation to the likelihood

3.2 Simulation and Application

- Time series of counts
- Stochastic volatility

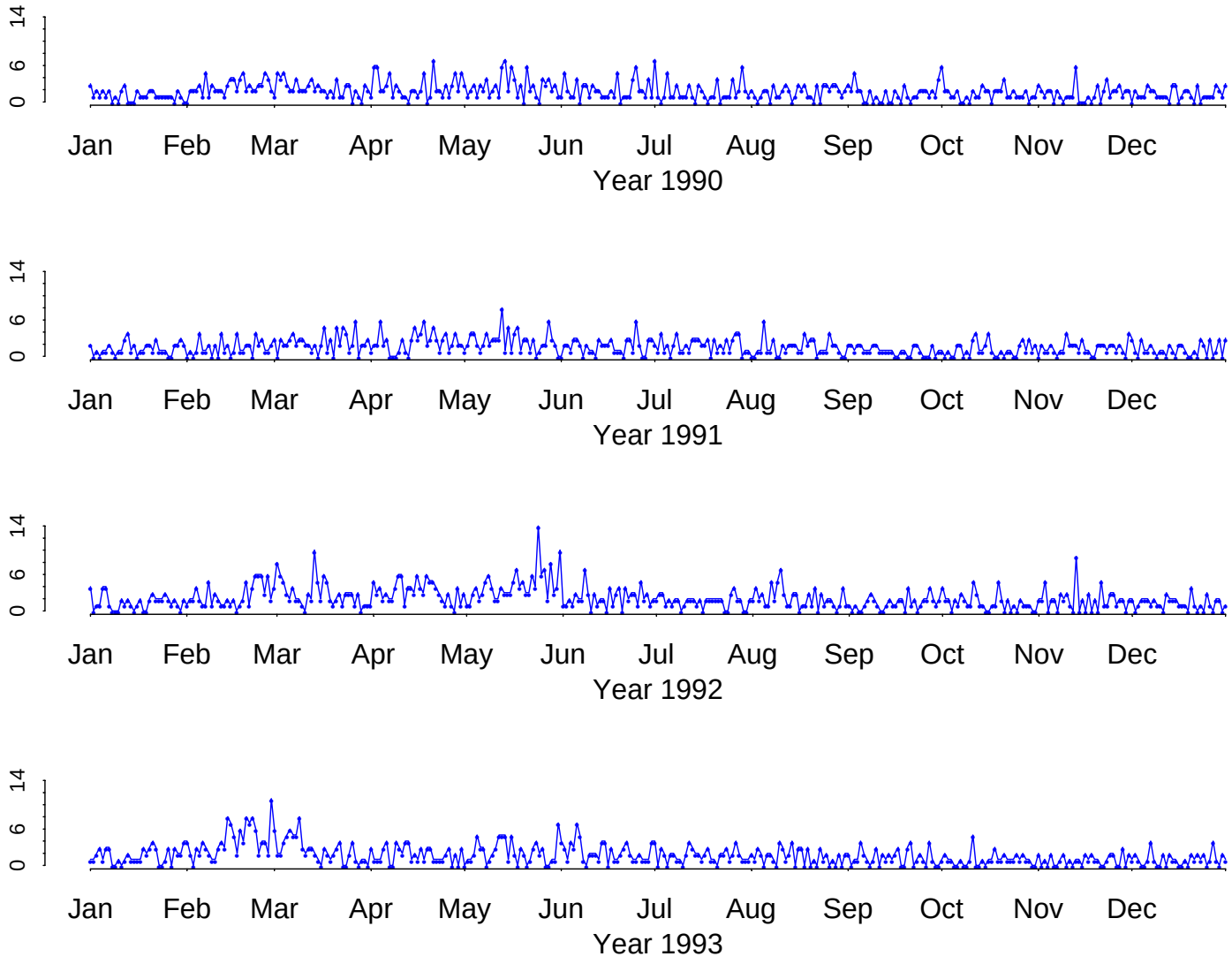
3.3 How good is the posterior approximation?

- Posterior mode vs posterior mean

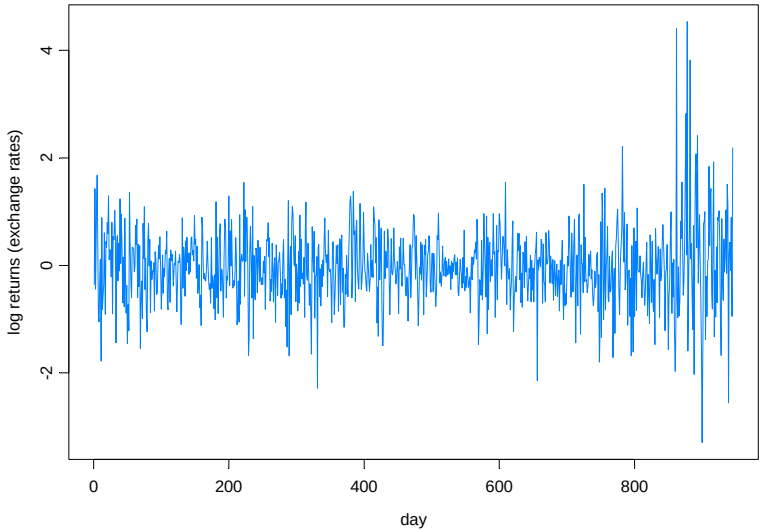
3.4 Application to estimating structural breaks

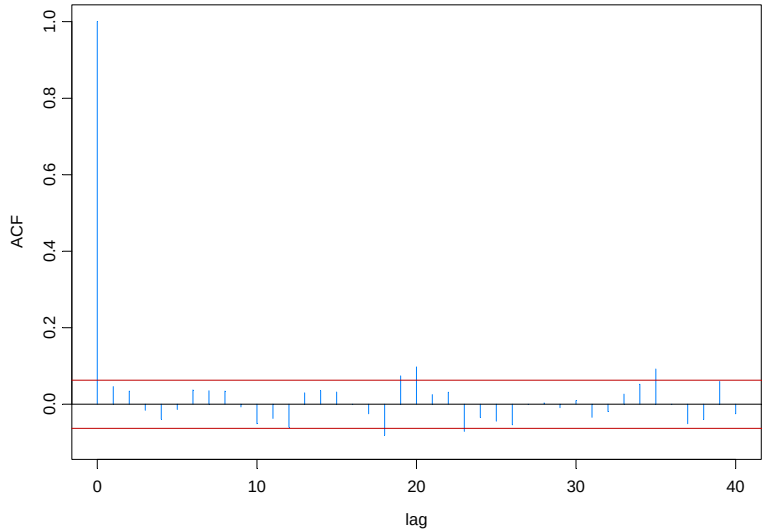
- Poisson model
- Stochastic volatility model

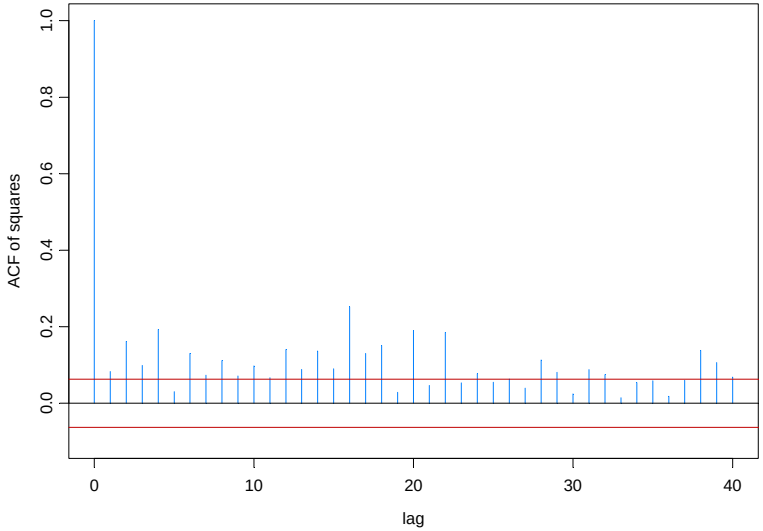
1.1 Motivating Examples: Daily Asthma Presentations (1990:1993)

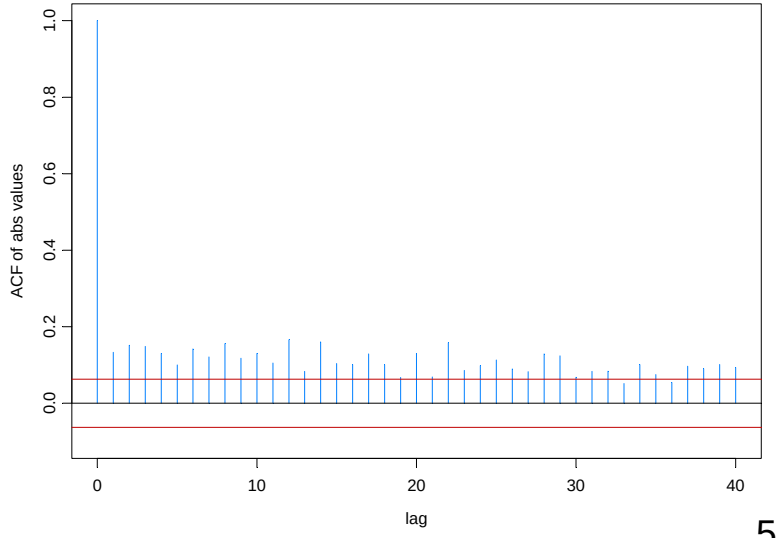


Example: Pound-Dollar Exchange Rates
(Oct 1, 1981 – Jun 28, 1985; Koopman website)









1.2 Linear State-Space Models: Setup

Observation equation:

(1) $\mathbf{Y}_t = G_t \mathbf{X}_t + \mathbf{W}_t, \quad \{\mathbf{W}_t\} \sim WN(0, R_t)$

State equation:

(2) $\mathbf{X}_{t+1} = F_t \mathbf{X}_t + \mathbf{V}_t, \quad \{\mathbf{V}_t\} \sim WN(0, Q_t)$

($\{\mathbf{W}_t\}$ and $\{\mathbf{V}_t\}$ are uncorrelated.)

Def: $\{\mathbf{Y}_t\}$ has a state-space representation if there exists a state-space model for $\{\mathbf{Y}_t\}$ given by (1) and (2).

Example (ARMA(1,1)):

Suppose $\{Y_t\}$ follows the recursions

$$Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim WN(0,\sigma^2).$$

Let $\{X_t\}$ be the AR(1) process defined by

$$X_t = \phi X_{t-1} + Z_t$$

so that

$$\begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \begin{bmatrix} X_{t-2} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ Z_t \end{bmatrix}$$

Setting $\mathbf{X}_{t+1}=[X_{t-1}, X_t]'$, we obtain the state-space representation

$$\mathbf{X}_{t+1} = \begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \mathbf{X}_t + \begin{bmatrix} 0 \\ Z_{t+1} \end{bmatrix} \text{(state equation)}$$

$$Y_t = \begin{bmatrix} \theta & 1 \end{bmatrix} \mathbf{X}_t \qquad \text{(observation equation)}$$

ARMA(1,1) cont (canonical observable representation):

Suppose $\{Y_t\}$ follows the recursions

$$Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim WN(0,\sigma^2).$$

and again let $\{X_t\}$ be the AR(1) process defined by

$$X_{t+1} = \phi X_t + (\phi + \theta) Z_t \quad \text{(state eqn)}$$

Then

$$Y_t = X_t + Z_t \quad \text{(observation eqn)}$$

To see this, note

$$\begin{aligned} Y_t - \phi Y_{t-1} &= X_t - \phi X_{t-1} + Z_t - \phi Z_{t-1} \\ &= (\phi + \theta) Z_{t-1} + Z_t - \phi Z_{t-1} \\ &= Z_t + \theta Z_{t-1} \end{aligned}$$

(Unlike the previous representation this state equation has dimension 1.)

Random walk plus noise

The observed process is modeled as

$$Y_t = X_t + W_t, \quad \{W_t\} \sim WN(0, \sigma_w^2),$$

where the *local level* X_t, follows the random walk

$$X_{t+1} = X_t + V_t, \{V_t\} \sim WN(0, \sigma_v^2).$$

Note that the *differenced data* follows a MA(1) process, i.e.,

$$D_t = Y_t - Y_{t-1} = X_t - X_{t-1} + W_t - W_{t-1} = V_{t-1} + W_t - W_{t-1}$$

has lagged 1-correlation given by

$$\rho(1) = \frac{-\sigma_w^2}{\sigma_v^2 + 2\sigma_w^2} = -\frac{1}{2}$$

if and only if $\sigma_v^2 = 0$.

This latter condition is equivalent to the signal being constant.

Random walk plus noise (cont)

The differenced process D_t can be expressed as the MA(1) process

$$D_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim WN(0,\sigma^2).$$

Matching up lag 1-correlations, we have

$$\rho(1) = \frac{-\sigma_w^2}{\sigma_v^2 + 2\sigma_w^2} = \frac{\theta}{1+\theta^2}$$

So that the signal is constant *if and only if* $\theta = -1$. This is referred to as the *unit root problem*.

Prediction. The *best linear predictor* of Y_{t+1} in terms of $Y_1, \ldots, Y_t$, is given by

$$\hat{Y}_{t+1} = Y_t + (a_t - 1)(Y_t - \hat{Y}_t) = (1 - a_t)\hat{Y}_t + a_t Y_t.$$

For large t, a_t converges to $\theta + 1$. The one-step predictor is given by *exponential smoothing*, which is known to be *optimal for ARIMA(0,1,1)* models!

Local linear trend plus seasonality model

This model takes the form

$$Y_t = M_t + S_t + W_t$$

trend seasonal noise

where

$$\left.\begin{aligned} M_t &= M_{t-1} + B_{t-1} + V_{t1}\,, \\ B_t &= B_{t-1} + V_{t2}\,, \end{aligned}\right\}$$ local linear trend

$$S_t = -S_{t-1} - \cdots - S_{t-11} + V_{t3}$$ seasonal component w/ period=12)

State variable $\mathbf{X}_t=[M_t, B_t, S_t, \ldots, S_{t-10}]'$,

$$\mathbf{X}_{t+1}=F\,\mathbf{X}_t + \mathbf{V}_t$$

$$Y_t = [1\ 0\ 1\ 0, \ldots]\,\mathbf{X}_t + W_t$$

Kalman Filter

Recursive method for producing:

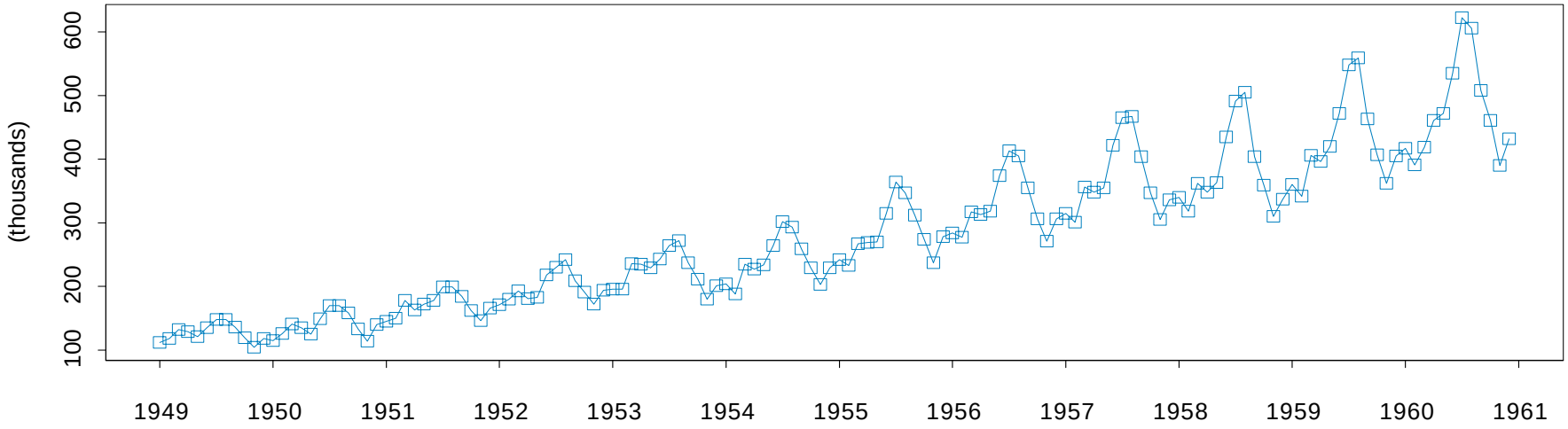
- predictors of the *signal* $P(X_{t+1} \mid Y_1, \ldots, Y_t)$.
- predictors of the *series* $P(Y_{t+1} \mid Y_1, \ldots, Y_t)$.
- *filtered values* of the signal $P(X_{t+1} \mid Y_1, \ldots, Y_{t+1})$.
- *Gaussian likelihood* of $(Y_1, \ldots, Y_n)$.

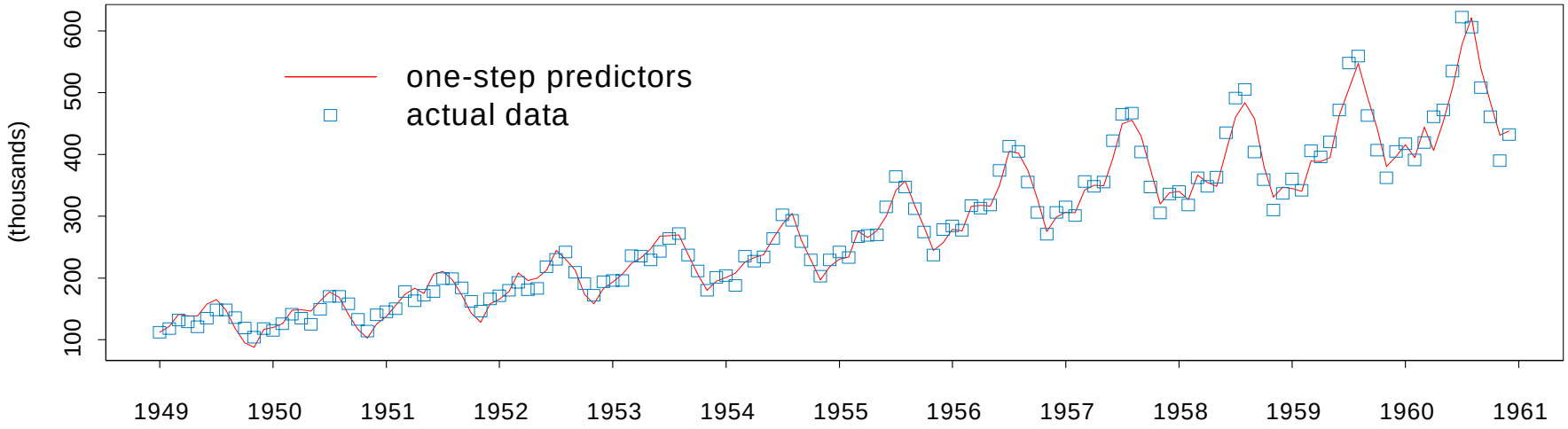
Remarks:

1. Excellent method for calculating best predictors of *missing observations.*

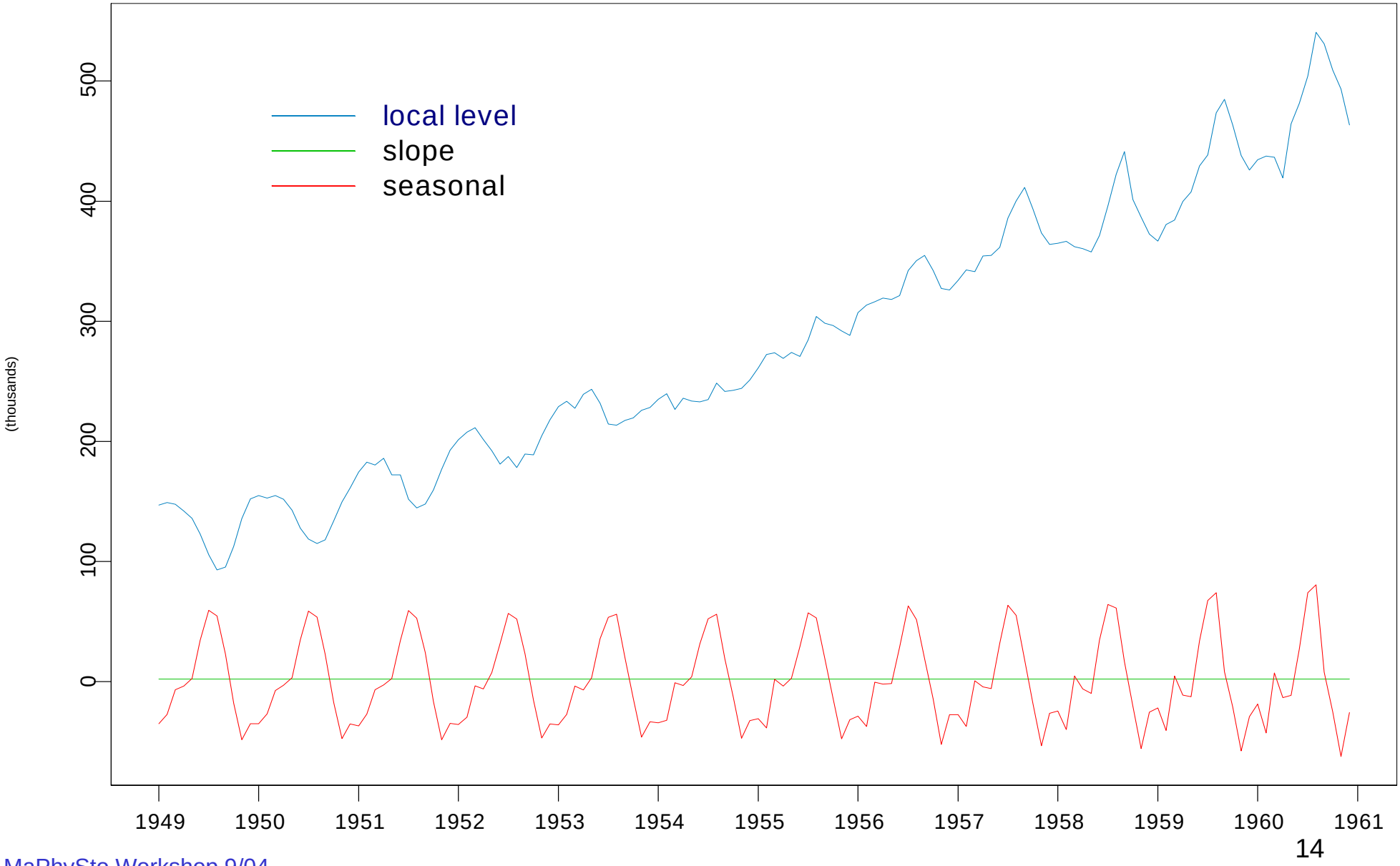
2. Can handle computation of Gaussian likelihood with *missing observations.*

Example (airpass data):

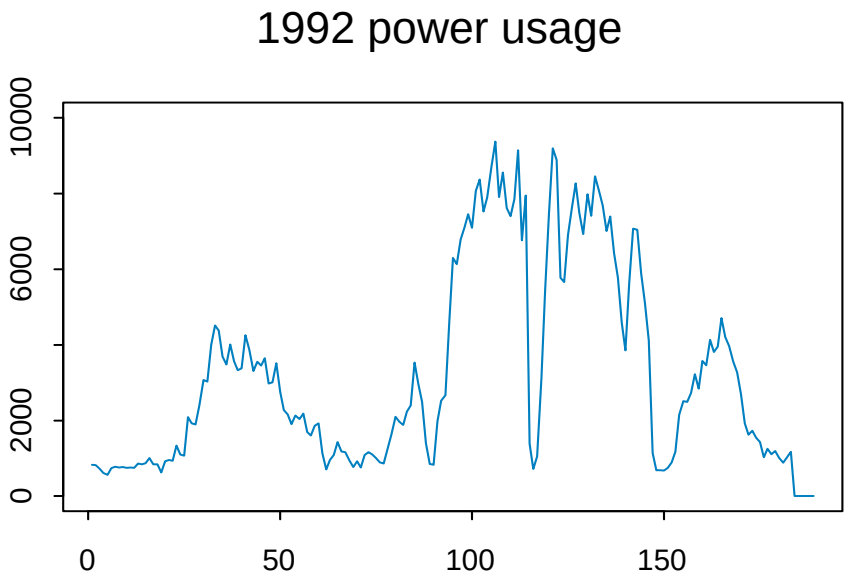


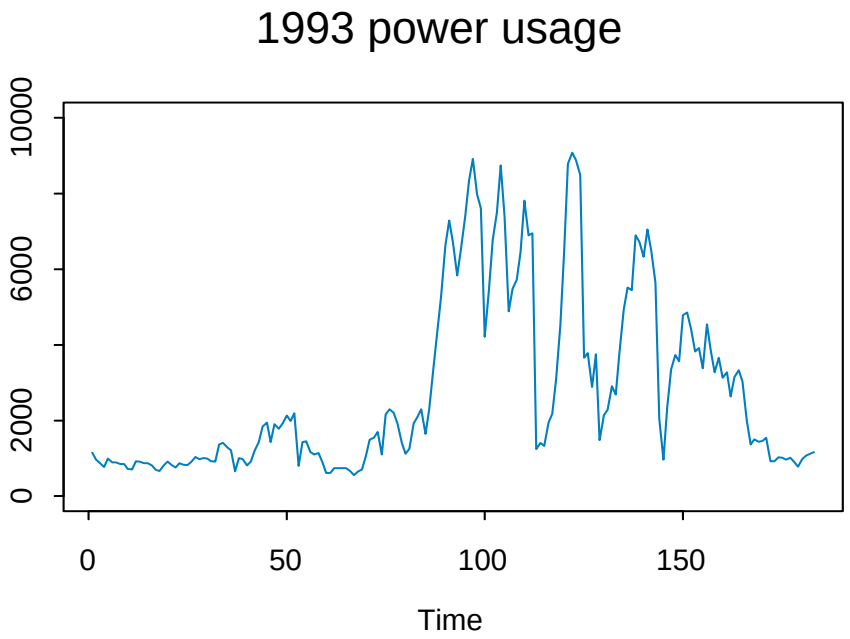


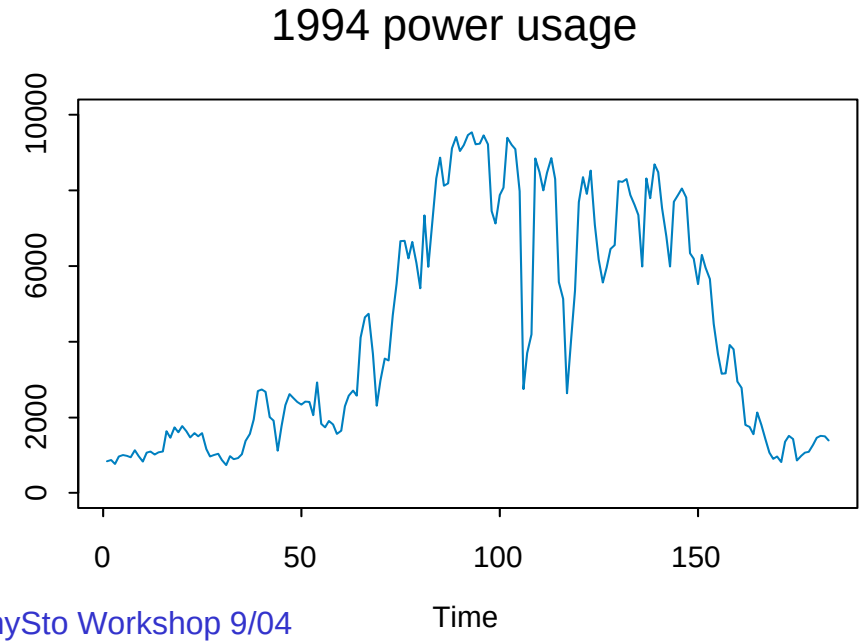
Predicted state components.

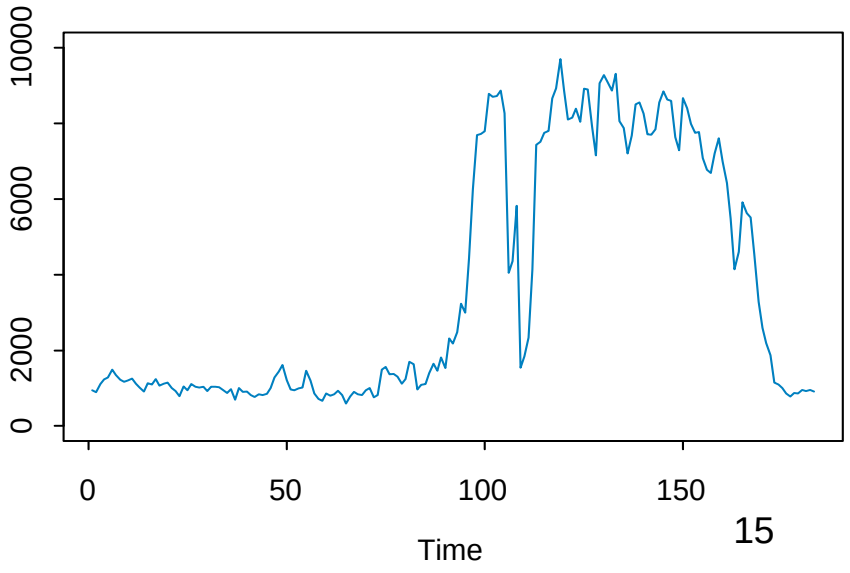


Example (power usage): (with Sarah Streett)









Goal: Forecast maximum power demand 1-2 weeks in advance

Legendre polynomials: orthogonal polyn wrt dx on [0,1]

$f_0(x) = 1,\ f_1(x) = 2x-1,\ f_2(x) = 6x^2-6x+1, \ldots$

Tentative regression model:

$$Y_t = \beta_0\, p_0(t) + \beta_1\, p_1(t) + \beta_2\, p_2(t) + \beta_3\, p_3(t) + \beta_4\, p_5(t) + \beta_5\, p_7(t) + \beta_6\, p_9(t) + \varepsilon_t, \quad t=1, 2, \ldots, 183,$$

where

$$p_j(t) = \frac{f_j(t/183)}{\left\| f_j(t/183) \right\|}$$

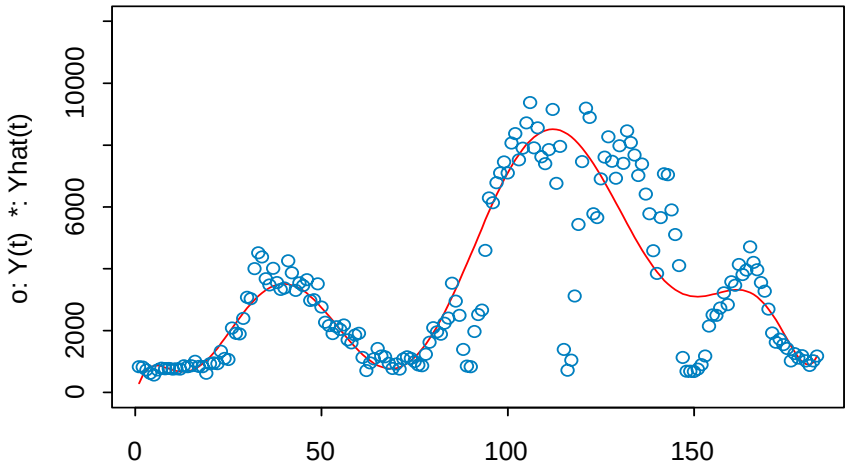
State-space formulation:

$Y_t = \mathbf{p}_t\, \beta_t + W_t$, observation equation

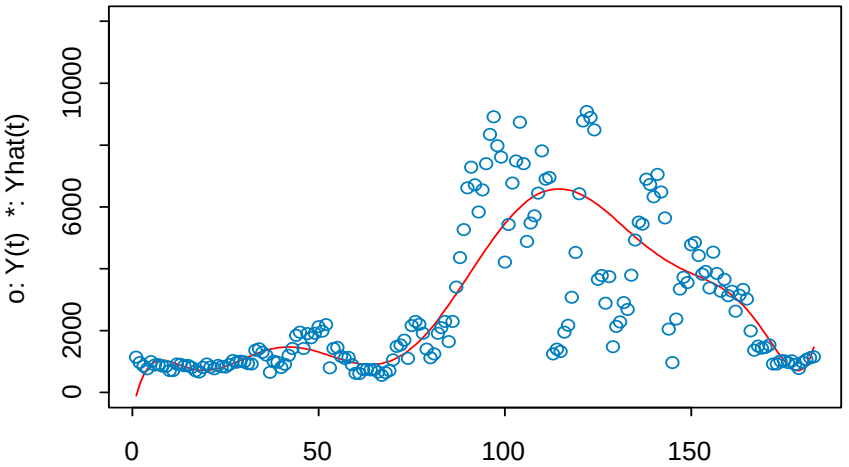
$\beta_t = \beta_{t-1} + \mathbf{V}_t$. state equation

Polynomial fits (using asymmetric loss fcn).

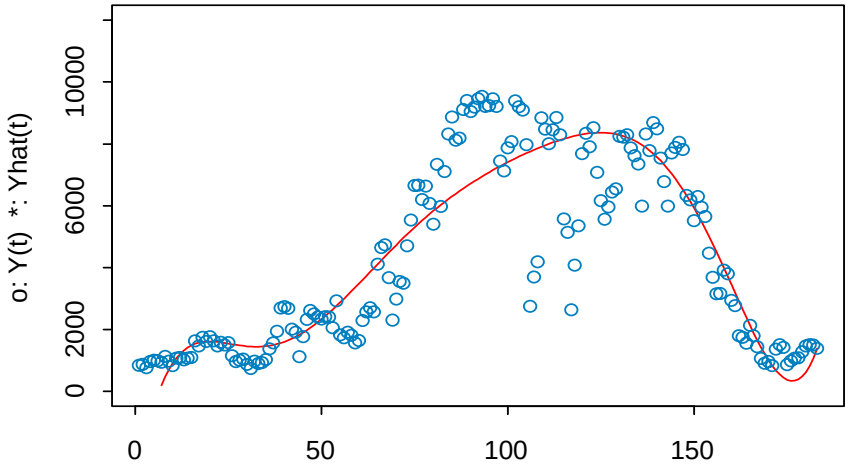
1992 Legendre polynomial fit



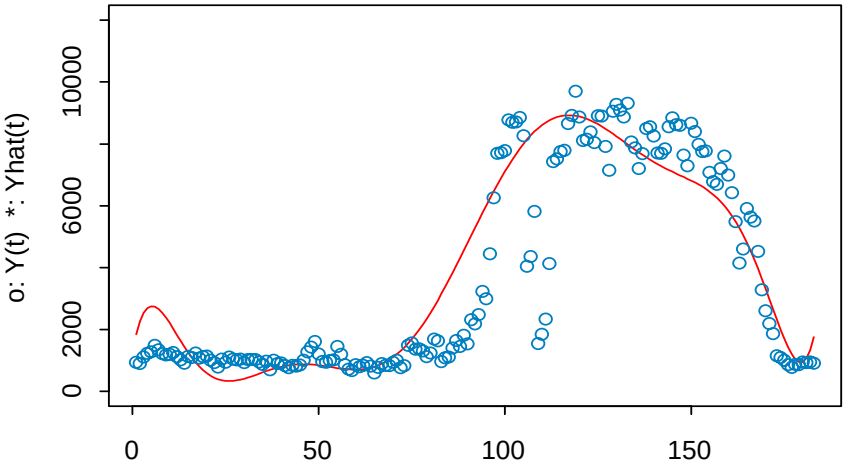
1993 Legendre polynomial fit



1994 Legendre polynomial fit



1995 Legendre polynomial fit



Example: power usage (cont)

The model:

$$
\begin{aligned}
Y_t &= \mathbf{p}_t\beta_t + W_t, \quad W_t \sim \text{IID N}(0, 10000) \\
\beta_{t+1} &= \beta_t + \mathbf{V}_t, \quad \mathbf{V}_t \sim \text{IID N}(\mathbf{0}, \sigma_Q^2(t) \cdot I_{8.8}) \\
t &= 1, \ldots, 183,
\end{aligned}
$$

where :

$$
\begin{aligned}
Y_t &= \text{the maximum daily power usage in kw} \\
\mathbf{p}_t &= \left[\begin{array}{cccccccc} p_0(t) & p_1(t) & p_2(t) & p_3(t) & p_5(t) & p_7(t) & p_9(t) & eta(t) \end{array} \right] \\
\beta_t &= \left[\begin{array}{cccccccc} \beta_0(t) & \beta_1(t) & \beta_2(t) & \beta_3(t) & \beta_4(t) & \beta_5(t) & \beta_6(t) & \beta_7(t) \end{array} \right]' \\
eta(t) &= \text{ETACorn(t)}
\end{aligned}
$$

with initial value:

$$
\beta_1 = \left[\begin{array}{cccccccc} 3793 & 1083 & -1167 & -1344 & 637 & -318 & 245 & 268 \end{array} \right]'
$$

Example: power usage (cont)

The model:

$$\begin{aligned} Y_t &= \mathbf{p}_t\boldsymbol{\beta}_t + W_t, \quad W_t \sim \text{IID N}(0, 10000) \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \mathbf{V}_t, \quad \mathbf{V}_t \sim \text{IID N}(\mathbf{0}, \sigma_Q^2(t) \cdot I_{8\cdot 8}) \end{aligned}$$

Also, the variance is modeled as

$$\begin{aligned} \sigma_Q^2(t) &= 100 \exp^{-3.014 \cdot h(t)} \\ h(t) &= 0.6R(t) + 0.4h(t-1), \end{aligned}$$

where *R(t)* is the amount of rainfall on day *t*. If there is a large rainfall on a given day, then only small changes in the coefficients are allowed for the next several days.

Example: power usage (cont)

Kalman recursions and prediction:

$$\begin{aligned} Y_t &= \mathbf{p}_t\boldsymbol{\beta}_t + W_t, \quad W_t \sim \text{IID N}(0, 10000) \\ \boldsymbol{\beta}_{t+1} &= \boldsymbol{\beta}_t + \mathbf{V}_t, \quad \mathbf{V}_t \sim \text{IID N}(\mathbf{0}, \sigma_Q^2(t) \cdot I_{8\cdot 8}) \end{aligned}$$

Under this modeling scheme, the standard Kalman recursions, which produce best MSE predictors, simplify as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{t+1} &= \hat{\boldsymbol{\beta}}_t + \Theta_t\Delta_t^{-1}(Y_t - \mathbf{p}_t\hat{\boldsymbol{\beta}}_t) \\ \Omega_{t+1} &= \Omega_t + Q - \Theta_t\Delta_t^{-1}\Theta_t' \end{aligned}$$

where

$$\begin{aligned} \Delta_t &= \mathbf{p}_t\Omega_t\mathbf{p}_t' + \sigma_w^2, \text{ and} \\ \Theta_t &= \Omega_t\mathbf{p}_t'. \end{aligned}$$

The h-step predictors are then given by

$$\begin{aligned} P_t\boldsymbol{\beta}_{t+h} &= P_t\boldsymbol{\beta}_{t+1} = \hat{\boldsymbol{\beta}}_{t+1} \text{and} \\ P_tY_{t+h} &= \mathbf{p}_{t+h}\hat{\boldsymbol{\beta}}_{t+1}. \end{aligned}$$

Example: power usage (cont)

Robust Kalman filter (Cipra and Romera 1991):

Instead of using least squares, we use a variation of the Huber influence function defined by

$$\psi_H(z) = \begin{cases} z, & for\ |z| \le c, \\ c\,\mathrm{sgn(z)}, & for\ |z| > c. \end{cases}$$

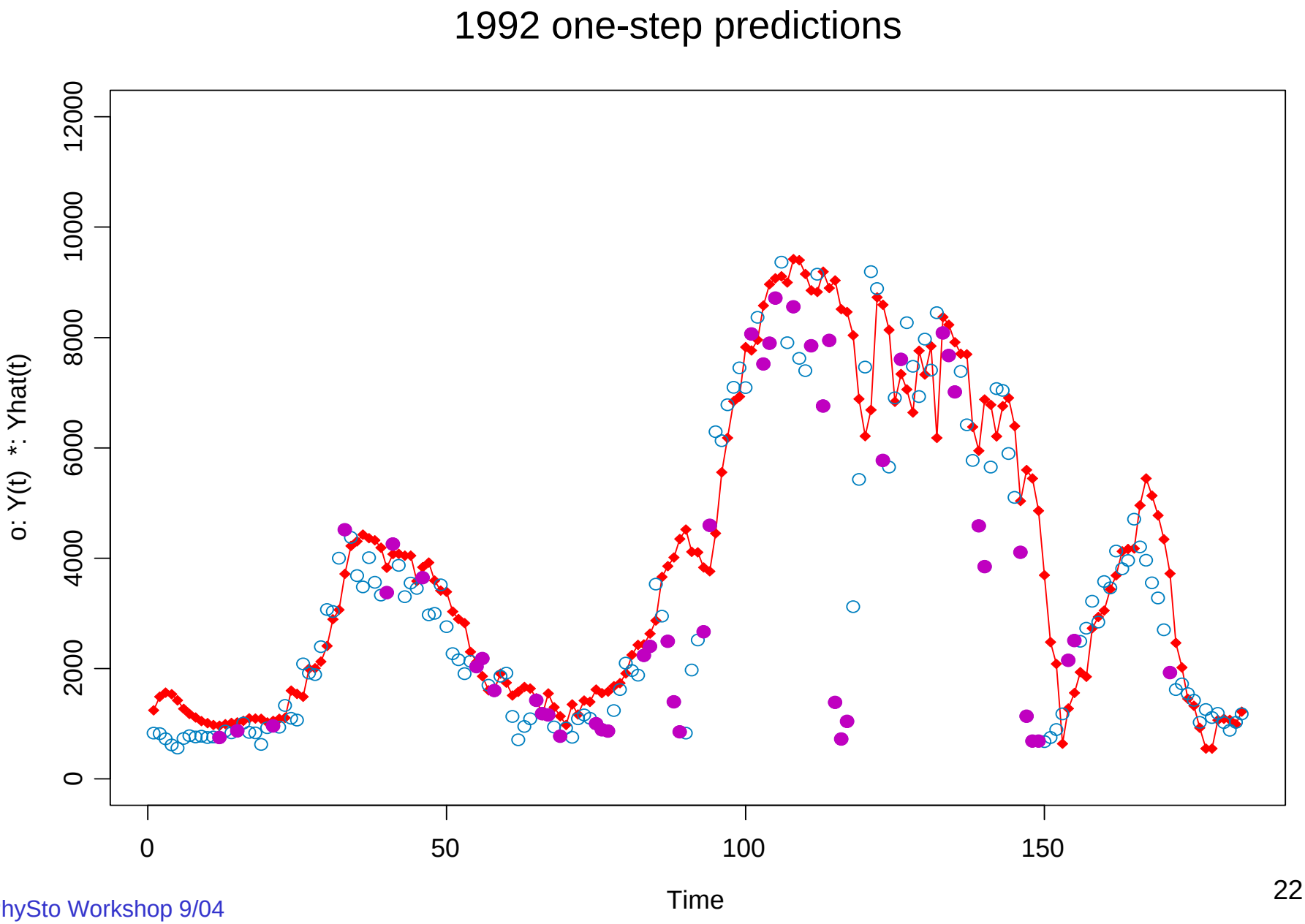
This function was not completely satisfactory. We were led to consider an asymmetric loss function which gives more weight to positive residuals than negative ones. The robust recursive formulas for $\hat{\boldsymbol{\beta}}_{t+1}$ are as follows:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{t+1} &= \hat{\boldsymbol{\beta}}_t + \frac{\Omega_t \mathbf{p}_t'}{\mathbf{p}_t \Omega_t \mathbf{p}_t' + \sigma_w^2/c}(Y_t - \mathbf{p}_t \hat{\boldsymbol{\beta}}_t) \\ \Omega_{t+1} &= \Omega_t + Q - \frac{\Theta_t \Theta_t'}{\mathbf{p}_t \Omega_t \mathbf{p}_t' + \sigma_w^2/c}, \end{aligned}$$

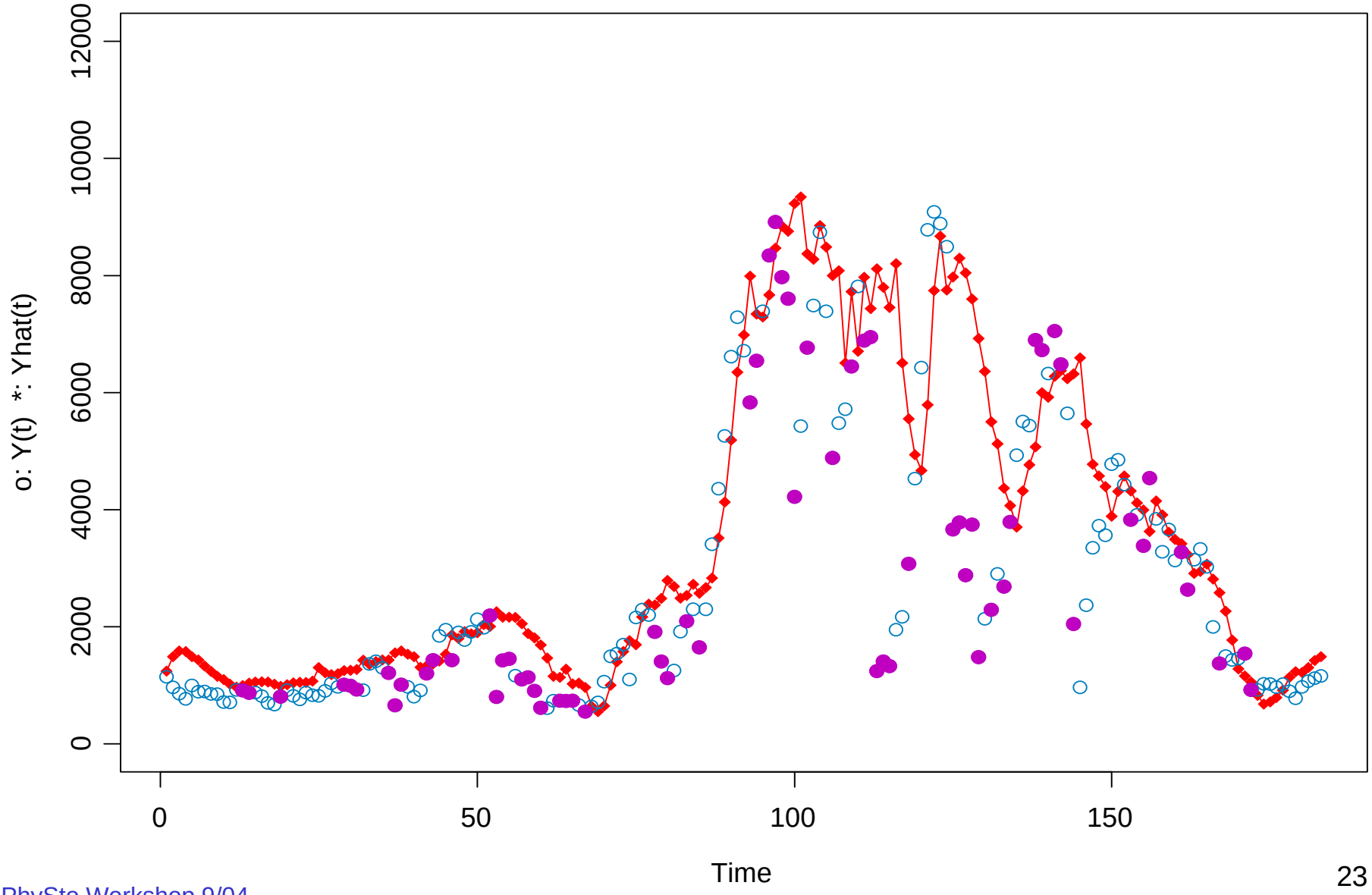
where

$$c = \begin{cases} 10, & \text{if } (Y_t - \hat{Y}_t) > 0, \\ 1, & \text{otherwise.} \end{cases}$$

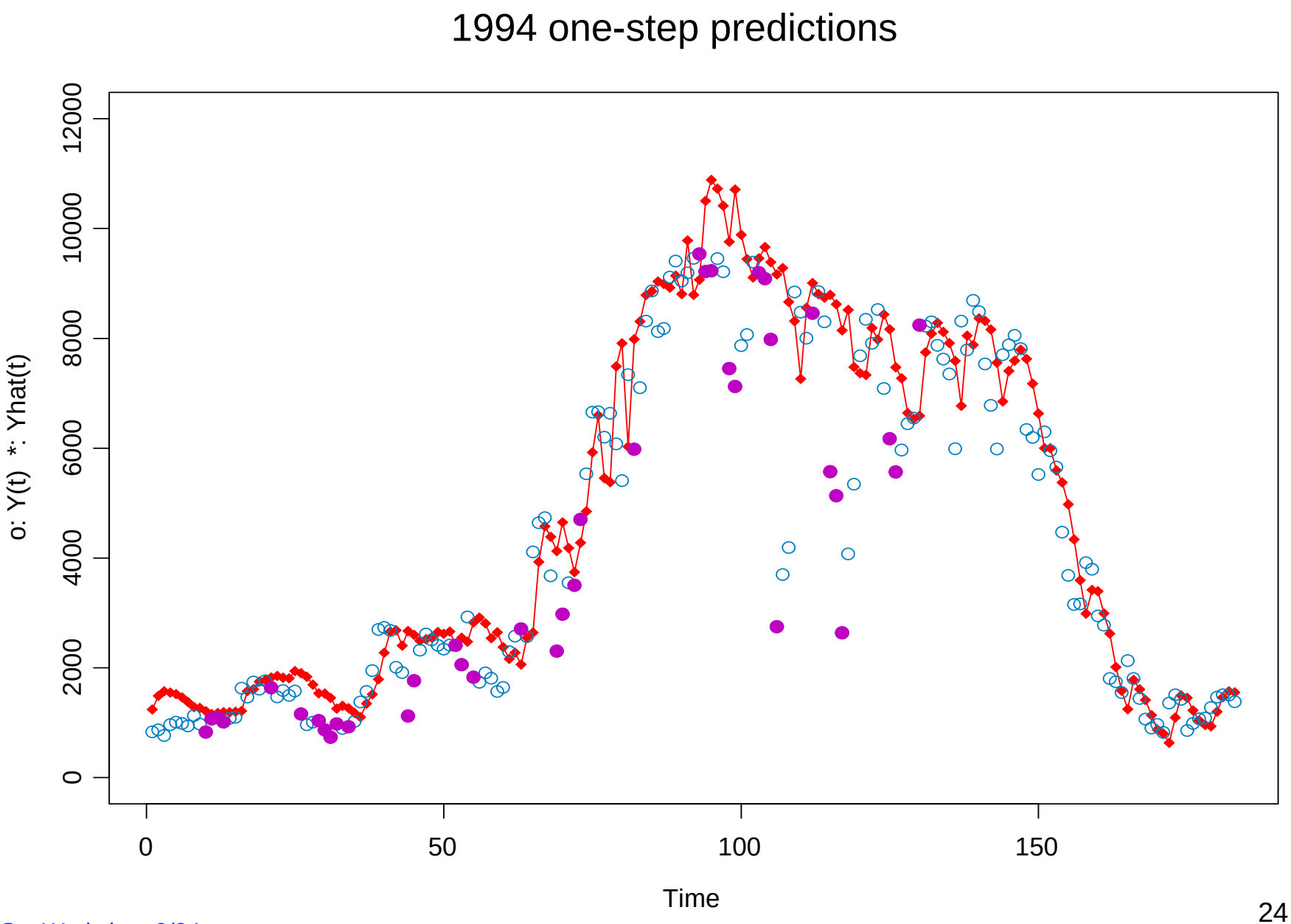
One-step predictions using robustified KF filter:



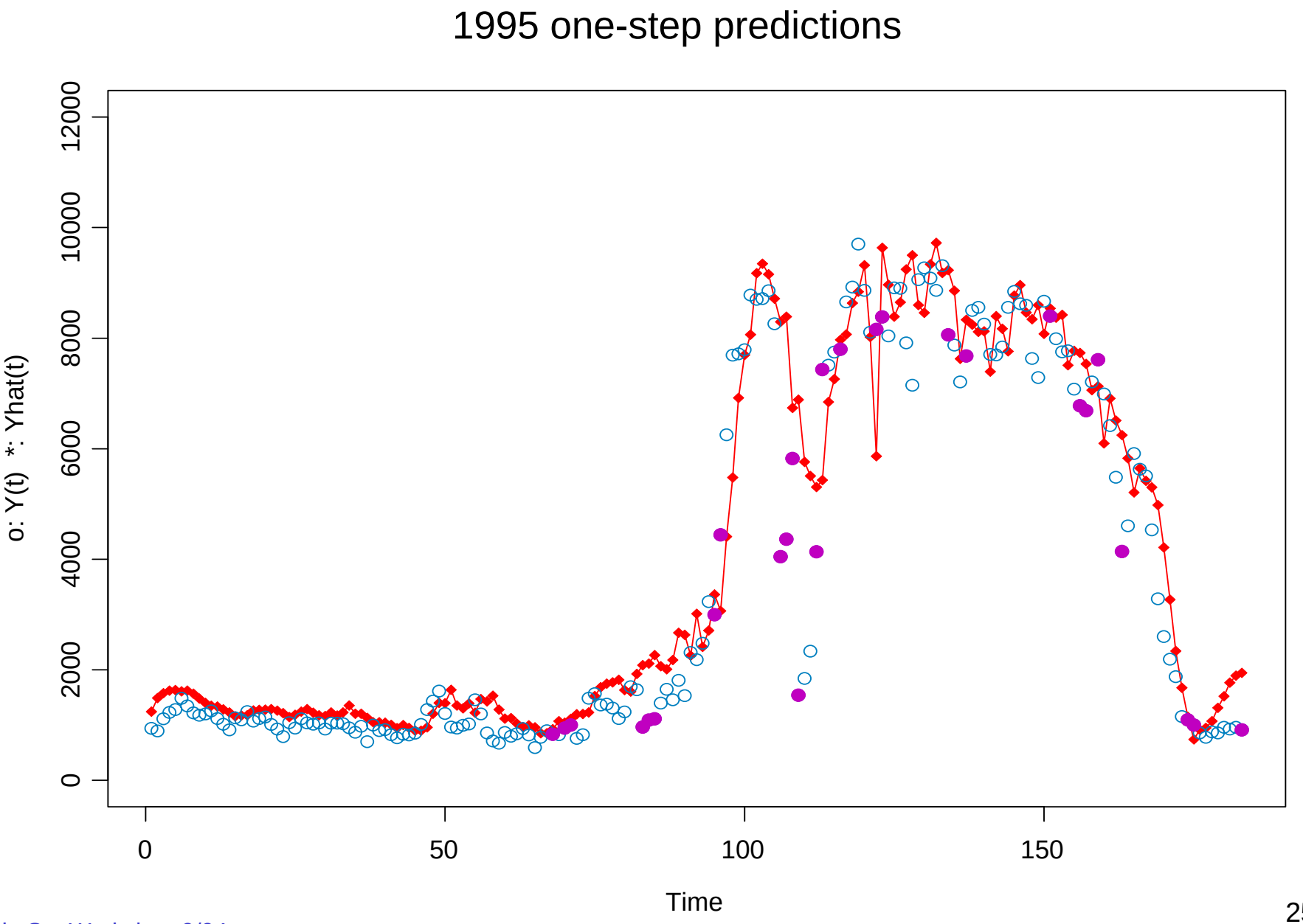
One-step predictions using robustified KF filter:



One-step predictions using robustified KF filter:



One-step predictions using robustified KF filter:



1.3 Generalized State-Space Models

Observations: $y^{(t)} = (y_1, \ldots, y_t)$

States: $\alpha^{(t)} = (\alpha_1, \ldots, \alpha_t)$

Observation equation:

$$p(y_t \mid \alpha_t) := p(y_t \mid \alpha_t, \alpha^{(t-1)}, y^{(t-1)})$$

State equation:

-observation driven

$$p(\alpha_{t+1} \mid y^{(t)}) := p(\alpha_{t+1} \mid \alpha_t, \alpha^{(t-1)}, y^{(t)})$$

-parameter driven

$$p(\alpha_{t+1} \mid \alpha_t) := p(\alpha_{t+1} \mid \alpha_t, \alpha^{(t-1)}, y^{(t)})$$

Example for Time Series of Counts

Count data: $Y_1, \ldots, Y_n$

Regression (explanatory) vector: x_t

Model: Distribution of the Y_t given x_t and a stochastic process α_t are indep Poisson distributed with mean

$$\mu_t = \exp(x_t^T \beta + \alpha_t).$$

The distribution of the stochastic process α_t may depend on a vector of parameters γ.

Note: $\alpha_t = 0$ corresponds to standard Poisson regression model.

Primary objective: Inference about β.

Parameter-Driven Specification

Parameter-driven specification: (Assume $Y_t \mid \mu_t$ is Poisson(μ_t))

$$\log \mu_t = x_t^T \beta + \alpha_t \ ,$$

where $\{\alpha_t\}$ is a stationary Gaussian process.

e.g. (AR(1) process)

$$(\alpha_t + \sigma^2/2) = \phi(\alpha_{t-1} + \sigma^2/2) + \varepsilon_t \ , \quad \{\varepsilon_t\} \sim \text{IID } N(0, \sigma^2(1-\phi^2)).$$

Advantages:

- properties of model (ergodicity and mixing) easy to derive.
- interpretability of regression parameters

$$E(Y_t) = \exp(x_t^T \beta) E\exp(\alpha_t) = \exp(x_t^T \beta), \text{ if } E[\exp(\alpha_t)] = 1.$$

Disadvantages:

- estimation is difficult-likelihood function not easily calculated (MCEM, importance sampling, estimating eqns).
- model building can be laborious
- prediction is more difficult.

Observation-Driven Specification

Observation-driven specification: (Assume $Y_t \mid \mu_t$ is Poisson(μ_t))

$$\log \mu_t = x_t^T \beta + \alpha_t \ ,$$

where ν_t is a function of past observations Y_s , $s < t$.

e.g. $\alpha_t = \gamma_1 Y_{t-1} + \ldots + \gamma_p Y_{t-p}$

Advantages:

- likelihood easy to calculate
- prediction is straightforward (at least one lead-time ahead).

Disadvantages:

- stability behavior, such as stationarity and ergodicty, is difficult to derive.
- $x_t^T \beta$ is not easily interpretable. In the special case above,

$$E(Y_t) = \exp(x_t^T \beta) E\exp(\gamma_1 Y_{t-1} + \ldots + \gamma_p Y_{t-p})$$

Financial Time Series Example-cont

A parameter-driven model for financial data (Taylor `86):

Model:

$$Y_t = \sigma_t Z_t \,,\ \{Z_t\} \sim \text{IID } N(0,1)$$

$$\alpha_t = \gamma + \phi\alpha_{t-1} + \varepsilon_t \,,\quad \{\varepsilon_t\} \sim \text{IID } N(0,\sigma^2),$$

where $\alpha_t = 2 \log \sigma_t$.

The resulting observation and state transition densities are

$$p(y_t \mid \alpha_t) = n(y_t\,;\, 0,\, \exp(\alpha_t))$$

$$p(\alpha_{t+1} \mid \alpha_t) = n(\alpha_{t+1}\,;\, \gamma + \phi\alpha_t \,,\, \sigma^2)$$

Properties:

- Martingale difference sequence.
- Stationary.
- Strongly mixing at a geometric rate.
- Estimation can be difficult-cannot calculate likelihood in closed form.

Financial Time Series Example

An observation driven model for financial data:

Model (GARCH(p,q)):

$$Y_t = \sigma_t Z_t, \{Z_t\} \sim \text{IID } N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \cdots + \alpha_p Y_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2$$

Special case (ARCH(1)=GARCH(1,0)): The resulting observation and state transition density/equations are

$$p(y_t | \sigma_t) = n(y_t; 0, \sigma_t^2)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2$$

Properties:

- Martingale difference sequence.
- Stationary for $\alpha_1 \in [0, 2e^E)$, E-Euler's constant.
- Strongly mixing at a geometric rate.
- For general ARCH (GARCH), properties are difficult to establish.

2 Observation-Driven Models

2.1 GLARMA models for TS models of counts

- Properties
- Existence and uniqueness of stationary distributions
- Maximum likelihood estimation and asymptotic theory
- Application to asthma data

2.2 GLARMA extensions

- Bernoulli case

2.3 Other

- Bin

2.1 Generalized Linear ARMA (GLARMA) Model for Poisson Counts

Model: $Y_t \mid \mu_t$ is Poisson(μ_t) with $\log \mu_t = x_t^T \beta + \alpha_t$

Two components in the specification of α_t (see also Shephard (1994)).

1. Uncorrelated (martingale difference sequence)

For $\lambda > 0$, define

$$e_t = (Y_t - \mu_t) / \mu_t^{\lambda}$$

(Specification of λ will be described later.)

2. Form a linear process driven by the MGD sequence $\{e_t\}$

where

$$\log \mu_t = x_t^T \beta + \alpha_t,$$

$$\alpha_t = \sum_{i=1}^{\infty} \psi_i e_{t-i}.$$

Since the conditional mean μ_t is based on the whole past, the model is no longer Markov.

Properties of the Model

$$e_t = (Y_t - \mu_t)/\mu_t^{\lambda}, \quad \log\mu_t = x_t^T\beta + \alpha_t, \; \alpha_t = \sum_{i=1}^{\infty}\psi_i e_{t-i}.$$

1. $\{e_t\}$ is a MG difference sequence $E(e_t \mid F_{t-1}) = 0$

2. $\{e_t\}$ is an uncorrelated sequence (follows from 1)

3. $E(e_t^2) = E(\mu_t^{1-2\lambda})$

   $= 1$ if $\lambda = .5$

4. Set,

$$W_t = \log\mu_t = x_t^T\beta + \alpha_t,$$

so that

$$E(W_t) = x_t^T\beta \qquad \text{and} \qquad \mathrm{Var}(W_t) = \sum_{i=1}^{\infty}\psi_i^2 E(\mu_{t-i}^{1-2\lambda})$$

$$= \sum_{i=1}^{\infty}\psi_i^2 \quad (\text{if } \lambda = .5)$$

Properties continued

5. $\mathrm{Cov}(W_t, W_{t+h}) = \sum_{i=1}^{\infty} \psi_i \psi_{i+h} E(\mu_{t-i}^{1-2\lambda})$

It follows that $\{W_t\}$ has properties similar to the latent process specification:

$$W_t = x_t^T \beta + \sum_{i=1}^{\infty} \psi_i e_{t-i}$$

which, by using the results for the latent process case and assuming the linear process part is nearly Gaussian, we obtain

$$\begin{aligned} E(e^{W_t}) &= E(e^{x_t^T \beta + \sum_i \psi_i e_{t-i}}) \\ &\approx e^{x_t^T \beta + Var(\alpha_t)/2} \\ &= e^{x_t^T \beta + \sum_{i=1}^{\infty} \psi_i^2 / 2}, \end{aligned}$$

By adjusting the intercept term, $E(\mu_t)$ can be interpreted as $\exp(x_t^T \beta)$.

Properties continued

6. (GLARMA model). Let $\{U_t\}$ be an ARMA process driven by the MGD sequence $\{e_t\}$, i.e.,

$$U_t = \phi_1 U_{t-1} + \ldots + \phi_p U_{t-p} + e_t + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q}$$

Then the best predictor of U_t based on the infinite past is

$$\hat{U}_t = \sum_{i=1}^{\infty} \psi_i e_{t-i}$$

where

$$\sum_{i=1}^{\infty} \psi_i z^i = \phi(z)^{-1}\theta(z) - 1.$$

The model for $\log \mu_t$ is then

$$W_t = x_t^T \beta + Z_t,$$

where

$$Z_t = \hat{U}_t = \phi_1(Z_{t-1} + e_{t-1}) + \cdots + \phi_p(Z_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}.$$

Existence and uniqueness of a stationary distr in the simple case.

Consider the simplest form of the model with $\lambda = 1$, given by

$$W_t = \beta + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-W_{t-1}}.$$

Theorem: The Markov process $\{W_t\}$ has a unique stationary distribution.

Idea of proof:

- State space is $[\beta-\gamma, \infty)$ (if $\gamma > 0$) and $(-\infty, \beta-\gamma]$ (if $\gamma < 0$).
- Satisfies Doeblin's condition:

  There exists a prob measure ν such for some $m > 1$, $\varepsilon > 0$, and $\delta > 0$,

  $\nu(A) > \varepsilon$ implies $P^m(x,A) \geq \delta$ for all x.
- Chain is strongly aperiodic.
- It follows that the chain $\{W_t\}$ is *uniformly ergodic* (Thm 16.0.2 (iv) in Meyn and Tweedie (1993))

Existence of Stationary Distr in Case .5 ≤ λ <1.

Consider the process

$$W_t = \beta + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}.$$

Proposition: The Markov process $\{W_t\}$ has at least one stationary distribution.

Idea of proof:

- $\{W_t\}$ is weak Feller.
- $\{W_t\}$ is bounded in probability on average, i.e., for each x, the sequence

  $\{k^{-1}\sum_{i=1}^{k} P^i(x,\cdot),\ k = 1,2,...,\}$ is tight.
- There exists at least one stationary distribution (Thm 12.0.1 in M&T)

Lemma: If a MC $\{X_t\}$ is weak Feller and $\{P(x, \bullet),\ x \in X\}$ is tight, then $\{X_t\}$ is bounded in probability on average and hence has a stationary distribution.

Note: For our case, we can show tightness of $\{P(x, \bullet),\ x \in X\}$ using a Markov style inequality.

Uniqueness of Stationary Distr in Case .5 ≤ λ <1?

Theorem (M&T `93): If the Markov process $\{X_t\}$ is an *e-chain* which is bounded in probability on average, then there exists a unique stationary distribution if and only if there exists a *reachable point* x^*.

For the process $W_t = \beta + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-\lambda W_{t-1}}$, we have

- $\{W_t\}$ is bounded in probability uniformly over the state space.
- $\{W_t\}$ has a reachable point x^* that is a zero of the equation

$$0 = x^* + \gamma \exp\{(1-\lambda)\, x^*\}$$

- e-chain?

Reachable point: x^* is a reachable point if for every open set O containing x^*,

$$\sum_{n=1}^{\infty} P^n(x, O) > 0 \quad \text{for all } x.$$

e-chain: For every continuous f with compact support, the sequence of functions $\{P^n f\,,\ n = 1,\ldots\}$ is equicontinuous, on compact sets.

Estimation for Poisson GLARMA

Let $\delta = (\beta^T, \gamma^T)^T$ be the parameter vector for the model (γ corresponds to the parameters in the linear process part).

Model: $Y_t \mid \mu_t$ is Poisson(μ_t)

$$\log \mu_t = x_t^T \beta + \alpha_t,$$

$$\alpha_t = \sum_{i=1}^{\infty} \psi_i e_{t-i}.$$

Log-likelihood:

$$L(\delta) = \sum_{t=1}^{n} (Y_t W_t(\delta) - e^{W_t(\delta)}),$$

where

$$W_t(\delta) = x_t \beta + \sum_{i=1}^{\infty} \psi_i(\delta) e_{t-i}.$$

First and second derivatives of the likelihood can easily be computed recursively and Newton-Raphson methods are then implementable. For example,

$$\frac{\partial L(\delta)}{\partial \delta} = \sum_{t=1}^{n} (Y_t - e^{W_t(\delta)}) \frac{\partial W_t(\delta)}{\partial \delta}$$

and the term $\partial W_t(\delta) / \partial \delta$ can be computed recursively.

Asymptotic Results for MLE

Define the array of random variables by

$$\eta_{nt} = n^{-1/2}(Y_t - e^{W_t(\delta)})\frac{\partial W_t(\delta)}{\partial \delta}.$$

Properties of $\{\eta_{nt}\}$:

- $\{\eta_{nt}\}$ is a martingale difference sequence.
- $\sum_{t=1}^{n} E(\eta_{nt}\eta_{nt}^T \mid F_{t-1}) \xrightarrow{P} V(\delta).$
- $\sum_{t=1}^{n} E(\eta_{nt}\eta_{nt}^T I(\mid \eta_{nt} \mid > \varepsilon) \mid F_{t-1}) \xrightarrow{P} 0.$

Using a MG central limit theorem, it "follows" that

$$n^{1/2}(\hat{\delta} - \delta) \xrightarrow{D} N(0, V^{-1}),$$

where $$V = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} e^{W_t(\delta)} \partial W_t(\delta) \partial W_t^T(\delta).$$

Simulation Results

Model 1: $W_t = \beta_0 + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-W_{t-1}},\ n = 500, nreps = 5000$

Parameter	Mean	SD	SD(from like)
$\beta_0 = 1.50$	1.499	0.0263	**0.0265**
$\gamma = 0.25$	0.249	0.0403	**0.0408**
$\beta_0 = 1.50$	1.499	0.0366	**0.0364**
$\gamma = 0.75$	0.750	0.0218	**0.0218**
$\beta_0 = 3.00$	3.000	0.0125	**0.0125**
$\gamma = 0.25$	0.249	0.0431	**0.0430**
$\beta_0 = 3.00$	3.000	0.0175	**0.0174**
$\gamma = 0.75$	0.750	0.0270	**0.0271**

Model 2: $W_t = \beta_0 + \beta_1 t/500 + \gamma(Y_{t-1} - e^{W_{t-1}})e^{-W_{t-1}},\ n = 500, nreps = 5000$

$\beta_0 = 1.00$	1.000	0.0286	**0.0284**
$\beta_1 = 0.50$	0.500	0.0035	**0.0034**
$\gamma = 0.25$	0.248	0.0420	**0.0426**
$\beta_0 = 1.50$	0.998	0.0795	**0.0805**
$\beta_1 = -.15$	-.150	0.0171	**0.0173**
$\gamma = 0.25$	0.247	0.0337	**0.0339**

Application to Sydney Asthma Count Data

Data: $Y_1, \ldots, Y_{1461}$ daily asthma presentations in a Campbelltown hospital.

Preliminary analysis identified.

- no upward or downward trend
- annual cycle modeled by $\cos(2\pi t/365)$, $\sin(2\pi t/365)$
- seasonal effect modeled by

$$P_{ij}(t) = \frac{1}{B(2.5,5)} \left(\frac{t - T_{ij}}{100} \right)^{2.5} \left(1 - \frac{t - T_{ij}}{100} \right)^{5}$$

  where $B(2.5,5)$ is the beta function and T_{ij} is the start of the j^{th} school term in year i.

- day of the week effect modeled by separate indicator variables for Sunday and Monday (increase in admittance on these days compared to Tues-Sat).
- Of the meteorological variables (max/min temp, humidity) and pollution variables (ozone, NO, NO_2), only humidity at lags of 12-20 days and NO_2(max) appear to have an association.

Model for Asthma Data

Trend function.

$x_t^T = (1, S_t, M_t, \cos(2\pi t/365), \sin(2\pi t/365), P_{11}(t), P_{12}(t),$
$P_{21}(t), P_{22}(t), P_{31}(t), P_{32}(t), P_{41}(t), P_{42}(t), H_t, N_t)$

($H_t = \frac{1}{7}\sum_{i=0}^{6} h_{t-12-i}$ and h_t is the residual from an annual cycle fitted to the daily average humidity at 0900 and 1500.)

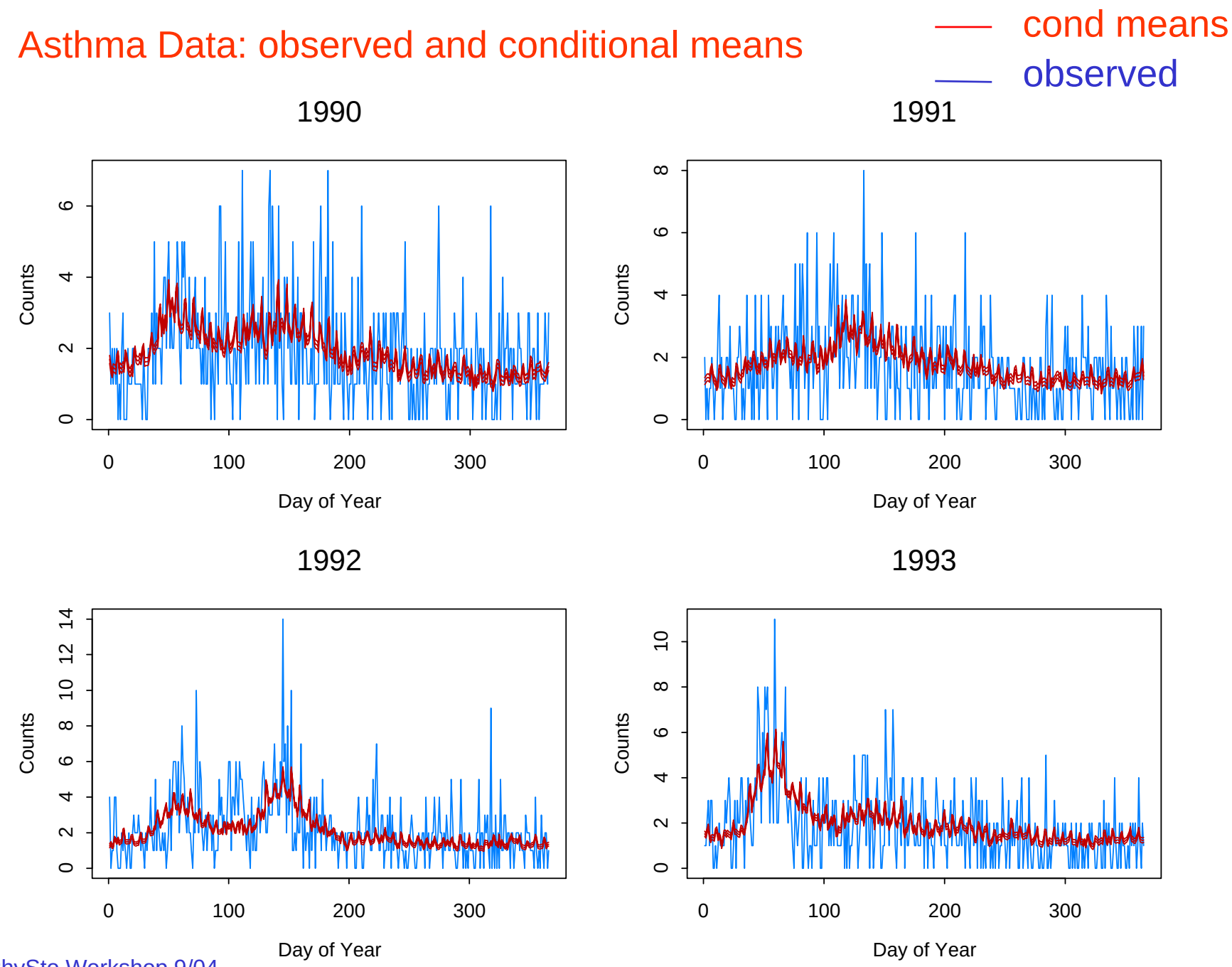
Model for $\{\alpha_t\}$.

MA(7): $\alpha_t = \theta_7 e_{t-7}$

Results for Asthma Data

Term	Est	SE	T-ratio
Intercept	0.583	0.062	9.46
Sunday effect	0.197	0.056	3.53
Monday effect	0.230	0.055	4.20
$\cos(2\pi t/365)$	-0.214	0.039	-5.54
$\sin(2\pi t/365)$	0.176	0.040	4.35
Term 1, 1990	0.200	0.056	3.54
Term 2, 1990	0.132	0.057	2.31
Term 1, 1991	0.087	0.066	1.32
Term 2, 1991	0.172	0.057	2.99
Term 1, 1992	0.254	0.055	4.66
Term 2, 1992	0.308	0.049	6.31
Term 1, 1993	0.439	0.050	8.77
Term 2, 1993	0.116	0.061	1.91
Humidity $H_t/20$	0.169	0.055	3.09
NO_2 max	-0.104	0.033	-3.16
MA, lag 7 θ_7	0.042	0.018	2.32

Asthma Data: observed and conditional means



Summary Remarks for Observation-Driven Models

The observation model for the Poisson counts proposed here has the following properties.

1. Easily interpretable on the *linear predictor scale* and on the scale of the mean μ_t with the regression parameters directly interpretable as the amount by which the mean of the count process at time t will change for a unit change in the regressor variable.

2. An approximately unbiased plot of the μ_t can be generated by

$$\hat{\mu}_t = \exp(\hat{W}_t - .5\sum_{i=1}^{\infty}\hat{\psi}_i^2).$$

3. Is easy to predict with.

4. Provides a mechanism for adjusting the inference about the regression parameter β for a form of serial dependence.

5. Generalizes to ARMA type lag structure.

6. Estimation (approx MLE) is easy to carry out.

2.2 GLARMA Extensions (Binary data)

Binary data: $Y_1, \ldots, Y_n$

Regression (explanatory) variable: x_t

Model: Distribution of the Y_t given x_t and the past is Bernoulli(p_t), i.e.,

$$P(Y_t = 1 | F_{t-1}) = p_t \text{ and } P(Y_t = 0 | F_{t-1}) = 1 - p_t.$$

As before construct a MGD sequence

$$e_t = (Y_t - p_t)/(p_t(1-p_t))^{1/2}$$

and using the logistic link function, the GLARMA model becomes

$$W_t = \log \frac{p_t}{1-p_t} \text{ with } W_t = x_t^T \beta + Z_t,$$

and

$$Z_t = \hat{U}_t = \phi_1(Z_{t-1} + e_{t-1}) + \cdots + \phi_p(Z_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}.$$

A Simple GLARMA Model for Price Activity (R&S)

Model for price change: The price change C_i of the i^{th} transaction has the following components:

- Y_t activity {0,1}
- D_t direction {-1,1}
- S_t size {1, 2, 3, . . .}

Rydberg and Shephard consider a model for these components. An autologistic model is used for Y_t .

Simple GLARMA(0,1) model for price activity: Y_t is a Bernoulli rv representing a price change at the t^{th} transaction. Assume Y_t given F_{t-1} is Bernoulli(p_t), i.e.,

$$P(Y_t = 1 \mid F_{t-1}) = p_t = 1 - P(Y_t = 0 \mid F_{t-1}),$$

where

$$p_t = \frac{e^{\sigma Z_t}}{(1+e^{\sigma Z_t})} \text{ and } Z_t = \frac{Y_{t-1} - p_{t-1}}{\sqrt{p_{t-1}(1-p_{t-1})}} = e_{t-1}.$$

Existence of Stationary Solns for the Simple GLARMA Model

Consider the process

$$Z_t = \frac{Y_{t-1} - p_{t-1}}{\sqrt{p_{t-1}(1-p_{t-1})}},$$

where Y_{t-1} is Bernoulli with parameter $p_{t-1} = e^{\sigma Z_{t-1}}(1+e^{\sigma Z_{t-1}})^{-1}$.

Propostion: The Markov process $\{Z_t\}$ has a unique stationary distribution.

Idea of proof:

- $\{Z_t\}$ is an e-chain.
- $\{Z_t\}$ is bounded in probability on uniformly over the state space
- Possesses a reachable point (x^* is soln to $x+e^{\sigma x/2}=0$)

2.3 BIN Models: A Modeling Framework for Stock Prices (Davis, Rydberg, Shephard, Streett)

Consider the model of a price of an asset at time *t* given by

where
$$p(\mathrm{t}) = p(0) + \sum_{i=1}^{N(t)} Z_i,$$

- *N(t)* is the number of trades up to time *t*
- Z_i is the price change of the i^{th} transaction.

Then for a fixed time period Δ,

$$p_t := p((t+1)\Delta-) - p(t\Delta) = \sum_{i=N(t\Delta)+1}^{N((t+1)\Delta-)} Z_i \ ,$$

denotes the rate of return on the investment during the t^{th} time interval and

$$N_t := N((t+1)\Delta-) - N(t\Delta)$$

denotes the number of trades in $[t\,\Delta, (t+1)\,\Delta)$.

The Bin Model for the Number of Trades

Bin(p,q) model: The distribution of the number of trades N_t in $[t\,\Delta, (t+1)\,\Delta)$, conditional on information up to time $t\,\Delta-$ is Poisson with mean

$$\lambda_t = \alpha + \sum_{j=1}^{p} \gamma_j N_{t-j} + \sum_{j=1}^{q} \delta_j \lambda_{t-j}, \alpha \geq 0, 0 \leq \gamma_j, \delta_j < 1.$$

Proposition: For the Bin(1,1) model,

$$\lambda_t = \alpha + \gamma N_{t-1} + \delta \lambda_{t-1},$$

there exists a unique stationary solution.

Idea of proof:

- $\{\lambda_t\}$ is an e-chain.
- $\{\lambda_t\}$ is bounded in probability on average.
- Possesses a reachable point ($x^* = \alpha/(1-\gamma)$)

3 Parameter Driven Models

3.1 Estimation

- GLM
- Importance sampling
- Approximation to the likelihood

3.2 Simulation and Application

- Time series of counts
- Stochastic volatility

3.3 How good is the posterior approximation?

- Posterior mode vs posterior mean

3.4 Application to estimating structural breaks (tomorrow)

- Poisson model
- Stochastic volatility model

Exponential Family Setup for Parameter-Driven Model

Time series data: $Y_1, \ldots, Y_n$

Regression (explanatory) vector: x_t

Observation equation:

$$p(y_t \mid \alpha_t) = \exp\{(\alpha_t + \beta^T x_t)\, y_t - b(\alpha_t + \beta^T x_t) + c(y_t)\}.$$

State equation: $\{\alpha_t\}$ follows an autoregressive process satisfying the recursions

$$\alpha_t = \gamma + \phi_1 \alpha_{t-1} + \phi_2 \alpha_{t-2} + \ldots + \phi_p \alpha_{t-p} + \varepsilon_t ,$$

where $\{\varepsilon_t\} \sim$ IID $N(0,\sigma^2)$.

Note: $\alpha_t = 0$ corresponds to standard generalized linear model.

Original primary objective: Inference about β.

Estimation Methods for Parameter Driven Models

- Estimating equations (Zeger `88):
- Monte Carlo EM (Chan and Ledolter `95)

- GLM (ignores the presence of the latent process, i.e., $\alpha_t = 0$.)
- Importance sampling (Durbin & Koopman `01, Kuk `99, Kuk & Chen `97):
- Approximate likelihood (Davis, Dunsmuir & Wang '98)

Estimation Methods — Estimating Equations

Estimating equations (Zeger `88): Let $\hat{\boldsymbol{\beta}}$ be the solution to the equation

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \Gamma_n (\mathbf{y}_n - \boldsymbol{\mu}) = 0,$$

where $\mu = \exp(X\,\beta)$ and $\Gamma_n = \mathrm{var}(\mathbf{Y}_n)$.

Iterative weighted least squares can be used to compute $\hat{\boldsymbol{\beta}}$. (See Zeger for details and asymptotic results.)

Estimation Methods — MCEM

Monte Carlo EM (Chan and Ledolter `95): Given $\psi^{(k)}$ from the k-th iteration, $\psi^{(k+1)}$ is computed in the two steps:

E-Step: Compute $Q(\psi|\ \psi^{(k)}) = E(L(\psi;\ \mathbf{y}_n,\ \alpha_n)\ |\ \mathbf{y}_n,\ \psi^{(k)})$,

- $L(\psi;\ \mathbf{y}_n,\ \alpha_n)$ is the log-likelihood based on $\mathbf{y}_n,\ \alpha_n$
- expectation taken with respect to $p(\alpha_n\ |\ \mathbf{y}_n,\ \psi^{(k)})$

M-Step: Update $\psi^{(k)}$ by maximizing $Q(\psi|\ \psi^{(k)})$ with respect to ψ

Note: This procedure is relatively straightforward except for drawing samples from $p(\alpha_n\ |\ \mathbf{y}_n,\ \psi^{(k)})$ in the E-step. Chan and Ledolter use a Gibbs sampler for this.

GLM estimation – (Linear Regression Model)

Suppose $\{Y_t\}$ follows the linear model with time series errors given by

$$Y_t = \mathbf{x}_t^T \beta + W_t ,$$

where $\{W_t\}$ is a stationary (ARMA) time series.

- Estimate β by ordinary least squares (OLS).
- OLS estimate has same asymptotic efficiency as MLE.
- Asymptotic covariance matrix of $\hat{\beta}_{OLS}$ depends on ARMA parameters.
- Identify and estimate ARMA parameters using the estimated residuals,

$$W_t = Y_t - \mathbf{x}_t^T \hat{\beta}_{OLS}$$

- Re-estimate β and ARMA parameters using full MLE.

Estimation Methods Specialized to Poisson Example— GLM estimation

Model: $Y_t \mid \alpha_t , x_t \sim Pois(\exp(x_t^T \beta + \alpha_t))$.

GLM log-likelihood:

$$l(\beta) = -\sum_{t=1}^{n} e^{x_t^T \beta} + \sum_{t=1}^{n} y_t x_t^T \beta - \log\left[\prod_{t=1}^{n} y_t!\right]$$

(This *likelihood* ignores presence of the latent process.)

Assumptions on regressors:

$$\Omega_{I,n} = n^{-1} \sum_{t=1}^{n} x_t x_t^T \mu_t \to \Omega_I(\beta),$$

$$\Omega_{II,n} = n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} x_t x_s^T \mu_t \mu_s \gamma_\varepsilon(s-t) \to \Omega_{II}(\beta),$$

Theory of GLM Estimation in Presence of Latent Process

Theorem (Davis, Dunsmuir, Wang `00). Let $\hat{\beta}$ be the GLM estimate of β obtained by maximizing $l(\beta)$ for the Poisson regression model with a stationary lognormal latent process. Then

$$n^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} \mathrm{N}(0, \Omega_{\mathrm{I}}^{-1} + \Omega_{\mathrm{I}}^{-1}\Omega_{\mathrm{II}}\,\Omega_{\mathrm{I}}^{-1}).$$

Notes:

1. $n^{-1}\Omega_{\mathrm{I}}^{-1}$ is the asymptotic cov matrix from a std GLM analysis.

2. $n^{-1}\Omega_{\mathrm{I}}^{-1}\,\Omega_{\mathrm{II}}\,\Omega_{\mathrm{I}}^{-1}$ is the additional contribution due to the presence of the latent process.

3. Result also valid for more general latent processes (mixing, etc),

4. The x_t can depend on the sample size n.

When Does CLT Apply?

Conditions on the regressors hold for:

1. Trend functions.

$$x_{nt} = \mathbf{f}(t/n)$$

where **f** is a continuous function on [0,1]. In this case,

$$n^{-1}\sum_{t=1}^{n} x_t x_t^{\mathrm{T}} \mu_t \to \int_0^1 f(t) f^{\mathrm{T}}(t) e^{f^{\mathrm{T}}(t)\beta} dt,$$

$$n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} x_t x_s^{\mathrm{T}} \mu_t \mu_s \gamma_\varepsilon(s-t) \to \int_0^1 f(t) f^{\mathrm{T}}(t) e^{2 f^{\mathrm{T}}(t)\beta} dt \sum_h \gamma_\varepsilon(h)\,.$$

Remark. $x_{nt} = (1, t/n)$ corresponds to linear regression and works. However $x_t = (1, t)$ does **not** produce consistent estimates say if the *true slope* is negative.

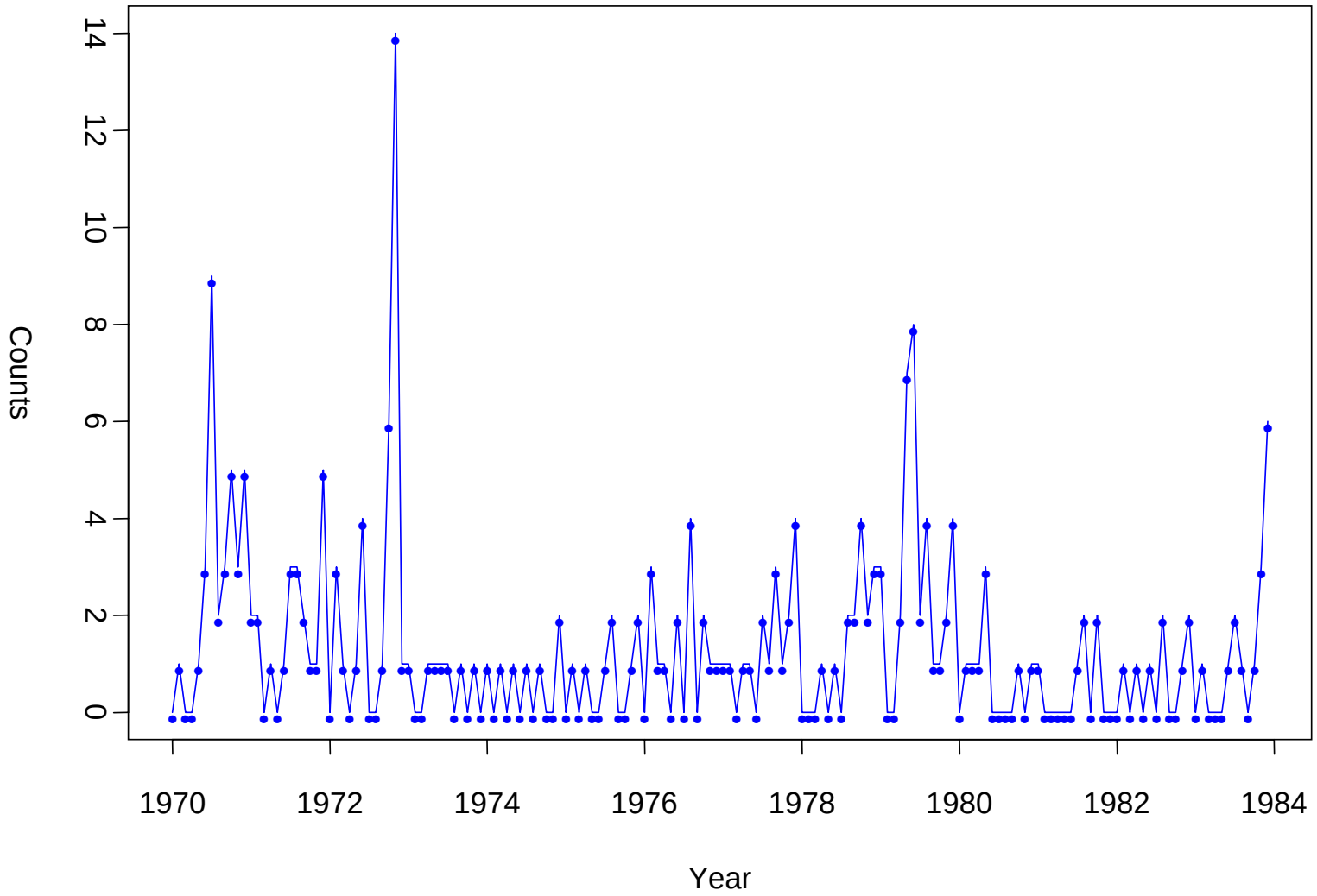
When Does CLT Apply? (cont)

2. Harmonic functions to specify annual or weekly effects, e.g.,

$$x_t = \cos(2\pi t/7)$$

3. Stationary process. (e.g. seasonally adjusted temperature series.)

Application to Polio Data



Application to Model for Polio Data

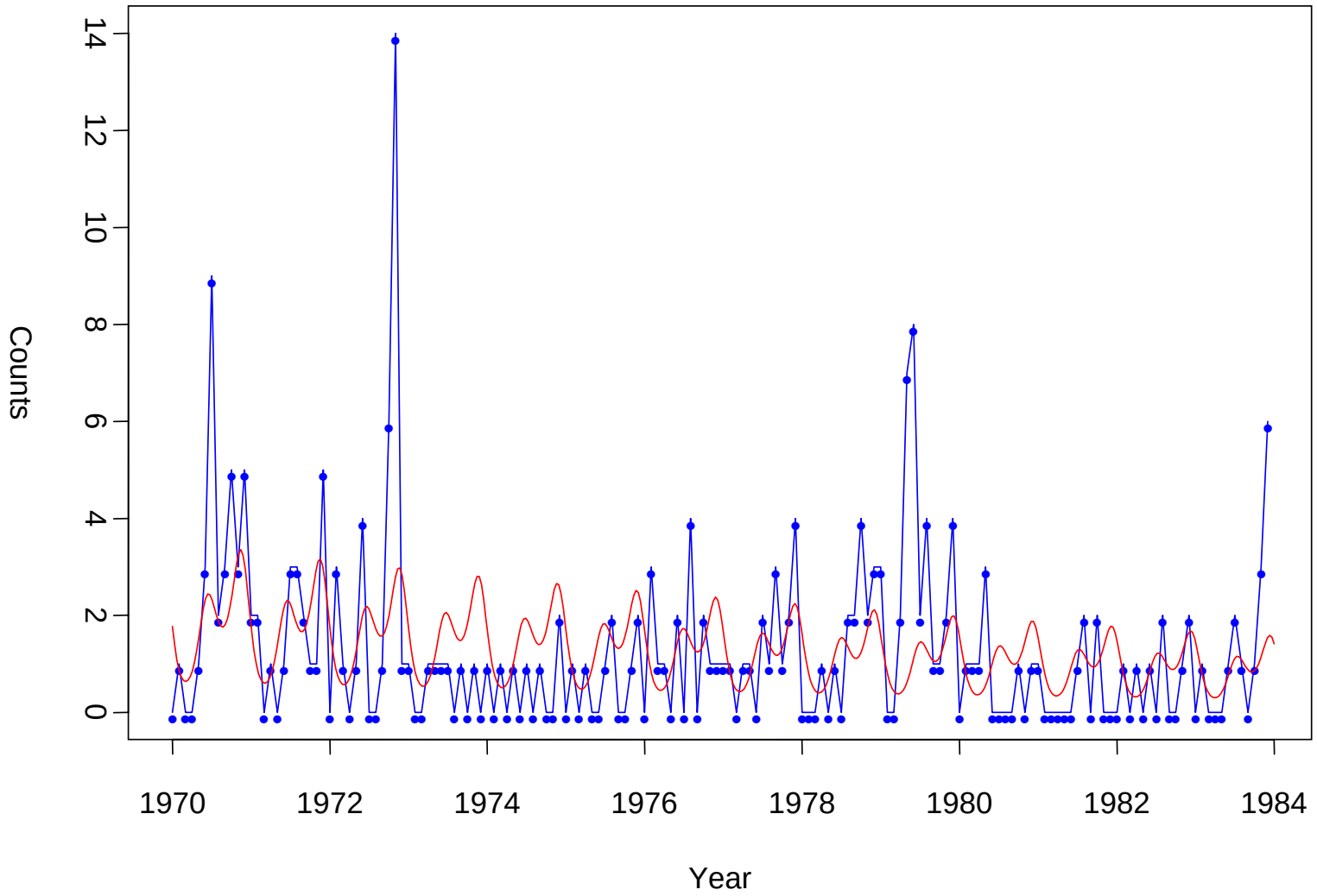
Assume the $\{\alpha_t\}$ follows a log-normal AR(1), where

$$(\alpha_t+\sigma^2/2) = \phi(\alpha_{t-1}+ \sigma^2/2) +\eta_t\ ,\ \ \{\eta_t\}\sim \text{IID } N(0, \sigma^2(1-\phi^2)),$$

with $\phi = .82$, $\sigma^2 = .57$.

	Zeger		GLM Fit		Asym	Simulation	
	$\hat{\beta}_Z$	s.e.	$\hat{\beta}_{GLM}$	s.e.	s.e.	$\hat{\beta}_{GLM}$	s.e.
Intercept	0.17	0.13	.207	.075	.205	.150	.213
Trend($\times 10^{-3}$)	-4.35	2.68	-4.80	1.40	4.12	-4.89	3.94
cos(2πt/12)	-0.11	0.16	-0.15	.097	.157	-.145	.144
sin(2πt/12)	-.048	0.17	-0.53	.109	.168	-.531	.168
cos(2πt/6)	0.20	0.14	.169	.098	.122	.167	.123
σιν(2ππ/6)	-0.41	0.14	-.432	.101	.125	-.440	.125

Polio Data With Estimated Regression Function



Estimation Methods — Importance Sampling (Durbin and Koopman)

Model:

$$Y_t \mid \alpha_t , x_t \sim Pois(\exp(x_t^T \beta + \alpha_t))$$

$$\alpha_t = \phi \alpha_{t-1} + \varepsilon_t , \ \{\varepsilon_t\} \sim \text{IID } N(0, \sigma^2)$$

Relative Likelihood: Let $\psi=(\beta, \phi, \sigma^2)$ and suppose $g(y_n, \alpha_n; \psi_0)$ is an approximating joint density for $Y_n= (Y_1, \ldots, Y_n)'$ and $\alpha_n= (\alpha_1, \ldots, \alpha_n)'$.

$$L(\psi) = \int p(y_n \mid \alpha_n) p(\alpha_n) d\alpha_n$$

$$= \int \frac{p(y_n \mid \alpha_n) p(\alpha_n)}{g(y_n, \alpha_n; \psi_0)} g(y_n, \alpha_n; \psi_0) d\alpha_n$$

$$= \int \frac{p(y_n \mid \alpha_n) p(\alpha_n)}{g(y_n, \alpha_n; \psi_0)} g(\alpha_n \mid y_n; \psi_0) g(y_n; \psi_0) d\alpha_n$$

$$\frac{L(\psi)}{L_g(\psi_0)} = \int \frac{p(y_n \mid \alpha_n) p(\alpha_n)}{g(y_n, \alpha_n; \psi_0)} g(\alpha_n \mid y_n; \psi_0) d\alpha_n$$

Importance Sampling (cont)

$$\frac{L(\psi)}{L_g(\psi_0)} = \int \frac{p(y_n \mid \alpha_n)p(\alpha_n)}{g(y_n, \alpha_n; \psi_0)} g(\alpha_n \mid y_n; \psi_0) d\alpha_n$$

$$= E_g\left[\frac{p(y_n \mid \alpha_n)p(\alpha_n)}{g(y_n, \alpha_n; \psi_0)} \mid y_n; \psi_0 \right]$$

$$\sim \frac{1}{N} \sum_{j=1}^{N} \frac{p(y_n \mid \alpha_n^{(j)})p(\alpha_n^{(j)})}{g(y_n, \alpha_n^{(j)}; \psi_0)},$$

where $\{\alpha_n^{(j)}; j = 1, \ldots, N\} \sim \text{iid } g(\alpha_n \mid y_n; \psi_0)$.

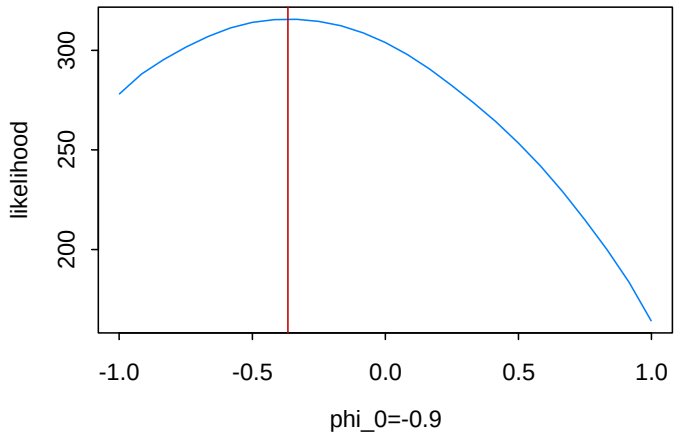
Notes:

- This is a "one-sample" approximation to the relative likelihood. That is, for one realization of the α's, we have, in principle, an approximation to the whole likelihood function.
- Approximation is only good in a neighborhood of ψ_0. Geyer suggests maximizing ratio wrt ψ and iterate replacing ψ_0 with $\hat{\psi}$.

Importance Sampling — example

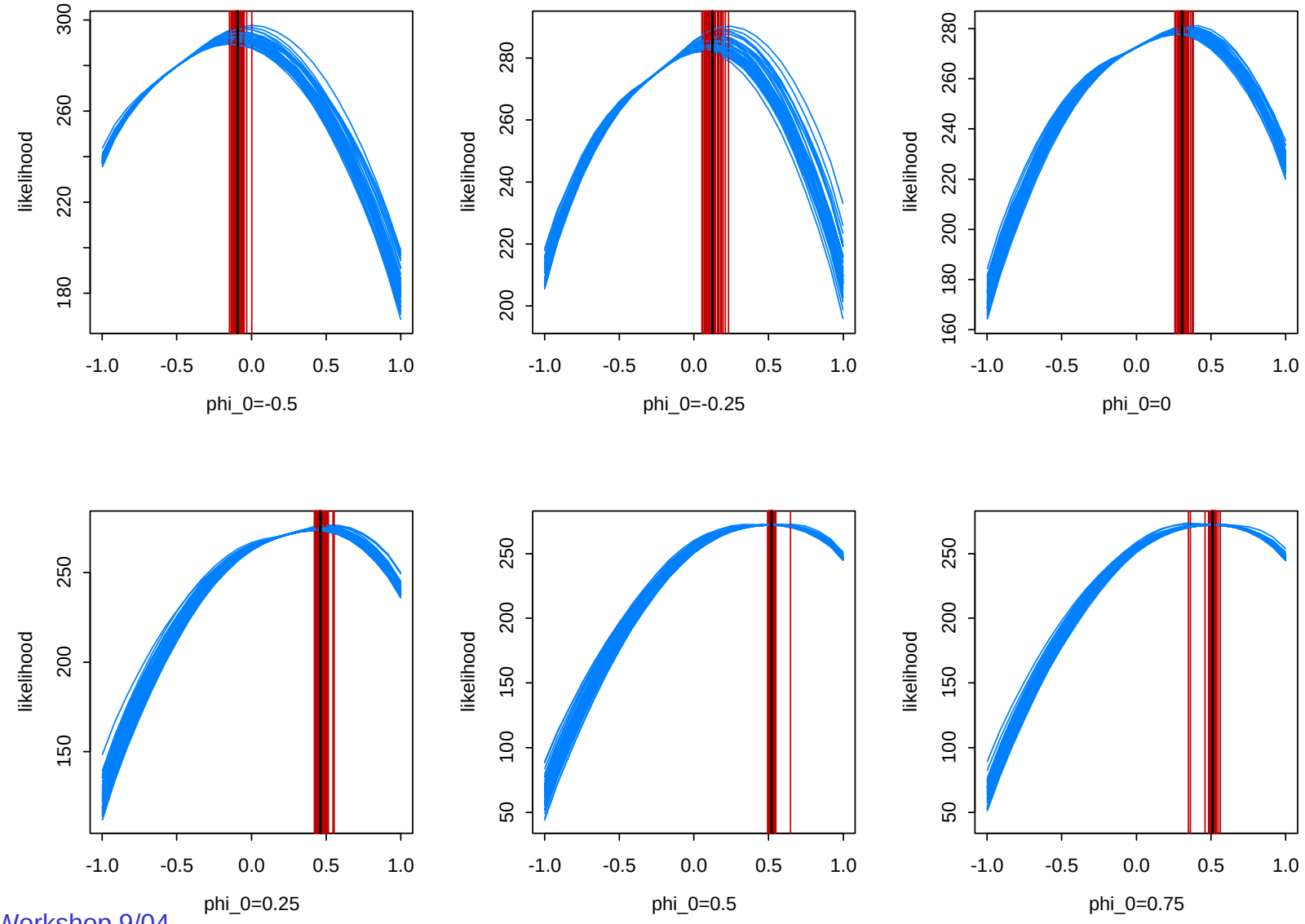
Simulation example: $Y_t \mid \alpha_t \sim Pois(\exp(.7 + \alpha_t))$,

$\alpha_t = .5\ \alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID N(0, .3), $n = 200$, $N = 1000$



Importance Sampling — example

Simulation example: $\beta = .7$, $\phi = .5$, $\sigma^2 = .3$, $n = 200$, $N = 1000$, 50 realizations plotted



Importance Sampling (cont)

- Instead one can use $\psi_0 = \psi$ to get

$$L(\psi) = L_g(\psi)\int \frac{p(y_n \mid \alpha_n)p(\alpha_n)}{g(y_n, \alpha_n; \psi)} g(\alpha_n \mid y_n; \psi) d\alpha_n$$

$$= L_g(\psi) E_g\left[\frac{p(y_n \mid \alpha_n)}{g(y_n \mid \alpha_n; \psi)} \mid y_n; \psi \right]$$

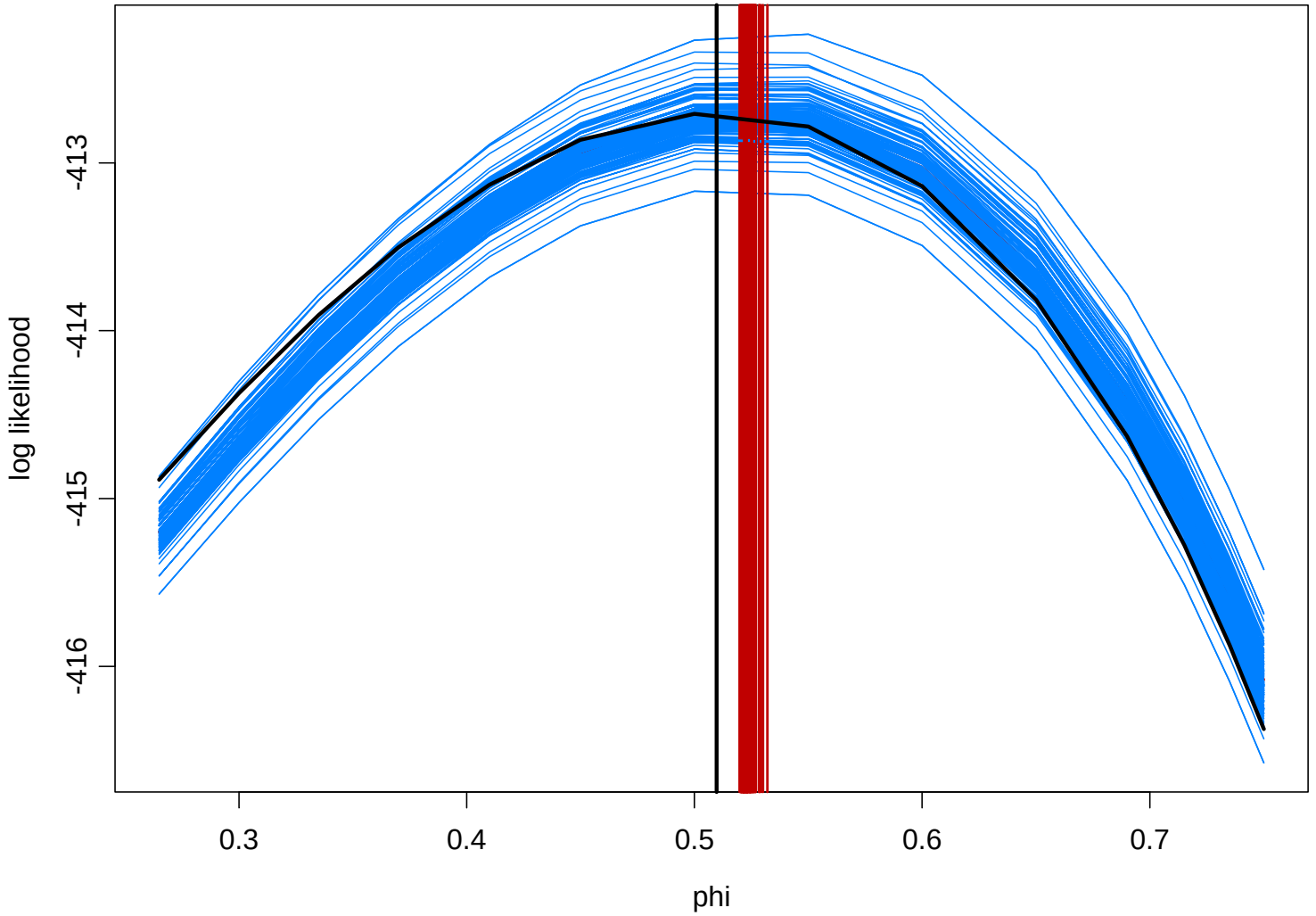
$$\sim \frac{L_g(\psi)}{N} \sum_{j=1}^{N} \frac{p(y_n \mid \alpha_n^{(j)})}{g(y_n \mid \alpha_n^{(j)}; \psi)},$$

where for each ψ, $\{\alpha_n^{(j)}; j = 1, \ldots, N\} \sim \text{iid } g(\alpha_n \mid y_n; \psi)$, are generated using the same noise sequence. This is the idea behind Durbin and Koopman's implementation.

Importance Sampling — example

Simulation example: $Y_t \mid \alpha_t \sim Pois(\exp(.7 + \alpha_t))$,

$\alpha_t = .5\ \alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID N(0, .3), $n = 200$, $N = 1000$



Importance Sampling (cont)

Choice of *importance density* g:

Durbin and Koopman suggest a linear state-space approximating model

$$Y_t = \mu_t + x_t^T \beta + \alpha_t + Z_t, \quad Z_t \sim N(0, H_t),$$

with

$$\mu_t = y_t - \hat{\alpha}_t - x'_t \, y_t e^{-(\hat{\alpha}_t + x'_t \beta)} + 1,$$

$$H_t = e^{-(\hat{\alpha}_t + x'_t \beta)},$$

where the $\hat{\alpha}_t = E_g(\alpha_t \mid y_n)$ are calculated recursively under the approximating model until convergence.

With this choice of approximating model, it turns out that

$$g(\alpha_n \mid y_n; \psi_0) \sim N(\Gamma_n^{-1} \tilde{y}_n, \Gamma_n^{-1}),$$

where

$$\tilde{y}_n = y_n - e^{X\beta + \hat{\alpha}_n} + e^{X\beta + \hat{\alpha}_n} \hat{\alpha}_n,$$

$$\Gamma_n = \mathrm{diag}(e^{X\beta + \hat{\alpha}_n}) + (E(\alpha_n \alpha'_n))^{-1}.$$

Importance Sampling (cont)

Components required in the calculation.

- $g(y_n,\alpha_n)$
  - $\tilde{y}'_n \Gamma_n^{-1} \tilde{y}_n$
  - $\det(\Gamma_n)$
- simulate from $N(\Gamma_n^{-1}\tilde{y}_n, \Gamma_n^{-1})$
  - compute $\Gamma_n^{-1}\tilde{y}_n$
  - simulate from $N(0, \Gamma_n^{-1})$

Remark: These quantities can be computed quickly using a version of the *innovations algorithm* or the *Kalman smoothing recursions*.

Estimation Methods — Approximation to the likelihood

General setup:

$$p(\mathrm{y}_n, \alpha_n) \propto p(\mathrm{y}_n \mid \alpha_n) \mid G_n \mid^{1/2} \exp\{-(\alpha_n - \mu)^T G_n (\alpha_n - \mu)/2\}$$

where

$$G_n^{-1} = E(\alpha_n - \mu)^T (\alpha_n - \mu)$$

Likelihood:

$$L(\psi) = \int p(\mathrm{y}_n \mid \alpha_n) p(\alpha_n) d\alpha_n$$

Log-likelihood based on y_n and α_n :

$$l(\psi; \mathrm{y}_n, \alpha_n) = \log p(\mathrm{y}_n, \alpha_n; \psi)$$

$$= -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log \mid \mathrm{G}_n \mid + l(\theta; \mathrm{y}_n \mid \alpha_n) - \frac{1}{2}(\alpha_n - \mu)^T G_n (\alpha_n - \mu)$$

Estimation Methods — Approximation to the likelihood (cont)

If $T(\alpha;\alpha^*)$ denotes the second-order Taylor series expansion of $l(\theta; y_n|\alpha_n)$ around α^* with remainder

$$R(\alpha;\alpha^*) = l(\theta; y_n|\alpha_n) - T(\alpha;\alpha^*),$$

where α^* is the mode of $l(\psi; y_n, \alpha_n)$, we have

$$l(\psi; y_n, \alpha_n) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|G_n| + h^* - \frac{1}{2}(\alpha^* - \mu)^T G_n(\alpha^* - \mu)$$
$$-\frac{1}{2}(\alpha_n - \alpha^*)^T(K^* + G_n)(\alpha_n - \alpha^*) + R(\alpha_n;\alpha^*),$$

where

$$h^* = l(\theta; y_n|\alpha_n)\Big|_{\alpha_n=\alpha_n^*}$$

$$K^* = -\frac{\partial^2}{\partial\alpha_n \partial\alpha_n^T} l(\theta; y_n|\alpha_n)\Big|_{\alpha_n=\alpha_n^*}$$

Estimation Methods — Approximation to the likelihood (cont)

Let $p_a(\alpha_n \mid y_n, \psi)$ be the posterior based on the log likelihood neglecting $R(\alpha ;\alpha^*)$,

i.e., $$p_a(\alpha_n \mid y_n, \psi) = \phi(\alpha_n; \alpha^*, (K^* + G_n)^{-1})$$

Then

$$L(\psi; y_n) = \int L(\psi; y_n, \alpha_n)\, d\alpha_n$$
$$= L_a(\psi; y_n) Er_a(\psi),$$

where

$$L_a(\psi; y_n) = \frac{|G_n|^{1/2}}{|K^* + G_n|^{1/2}} \exp\{h^* - (\alpha^* - \mu)^T G_n (\alpha^* - \mu)/2\}$$

and

$$Er_a(\psi) = \int \exp\{R(\alpha_n; \alpha^*)\} p_a(\alpha_n \mid y_n; \psi)\, d\alpha_n$$

Estimation Methods — Approximation to the likelihood (cont)

Notes:

1. The approximating posterior,

$$p_a(\alpha_n \mid y_n, \psi) = \phi(\alpha_n; \alpha^*, (K^* + G_n)^{-1}),$$

is identical to the importance sampling density used by Durbin and Koopman for the case of exponential families.

2. In traditional Bayesian setting, posterior is approximately p_a for large n (see Bernardo and Smith, 1994).

Estimation Methods — Approximation to the likelihood (cont)

Recall that

$$L(\psi; \mathrm{y}_n) = L_a(\psi; \mathrm{y}_n) Er_a(\psi)$$

1. Importance sampling (IS).

$$\hat{L}(\psi; \mathrm{y}_n) = L_a(\psi; \mathrm{y}_n) \hat{E}r_a(\psi),$$

where

$$\hat{E}r_a = \frac{1}{N} \sum_{i=1}^{N} \exp(R(\alpha_n^{(i)}, \alpha^*)),$$

and $\{\alpha_n^{(j)}; j = 1, \ldots, N\} \sim \text{iid } \mathrm{p}_\mathrm{a}(\alpha_n \mid \mathrm{y}_n; \psi).$

Estimation Methods — Approximation to the likelihood (cont)

$$L(\psi; y_n) = L_a(\psi; y_n) Er_a(\psi)$$

2. Approximate likelihood (AL).

If $p_a(\alpha_n \mid y_n, \psi)$ is concentrated around $\alpha_n = \alpha*$, then the error term $Er_a(\psi)$ is close to 1. This suggests ignoring the error term and approximating the likelihood by $\hat{L}(\psi; y_n) = L_a(\psi; y_n)$

3. Hybrid estimate (AIS).

Approximate $e(\psi) = \log Er_a(\psi)$ by a linear function. Let $\hat{\psi}_{AL}$ be the maximum likelihood estimator based on the objective function $L_a(\psi)$. Then

$$e(\psi) \sim e(\hat{\psi}_{AL}) + \dot{e}(\hat{\psi}_{AL})(\psi - \hat{\psi}_{AL})$$

Ignoring the constant term, we have

$$\hat{L}(\psi; y_n) \propto L_a(\psi; y_n) \exp\{\dot{e}(\hat{\psi}_{AL})(\psi - \hat{\psi}_{AL})\}.$$

Use one run of importance sampling to estimate $\dot{e}(\hat{\psi}_{AL})$.

Estimation Methods — Approximation to the likelihood (cont)

Case of exponential family:

$$L_a(\psi; \mathrm{y_n}) = \frac{|G_n|^{1/2}}{(K+G_n)^{1/2}} \exp\{\mathrm{y}_n^T \alpha^* - 1^T\{b(\alpha^*) - c(\mathrm{y}_n)\} - (\alpha^* - \mu)^T G_n (\alpha^* - \mu)/2\},$$

where

$$K^* = \mathrm{diag}\{\frac{\partial^2}{\partial \alpha_t^2} b_t(\alpha_t)\Big|_{\alpha_t^*}\},$$

and α^* is the solution to the equation

$$\mathrm{y}_n - \frac{\partial}{\partial \alpha_n} \mathrm{b}(\alpha_n) - \mathrm{G_n}(\alpha_n - \mu) = 0.$$

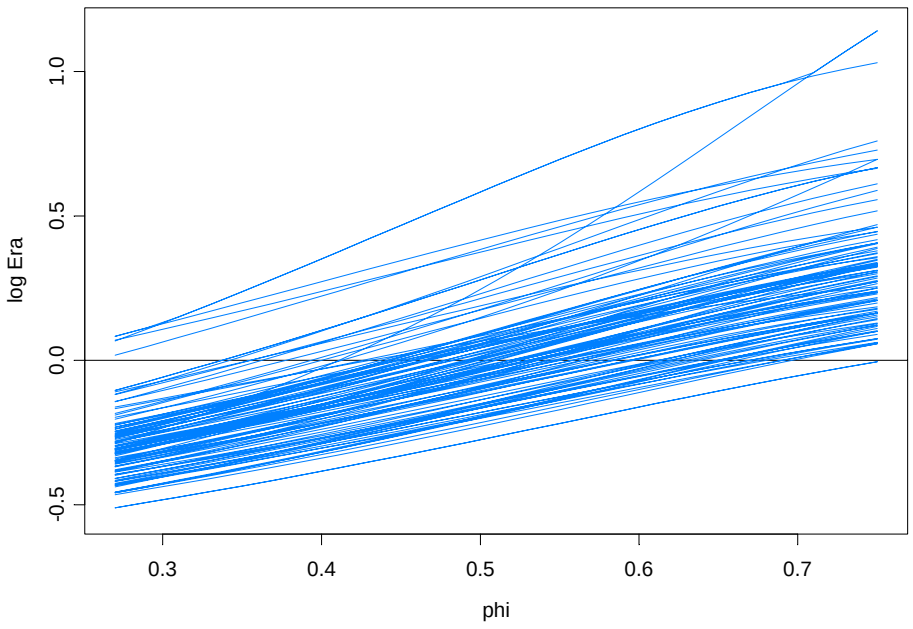
Using a Taylor expansion, the latter equation can be solved iteratively.

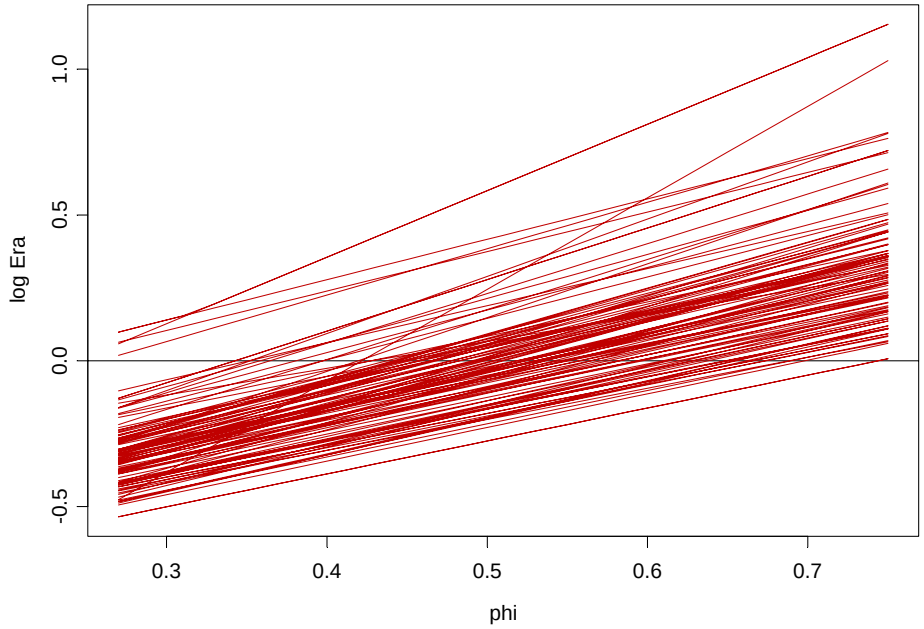
A closer look at log ($Er_a(\psi)$)

$$Er_a(\psi) = \int \exp\{R(\alpha_n;\alpha^*)\}\, p_a(\alpha_n \mid y_n;\psi)\, d\alpha_n$$

$$\log(\hat{Er}_a(\psi)) = \log\left(\frac{1}{N}\sum_{i=1}^{N}\exp(R(\alpha_n^{(i)},\alpha^*))\right)$$

$$e(\psi) \sim e(\hat{\psi}_{AL}) + \dot{e}(\hat{\psi}_{AL})(\psi - \hat{\psi}_{AL})$$

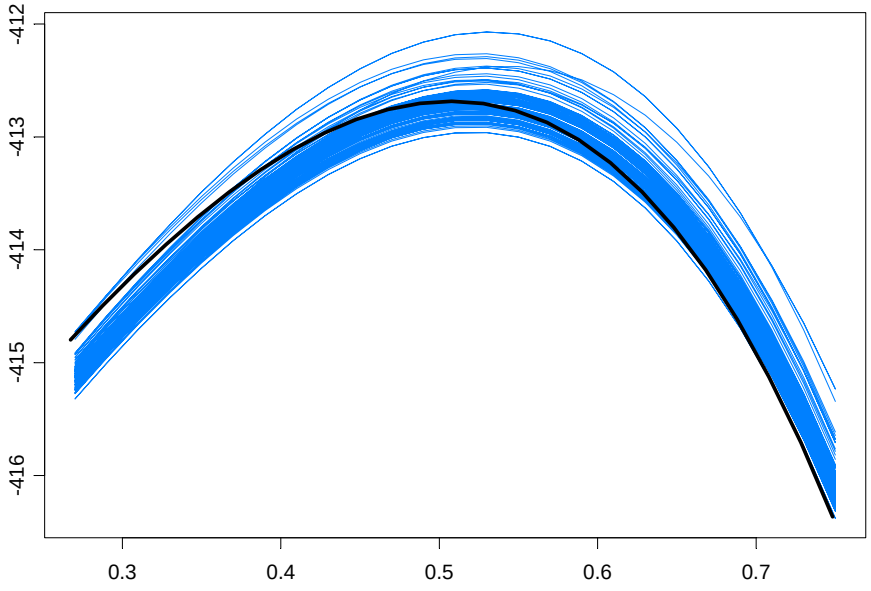




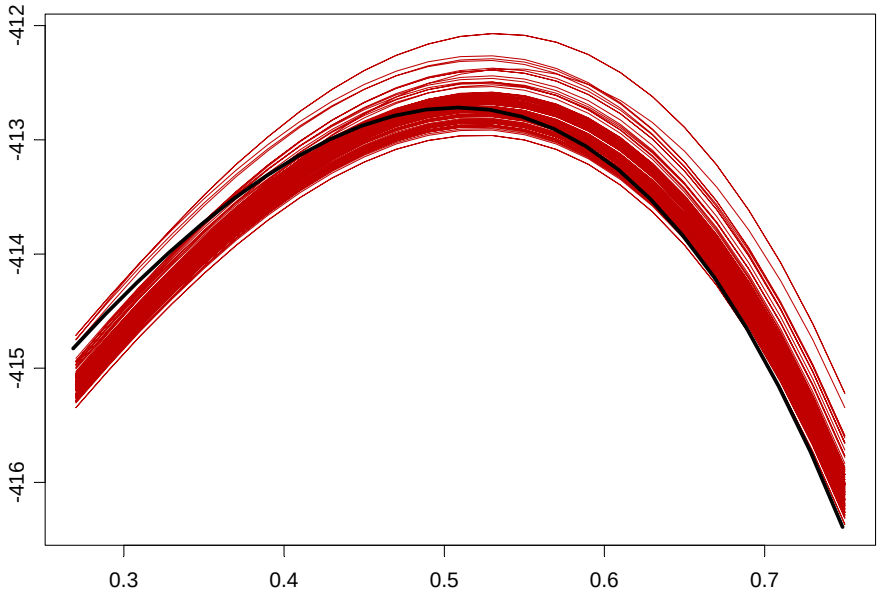
Comparison between IS and AIS

Model: $Y_t \mid \alpha_t \sim Pois(\exp(.7 + \alpha_t))$, $\alpha_t = .5\,\alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID N(0, .3), $n = 200$

Importance Sampling

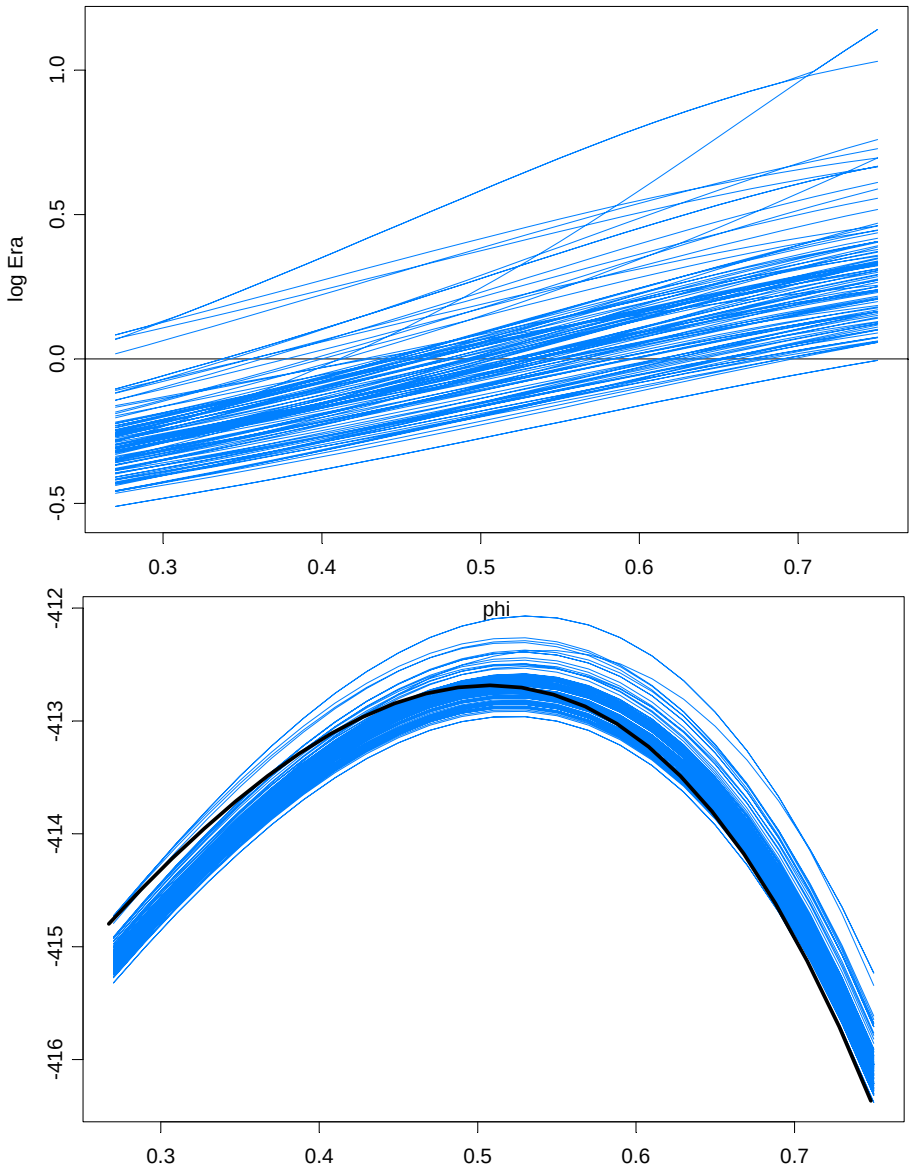


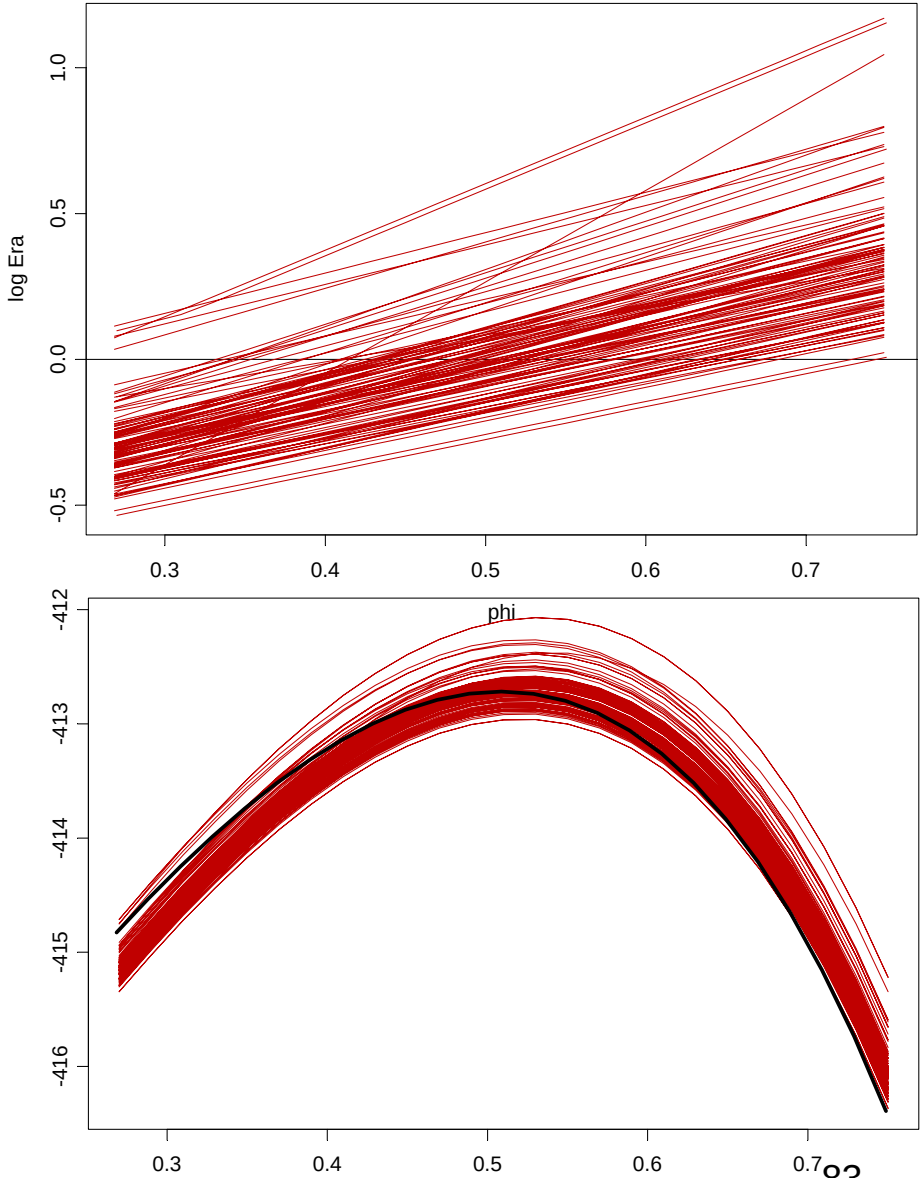
Hybrid: AL + IS



$$\log(\hat{E}r_a(\psi)) = \log\left(\frac{1}{1000}\sum_{i=1}^{1000}\exp(R(\alpha_n^{(i)}, \alpha^*))\right)$$

$$e(\psi) \sim e(\hat{\psi}_{AL}) + \dot{e}(\hat{\psi}_{AL})(\psi - \hat{\psi}_{AL})$$





Comparison between IS and AIS

Model: $Y_t \mid \alpha_t \sim Pois(\exp(1.0 + \alpha_t))$

$\alpha_t = 1.25\alpha_{t-1} - .75\alpha_{t-2} + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID N(0, .2), $n = 200$

Estimation methods:

- Importance sampling (N=100)

	β	ϕ_1	ϕ_2	σ^2
mean	0.9943	1.2520	-.7536	0.1925
rmse	**0.0832**	**0.0749**	**0.0668**	**0.0507**

- Approximate likelihood + IS

	β	ϕ_1	ϕ_2	σ^2
mean	0.9943	1.2520	-.7586	0.1925
rmse	**0.0832**	**0.0749**	**0.0669**	**0.0507**

Comparison between IS and AIS (cont)

Model: $Y_t \mid \alpha_t \sim Pois(\exp(1.0 + \alpha_t))$

$$\alpha_t = 1.25\alpha_{t-1} - .75\alpha_{t-2} + .2\alpha_{t-3} + \varepsilon_t, \quad \{\varepsilon_t\} \sim \text{IID } N(0, .2),\ n = 200$$

Estimation methods:

- Importance sampling (N=100)

	β	ϕ_1	ϕ_2	ϕ_3	σ^2
mean	1.0009	1.2473	-.7496	0.1896	0.1969
rmse	**0.1139**	**0.2765**	**0.3815**	**0.2011**	**0.0783**

- Approximate likelihood + IS

	β	ϕ_1	ϕ_2	ϕ_3	σ^2
mean	1.0009	1.2471	-.7493	0.1894	0.1969
rmse	**0.1139**	**0.2759**	**0.3809**	**0.2009**	**0.0783**

Simulation Results

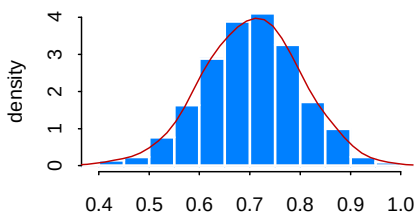
Model: $Y_t \mid \alpha_t \sim Pois(\exp(.7 + \alpha_t))$, $\alpha_t = .5\,\alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID N(0, .3), n = 200

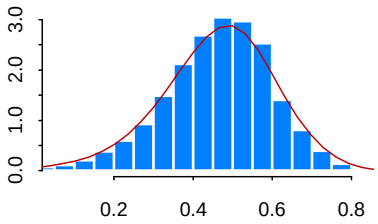
Estimation methods: Simulation results based on 1000 replications.

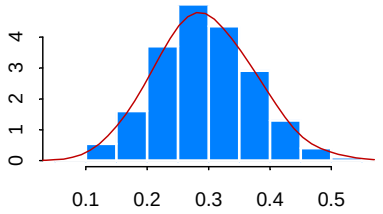
		beta	phi	sigma2
Importance sampling (N=100)	mean	0.7038	0.4783	0.2939
	std	**0.0970**	**0.1364**	**0.0789**
Approx likelihood	mean	0.7035	0.4627	0.2956
	std	**0.0972**	**0.1413**	**0.0785**
AL + IS (AIS)	mean	0.7039	0.4784	0.2938
	std	**0.0969**	**0.1364**	**0.0789**

Model: $Y_t \mid \alpha_t \sim Pois(\exp(.7 + \alpha_t))$, $\alpha_t = .5\,\alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\} \sim$ IID N(0, .3), $n = 200$

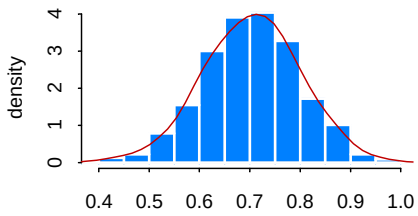
Approx likelihood

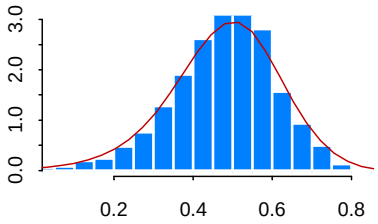


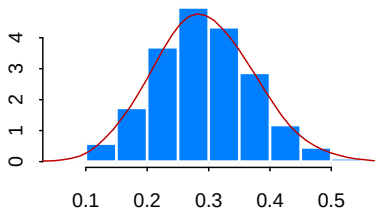




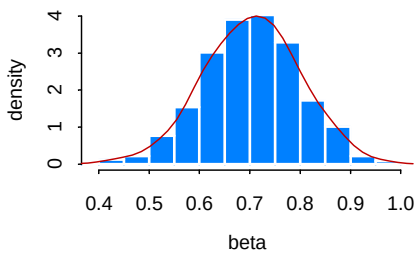
Importance Sampling

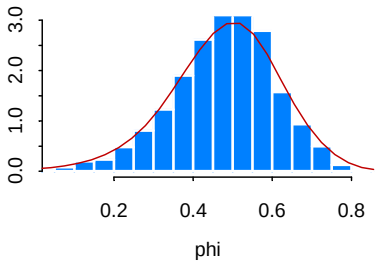


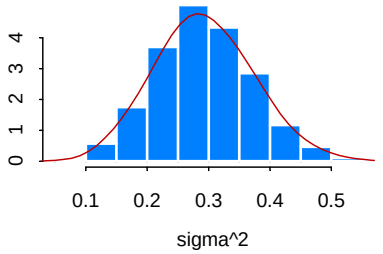




AL + IS (AIS)







Application to Model Fitting for the Polio Data

AR model order selection for $\{\alpha_t\}$:

Approx Like (AL)

p	0	1	2	3	4	5	13
AIC	518.0	**512.3**	512.3	513.9	**512.3**	514.2	527.4
log(like)	-252.0	**-248.1**	-247.1	-246.9	**-245.2**	-245.1	-243.7

Approx + IS (AIS)

p	0	1	2	3	4	5	13
AIC	519.8	**512.6**	512.1	513.9	**511.5**	514.3	527.4
log(like)	-252.9	**-248.3**	-247.0	-246.9	**-244.8**	-245.2	-243.7

Application to Model Fitting for the Polio Data

Model for $\{\alpha_t\}$:

$$\alpha_t = \phi\alpha_{t-1}+\varepsilon_t\ ,\ \ \{\varepsilon_t\}\sim \text{IID } N(0, \sigma^2).$$

Importance sampling (N=1000)

	Import Sampling		AL + IS		Approx Like	
	$\hat{\beta}_{IS}$	SE	$\hat{\beta}_{AIS}$	SE	$\hat{\beta}_{AL}$	SE
Intercept	**0.238**	0.283	**0.239**	0.285	**0.242**	0.273
Trend($\times 10^{-3}$)	**-3.744**	2.823	**-3.746**	2.867	**-3.814**	2.767
cos(2pt/12)	**0.160**	0.152	0.**161**	0.151	**0.162**	0.142
sin(2pt/12)	**-0.481**	0.168	-0.**480**	0.164	**-0.482**	0.166
cos(2pt/6)	**0.414**	0.128	0.**414**	0.122	**0.413**	0.128
sin(2pt/6)	**-0.011**	0.126	-0.**011**	0.127	**-0.011**	0.129
f	**0.661**	0.202	**0.661**	0.209	**0.627**	0.229
s^2	**0.272**	0.115	**0.272**	0.112	**0.289**	0.122

Simulation Results

Stochastic volatility model:

$Y_t = \sigma_t Z_t$, $\{Z_t\}$~IID N(0,1)

$\alpha_t = \gamma + \phi\alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\}$~IID N(0,σ^2), where $\alpha_t = 2 \log \sigma_t$; n=500, N=100,NR=500

CV=10

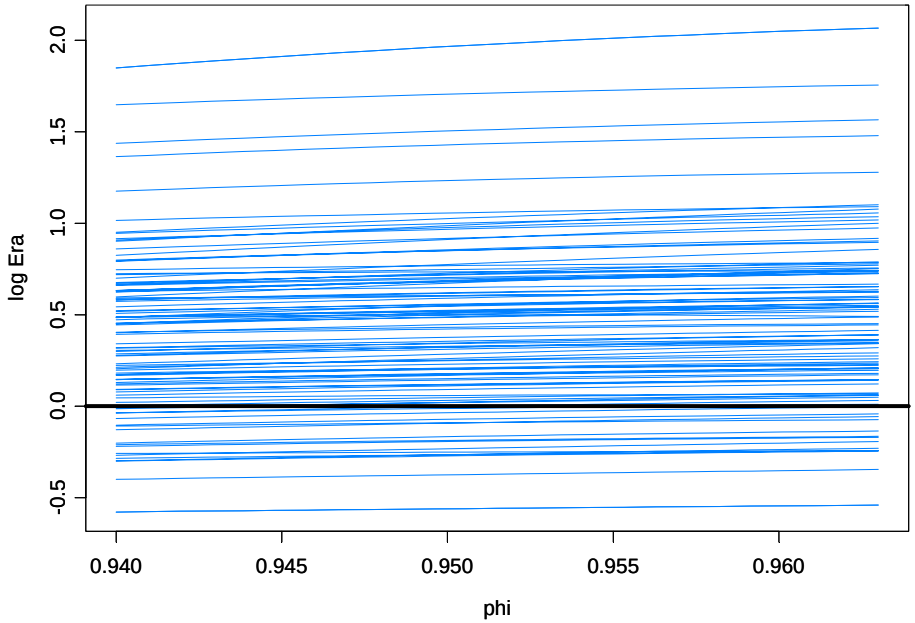
	True	AL	RMSE	AIS	RMSE
γ	−.411	**−.491**	.210	**−.481**	.195
ϕ	0.950	**0.940**	.025	**0.942**	.023
σ	0.484	**0.479**	.065	**0.477**	.064

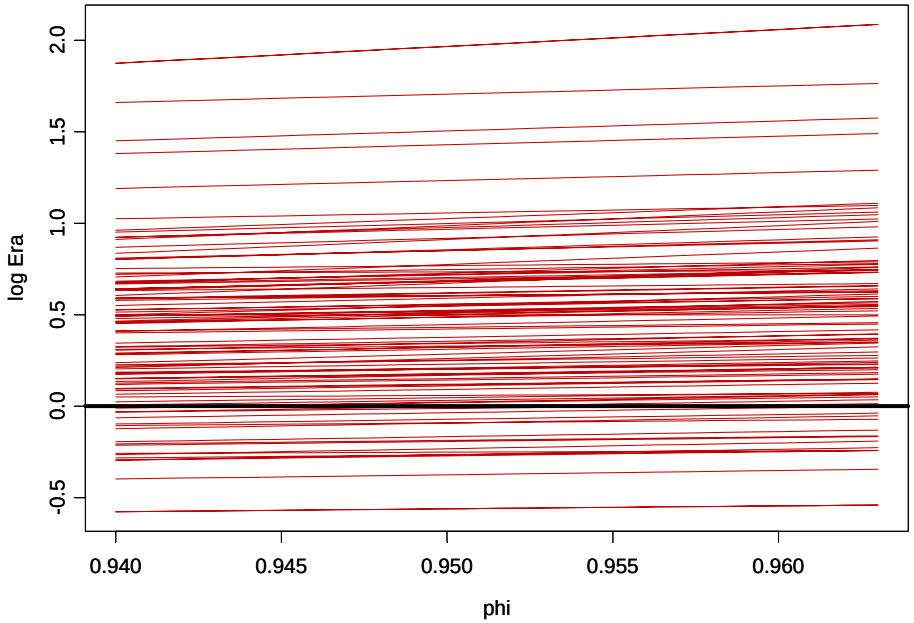
CV=1

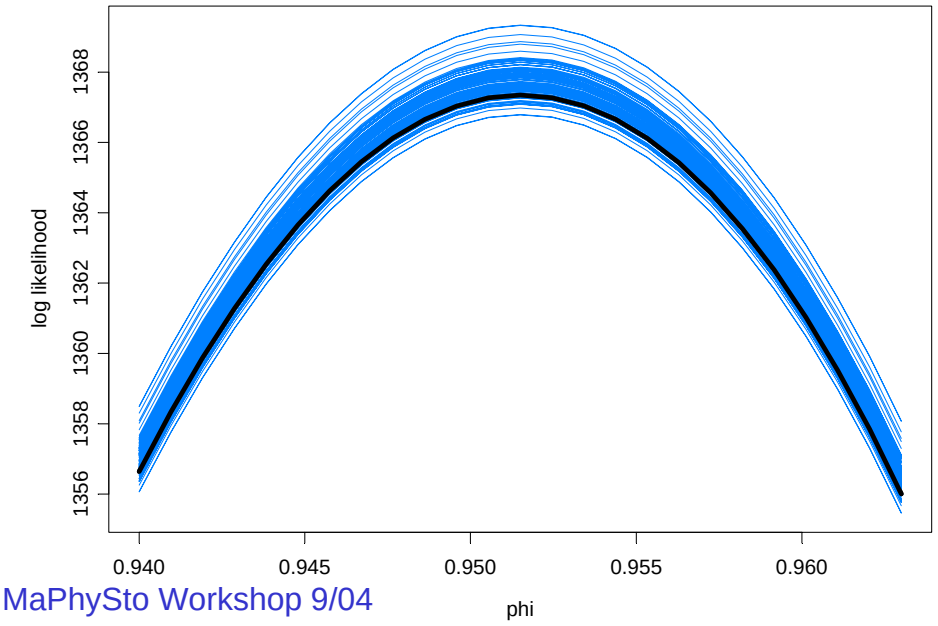
	True	AL	RMSE	AIS	RMSE
γ	−.368	**−.500**	.341	**−.483**	.299
ϕ	0.950	**0.932**	.046	**0.934**	.040
σ	0.260	**0.270**	.068	**0.269**	.064

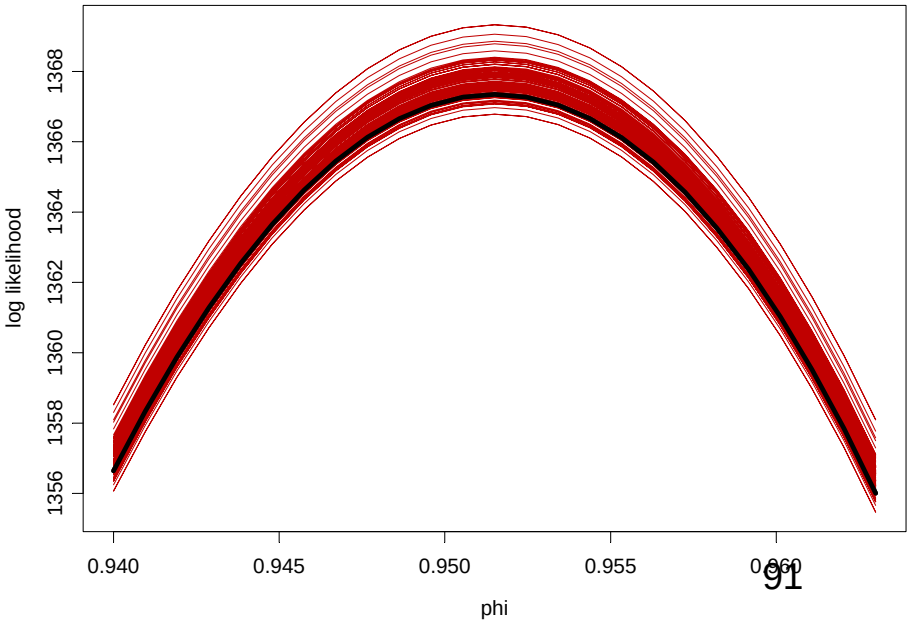
$$\log(\hat{E}r_a(\psi)) = \log\left(\frac{1}{500}\sum_{i=1}^{500}\exp(R(\alpha_n^{(i)}, \alpha^*))\right)$$

$$e(\psi) \sim e(\hat{\psi}_{AL}) + \dot{e}(\hat{\psi}_{AL})(\psi - \hat{\psi}_{AL})$$









Application to Pound-Dollar Exchange Rates

Stochastic volatility model:

$Y_t = \sigma_t Z_t$, $\{Z_t\}$~IID N(0,1)

$\alpha_t = \gamma + \phi\alpha_{t-1} + \varepsilon_t$, $\{\varepsilon_t\}$~IID N(0,σ^2), where $\alpha_t = 2 \log \sigma_t$; N=1000, B=500

	AL	BC	SE	AIS	BC	SE
γ	-0.0227			-0.0230		
ϕ	0.9750			0.9747		
σ^2	0.0267			0.0273		

Application to Sydney Asthma Count Data

Data: $Y_1, \ldots, Y_{1461}$ daily asthma presentations in a Campbelltown hospital.

Preliminary analysis identified.

- no upward or downward trend
- annual cycle modeled by $\cos(2\pi t/365)$, $\sin(2\pi t/365)$
- seasonal effect modeled by

$$P_{ij}(t) = \frac{1}{B(2.5,5)} \left(\frac{t - T_{ij}}{100} \right)^{2.5} \left(1 - \frac{t - T_{ij}}{100} \right)^{5}$$

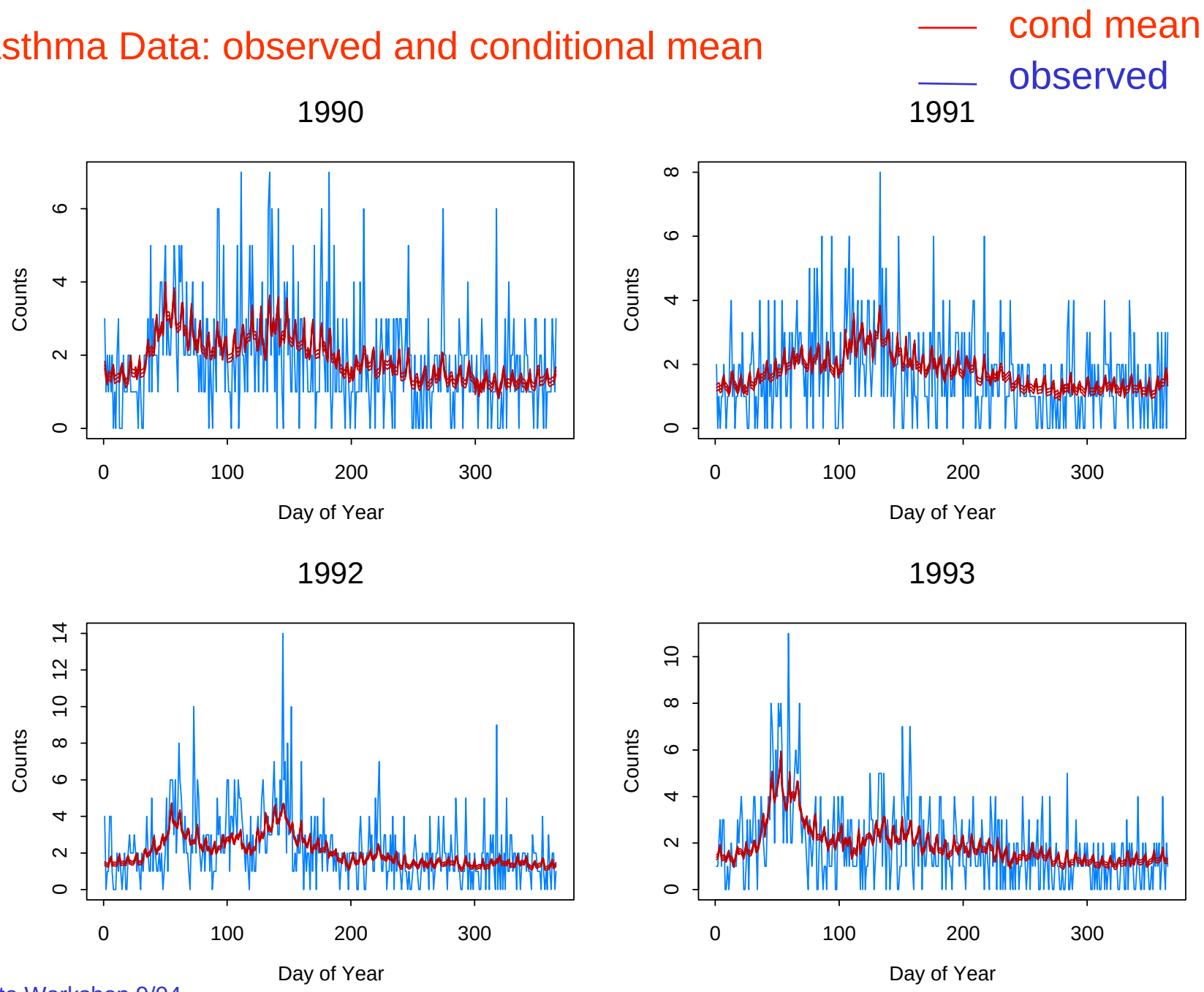
  where $B(2.5,5)$ is the beta function and T_{ij} is the start of the j^{th} school term in year i.

- day of the week effect modeled by separate indicator variables for Sunday and Monday (increase in admittance on these days compared to Tues-Sat).
- Of the meteorological variables (max/min temp, humidity) and pollution variables (ozone, NO, NO_2), only humidity at lags of 12-20 days and NO_2(max) appear to have an association.

Results for Asthma Data—(AL & AIS, N=100)

Term	AL	SE	AIS	SE
Intercept	0.568	**.064**	0.569	**.066**
Sunday effect	0.199	**.053**	0.199	**.051**
Monday effect	0.225	**.053**	0.226	**.052**
cos(2pt/365)	-0.214	**.041**	-0.214	**.041**
sin(2pt/365)	0.177	**.043**	0.177	**.048**
Term 1, 1990	0.199	**.063**	0.200	**.065**
Term 2, 1990	0.133	**.066**	0.133	**.063**
Term 1, 1991	0.085	**.070**	0.085	**.078**
Term 2, 1991	0.171	**.064**	0.171	**.069**
Term 1, 1992	0.249	**.061**	0.249	**.066**
Term 2, 1992	0.302	**.060**	0.302	**.058**
Term 1, 1993	0.431	**.060**	0.432	**.062**
Term 2, 1993	0.114	**.067**	0.114	**.075**
Humidity $H_t/20$	0.009	**.003**	0.009	**.003**
NO_2 max	-0.101	**.033**	-0.101	**.034**
AR(1), ϕ	0.774	**.327**	0.786	**.394**
σ^2	0.011	**.015**	0.010	**.013**

Asthma Data: observed and conditional mean



Is the posterior distribution close to normal?

Compare posterior mean with posterior mode: Can compute the posterior mean using *SIR* (sampling importance-resampling) or particle filtering.

Posterior mode: The mode of $p(\alpha_n \mid y_n)$ is α^* found at the last iteration of AL.

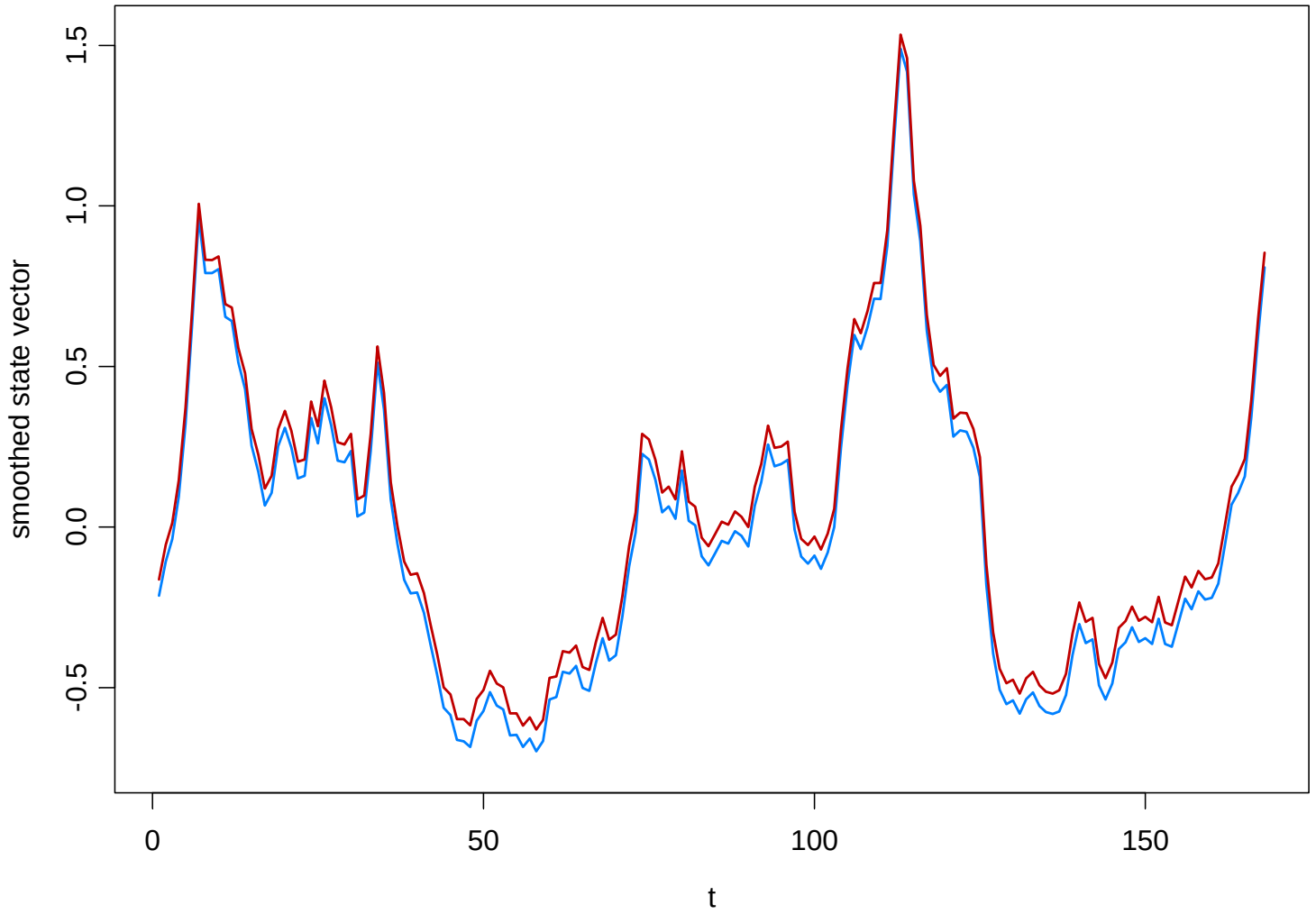
Posterior mean: The mean of $p(\alpha_n \mid y_n)$ can be found using SIR.

Let $\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(N)}$ be independent draws from the multivariate distr $p_a(\alpha_n \mid y_n)$. For N large, an approximate iid sample from $p(\alpha_n \mid y_n)$ can be obtained by drawing a random sample from $\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(N)}$ with probabilities

$$p_i = \frac{w_i}{\sum_{i=1}^{N} w_i}, \quad w_i = \frac{p(\alpha^{(i)} \mid y_n)}{p_a(\alpha^{(i)} \mid y_n)} \propto \exp\{R(\alpha^{(i)}; \alpha^*)\}, \quad i = 1, \ldots, N.$$

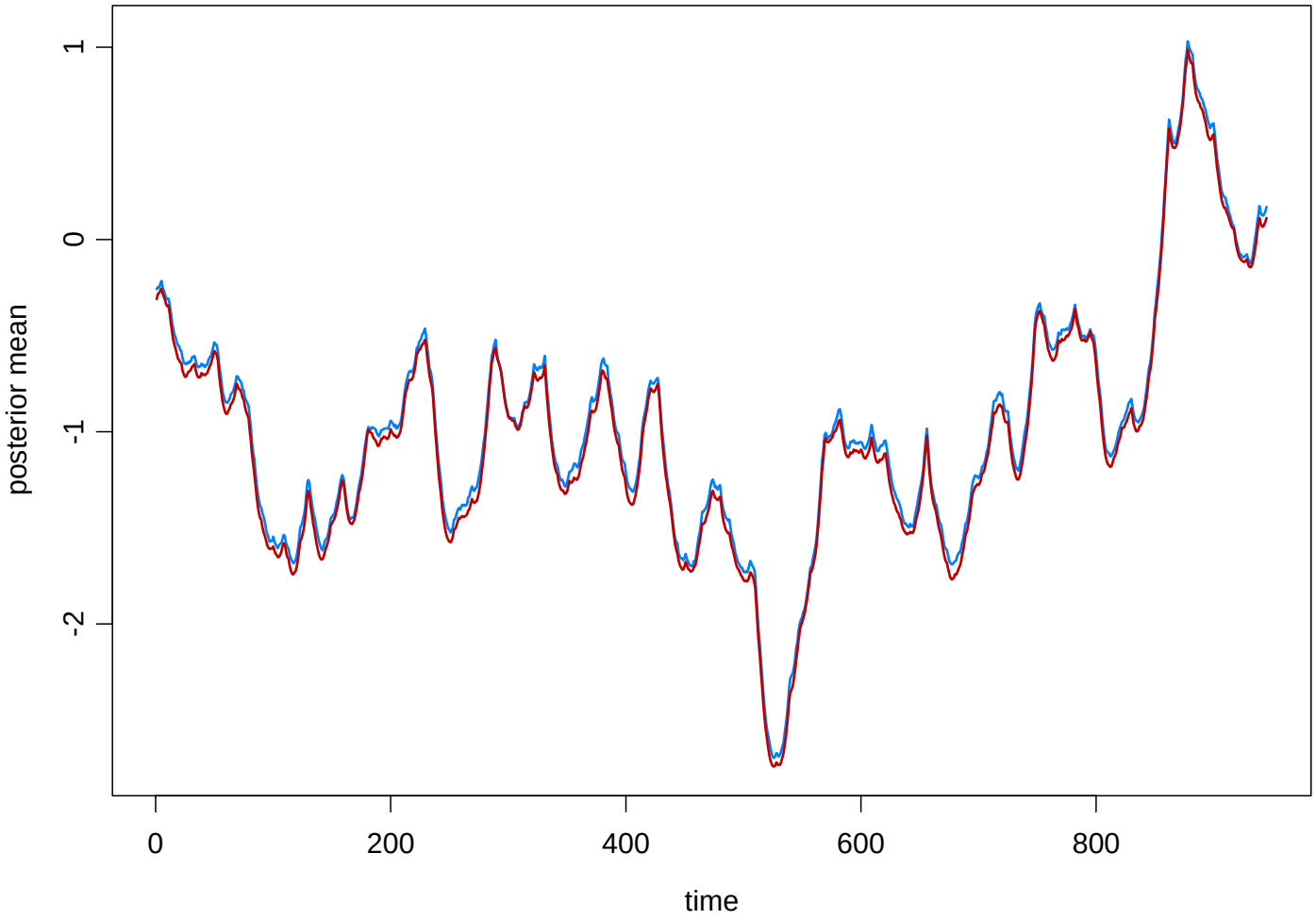
Posterior mean vs posterior mode?

Polio data: blue = mean, red = mode



Posterior mean vs posterior mode?

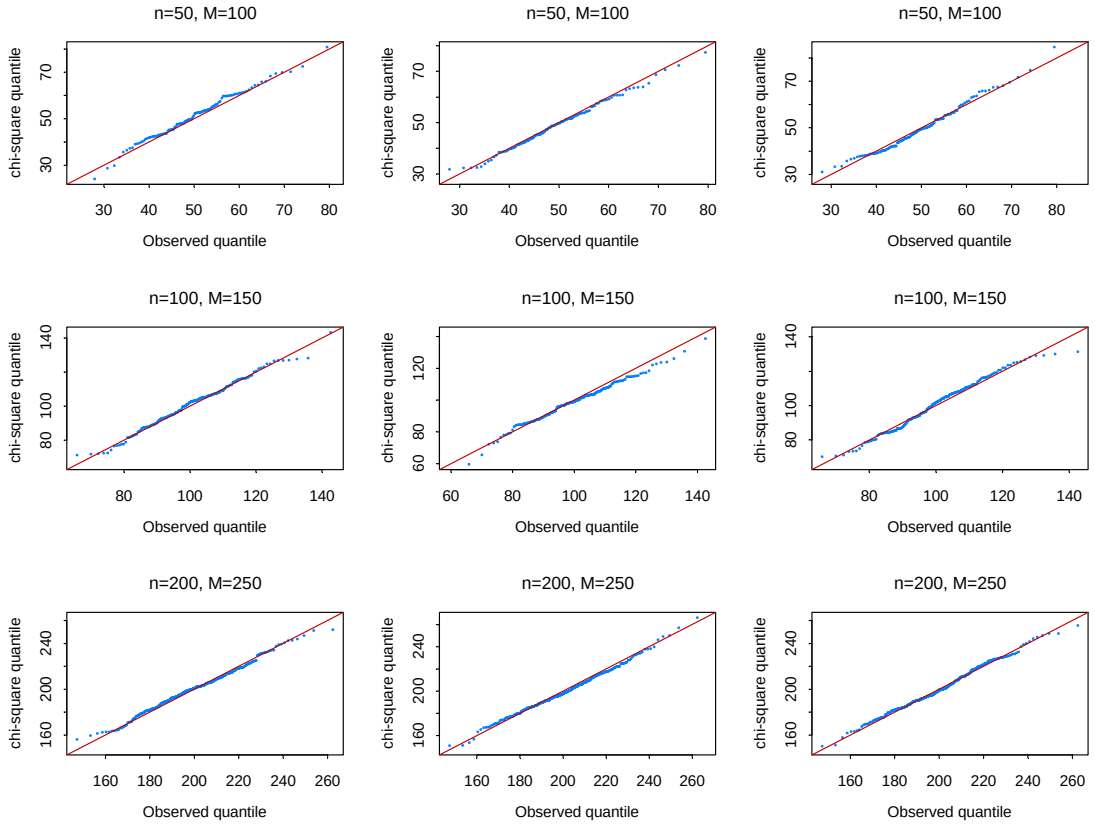
Pound/US exchange rate data: blue = mean, red = mode



Is the posterior distribution close to normal?

Suppose $\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(M)}$ are independent draws from the multivariate distr $p(\alpha_n \mid y_n)$, generated using SIR. Then

$$d_j^2 = (\alpha^{(j)} - \alpha^*)^T (K + G_n)(\alpha^{(j)} - \alpha^*) \overset{iid}{\sim} \chi^2_n$$



Correlations are *not significant.*

Summary Remarks

1. Importance sampling offers a nice clean method for estimation in parameter driven models.

2. Approximation to the likelihood is a non-simulation based procedure which may have great potential especially with large sample sizes and/or large number of explanatory variables.

3. The hybrid approach seems like a good compromise between importance sampling and the approximate likelihood methods.

5. Approximate likelihood and hybrid approaches are amenable to bootstrapping procedures for bias correction.

6. Posterior mode matches posterior mean reasonably well.

7. Approximate likelihood approach may be useful to the problem of structural break detection, the subject of tomorrow's lecture.