The Innovations Algorithm and Parameter Driven Models

Richard A. Davis
Colorado State University
(http://www.stat.colostate.edu/~rdavis/lectures)

Joint work with:
William Dunsmuir, University of New South Wales
Gabriel Rodriguez-Yam, Colorado State University
Ying Wang, Dept of Public Health, W. Virginia

- Generalized state-space models
  - Observation driven
  - Parameter driven
- Recursive one-step ahead prediction algorithms
  - Durbin-Levinson algorithm
  - Innovations algorithm
  - Applications
    - Gaussian likelihood calculations
    - simulation
    - generalized least squares estimation
- Time series of counts
  - Generalized linear models (GLM)
  - Estimating equations (Zeger)
  - MCEM (Chan and Ledolter)
  - Importance sampling
    - Durbin and Koopman
  - Approximation to the likelihood (Davis, Dunsmuir, and Wang)
- Examples

Generalized State-Space Models (observation driven)

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

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

Observation equation:

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

State equation:

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

Forecast density:

$$p(y_{t+1} \mid \mathbf{y}^{(t)}) = \int p(y_{t+1} \mid \alpha_{t+1})\, p(\alpha_{t+1} \mid \mathbf{y}^{(t)})\, d\mu(\alpha_{t+1}).$$

Joint density:

$$p(y_1, \ldots, y_n) = \prod_{t=1}^{n} p(y_t \mid \mathbf{y}^{(t-1)})$$

Examples of observation driven models

Poisson model for time series of counts

Observation equation:

$$p(y_t \mid \alpha_t) = \frac{\alpha_t^{y_t} e^{-\alpha_t}}{y_t!}, \quad y_t = 0, 1, \ldots,$$

State equation:

$$p(\alpha_{t+1} \mid \mathbf{y}^{(t)}) = f(\alpha_{t+1}; \nu_{t+1|t}, \lambda_{t+1|t}),$$

where

$$f(x;\nu,\lambda) = \exp(\nu x - \lambda e^x - \ln\Gamma(\nu) + \nu \ln \lambda), \quad \text{(log-gamma)}$$

and the $\alpha_{t+1|t}$ and $\lambda_{t+1|t}$ are functions of $\mathbf{y}^{(t)}$ (conjugate family of priors).

Remarks:

1. $\nu_{t+1|t} / \lambda_{t+1|t} = E(Y_{t+1} \mid \mathbf{y}^{(t)})$
2. $p(\alpha_t \mid \mathbf{y}^{(t)}) = f(\alpha_t ; \nu_t , \lambda_t)$, $\nu_t = y_t + \nu_{t|t-1}$ and $\lambda_t = 1 + \lambda_{t|t-1}$
3. $Y_t \to 0$ a.s. (Grunwald, et al. (199?) for power steady model.)

Examples of parameter driven models

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.

Generalized State-Space Models (parameter driven)

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

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

Observation equation:

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

State equation:

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

Joint density:

$$\begin{aligned} p(y_1, \ldots, y_n, \alpha_1, \ldots, \alpha_n) &= p(y_n \mid \alpha_n, \boldsymbol{\alpha}^{(n-1)}, \mathbf{y}^{(n-1)})\, p(\alpha_n, \boldsymbol{\alpha}^{(n-1)}, \mathbf{y}^{(n-1)}) \\ &= p(y_n \mid \alpha_n)\, p(\alpha_n \mid \boldsymbol{\alpha}^{(n-1)}, \mathbf{y}^{(n-1)})\, p(\boldsymbol{\alpha}^{(n-1)}, \mathbf{y}^{(n-1)}) \\ &= \cdots \\ &= \left(\prod_{j=1}^{n} p(y_j \mid \alpha_j)\right)\left(\prod_{j=2}^{n} p(\alpha_j \mid \alpha_{j-1})\right) p(\alpha_1) \end{aligned}$$

Parameter driven (cont)

Conditional independence:

$$p(y_1, \ldots, y_n \mid \alpha_1, \ldots, \alpha_n) = \prod_{j=1}^{n} p(y_j \mid \alpha_j)$$

Filtering or posterior density:

$$p(\alpha_t \mid \mathbf{y}^{(t)}) = p(y_t \mid \alpha_t)p(\alpha_t \mid \mathbf{y}^{(t-1)})/p(y_t \mid \mathbf{y}^{(t-1)})$$

Predictive densities:

$$p(\alpha_{t+1} \mid \mathbf{y}^{(t)}) = \int p(\alpha_t \mid \mathbf{y}^{(t)})\, p(\alpha_{t+1} \mid \alpha_t) d\mu(\alpha_t)$$

$$p(y_{t+1} \mid \mathbf{y}^{(t)}) = \int p(y_{t+1} \mid \alpha_{t+1})\, p(\alpha_{t+1} \mid \mathbf{y}^{(t)})\, d\mu(\alpha_{t+1})$$

Examples of parameter driven models

Poisson model for time series of counts

Observation equation:

$$p(y_t \mid \alpha_t) = \frac{e^{\alpha_t y_t} e^{-e^{\alpha_t}}}{y_t!}, \quad y_t = 0, 1, \ldots,$$

State equation: State variables follow a regression model with Gaussian AR(1) noise

$$\alpha_t = \boldsymbol{\beta}^T \mathbf{x}_t + W_t, \quad W_t = \phi W_{t-1} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$$

The resulting transition density of the state variables is

$$p(\alpha_{t+1} \mid \alpha_t) = n(\alpha_{t+1} ; \boldsymbol{\beta}^T \mathbf{x}_{t+1} + \phi\, (\alpha_t - \boldsymbol{\beta}^T \mathbf{x}_t), \sigma^2)$$

Remark: The case $\sigma^2 = 0$ corresponds to a log-linear model with Poisson noise.

Examples of parameter driven models

A stochastic volatility model for financial data (Taylor `86):

Model:

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

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

where $\alpha_t = \log \sigma_t$.

The resulting observation and state transition densities are

$$p(y_t | \alpha_t) = n(y_t ; 0, \exp(2\alpha_t))$$

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

Properties:

- Martingale difference sequence.
- Stationary.
- Strongly mixing at a geometric rate.

Recursive one-step ahead prediction algorithms (Durbin-Levinson)

Durbin-Levinson Algorithm: $\{X_t\}$ is a zero-mean stationary time series with ACF $\gamma(h)$ and write

$$\hat{X}_{t+1} = P_{sp\{1, X_1, \ldots, X_t\}} X_{t+1} = \phi_{t1} X_t + \cdots + \phi_{tt} X_1$$

Then the coefficients $\phi_{t1}, \ldots, \phi_{tt}$ and prediction errors v_{t-1} can be computed recursively from the equations,

$$\phi_{tt} = \left[\gamma(t) - \sum_{j=1}^{t-1} \phi_{t-1,t}\, \gamma(t-j) \right] v_{t-1}^{-1},$$

$$\begin{bmatrix} \phi_{t1} \\ \vdots \\ \phi_{t,t-1} \end{bmatrix} = \begin{bmatrix} \phi_{t-1,1} \\ \vdots \\ \phi_{t-1,t-1} \end{bmatrix} - \phi_{tt} \begin{bmatrix} \phi_{t-1,t-1} \\ \vdots \\ \phi_{t-1,1} \end{bmatrix},$$

and

$$v_t = v_{t-1}(1 - \phi_{tt}^2).$$

Recursive one-step ahead prediction algorithms (Innovations)

Innovations Algorithm (Brockwell and Davis `91): $\{X_t\}$ is a zero-mean time series with ACF $\kappa(i,j)$, then

$$\hat{X}_{t+1} = P_{\mathrm{sp}\{1, X_1, \ldots, X_t\}} X_{t+1} = \theta_{t1}(X_t - \hat{X}_t) + \cdots + \theta_{tt}(X_1 - \hat{X}_1)$$

The coefficients $\theta_{t1}, \ldots, \theta_{tt}$ and prediction errors v_{t-1} can be computed recursively from the equations,

$$v_0 = \kappa(1,1)$$

$$\theta_{t,t-k} = \left[\kappa(t+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{t,t-j} v_j \right] v_{k-1}^{-1}, \;\; k = 0, \ldots, t\text{-}1,$$

and

$$v_t = \kappa(t+1, t+1) - \sum_{j=0}^{t-1} \theta_{t,t-j}^2 v_j.$$

Recursive one-step ahead prediction algorithms (cont)

Remarks:

- D-L expresses one-step predictor in terms of previous *observations*, $X_1, \ldots, X_t$.
- Innovations algorithm expresses one-step predictor in terms of previous *innovations*, $X_1 - \hat{X}_1, \ldots, X_t - \hat{X}_t$, that are uncorrelated.
- If $\{X_t\}$ is an AR(p) process,

$$X_{t+1} = \phi_1 X_t + \cdots + \phi_p X_{t-p} + Z_{t+1}, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

then $(\phi_{t1}, \ldots, \phi_{tt}) = (\phi_1, \ldots, \phi_p, 0, \ldots, 0)$ for $t > p$.

- If $\{X_t\}$ is an MA(q) process

$$X_{t+1} = Z_{t+1} + \theta_1 Z_t + \cdots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

then $(\theta_{t1}, \ldots, \theta_{tt}) = (\theta_{t1}, \ldots, \theta_{tq}, 0, \ldots, 0)$ for all t.

- Innovations algorithm is well adapted for ARMA(p,q) models—only need to apply to MA(q) piece.
- Both D-L and IA can be used for preliminary estimation of ARMA models.

Recursive one-step ahead prediction algorithms — Applications

Likelihood calculation:

Using the IA representation,

$$\hat{X}_t = \theta_{t-1,1}(X_{t-1} - \hat{X}_{t-1}) + \cdots + \theta_{t-1,t-1}(X_1 - \hat{X}_1)$$

we have

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ \theta_{2,2} & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} X_1 - \hat{X}_1 \\ X_2 - \hat{X}_2 \\ X_3 - \hat{X}_3 \\ \vdots \\ X_n - \hat{X}_n \end{bmatrix}$$

$$\mathbf{X}_n = \mathrm{C}_n(\mathbf{X}_n - \hat{\mathbf{X}}_n)$$

By taking covariances of both sides it follows that

$$\Gamma_n = E(\mathbf{X}_n \mathbf{X}'_n) = C_n D_n C'_n, \quad D_n = \mathrm{diag}(v_0, \ldots, v_{\text{n-1}})$$

Recursive one-step ahead prediction algorithms — Applications

Quadratic form:

$$\mathbf{X}'_n \Gamma_n^{-1} \mathbf{X}_n = (\mathbf{X}_n - \hat{\mathbf{X}}_n)' C'_n (C'^{-1}_n D_n^{-1} C_n^{-1}) C_n (\mathbf{X}_n - \hat{\mathbf{X}}_n)$$

$$= (\mathbf{X}_n - \hat{\mathbf{X}}_n)' D_n^{-1} (\mathbf{X}_n - \hat{\mathbf{X}}_n)$$

$$= \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / v_{t-1}$$

Determinant:

$$\det(\Gamma_n) = \det(C_n D_n C'_n) = v_0 \cdots v_{\text{n-1}}$$

Gaussian likelihood:

$$L(\Gamma_n) = (2\pi)^{-n/2} (v_0 \cdots v_{n-1})^{-1/2} \exp\{-1/2 \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 / v_{t-1}\}$$

Simulation: If $\{Z_t\}$ ~ iid N(0,1), put $X_t = v_{t-1}^{-1/2} Z_t + \theta_{t-1,1} v_{t-2}^{-1/2} Z_{t-1} + \cdots + \theta_{t-1,t-1} v_0^{-1/2} Z_1$.

Then $\mathbf{X}_n = (X_1, \cdots, X_n)' = C'_n D_n^{-1/2} \mathbf{Z}_n$

has covariance matrix Γ_n.

Application to Regression With Time Series Errors

Data: $Y_1, \ldots, Y_n$

Regression model:

$$Y_t = \mathbf{x}_t^T \boldsymbol{\beta} + W_t, \;\; t = 1, \ldots, n,$$

$$\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})' \text{ (explanatory variables at time t)}$$

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' \quad \text{(regression coefficients)}$$

$$\{W_t\} \sim \text{(stationary time series, e.g., ARMA process)}$$

or in matrix notation

$$\mathbf{Y}_n = \mathrm{X}\boldsymbol{\beta} + \mathbf{W}_n$$

Generalized least squares: Minimize

$$(\mathbf{Y}_n - \mathrm{X}\boldsymbol{\beta})' \Gamma_n^{-1} (\mathbf{Y}_n - \mathrm{X}\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$, where Γ_n is the covariance matrix for $\mathbf{W}_n$. The GLS estimator is

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\mathrm{X}'\Gamma_n^{-1}\mathrm{X})^{-1}\mathrm{X}'\Gamma_n^{-1}\mathbf{Y}_n$$

Application to Regression With Time Series Errors (see B&D lite `02)

By transforming the model

$$\Gamma_n^{-1/2}\mathbf{Y}_n = \Gamma_n^{-1/2}\mathrm{X}\boldsymbol{\beta} + \Gamma_n^{-1/2}\mathbf{W}_n$$

$$\mathbf{Y}_n^* = \mathrm{X}_n^*\boldsymbol{\beta} + \mathbf{W}_n^*,$$

we see that

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \hat{\boldsymbol{\beta}}^*_{\mathrm{OLS}} = (\mathrm{X}^{*\prime}\mathrm{X}^*)^{-1}\mathrm{X}^{*\prime}\mathbf{Y}_n^* \quad \text{and} \quad \mathrm{cov}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) = (\mathrm{X}^{*\prime}\mathrm{X}^*)^{-1}.$$

But,

$$\mathbf{Y}_n^* = D_n^{-1/2}(\mathbf{Y}_n - \hat{\mathbf{Y}}_n)$$

$$\mathrm{X}_n^* = D_n^{-1/2}(\mathrm{X}_n - \hat{\mathrm{X}}_n)$$

which can be computed by applying the innovations algorithm to $\mathbf{Y}_{\mathrm{n}}$ and each column of the design matrix X.

Profile likelihood: Set $U_t = Y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$, then

$$L(\Gamma_n) = (2\pi)^{-n/2}(v_0\cdots v_{n-1})^{-n/2}\exp\{-1/2\sum_{t=1}^{n}(U_t - \hat{U}_t)^2/v_{t-1}\}$$

Time Series of Counts—Notation and Setup

Count data: $Y_1, \ldots, Y_n$

Regression (explanatory) variable: $\mathbf{x_t}$

Model: Distribution of the Y_t given $\mathbf{x_t}$ and a stochastic process α_t are indep Poisson distributed with mean

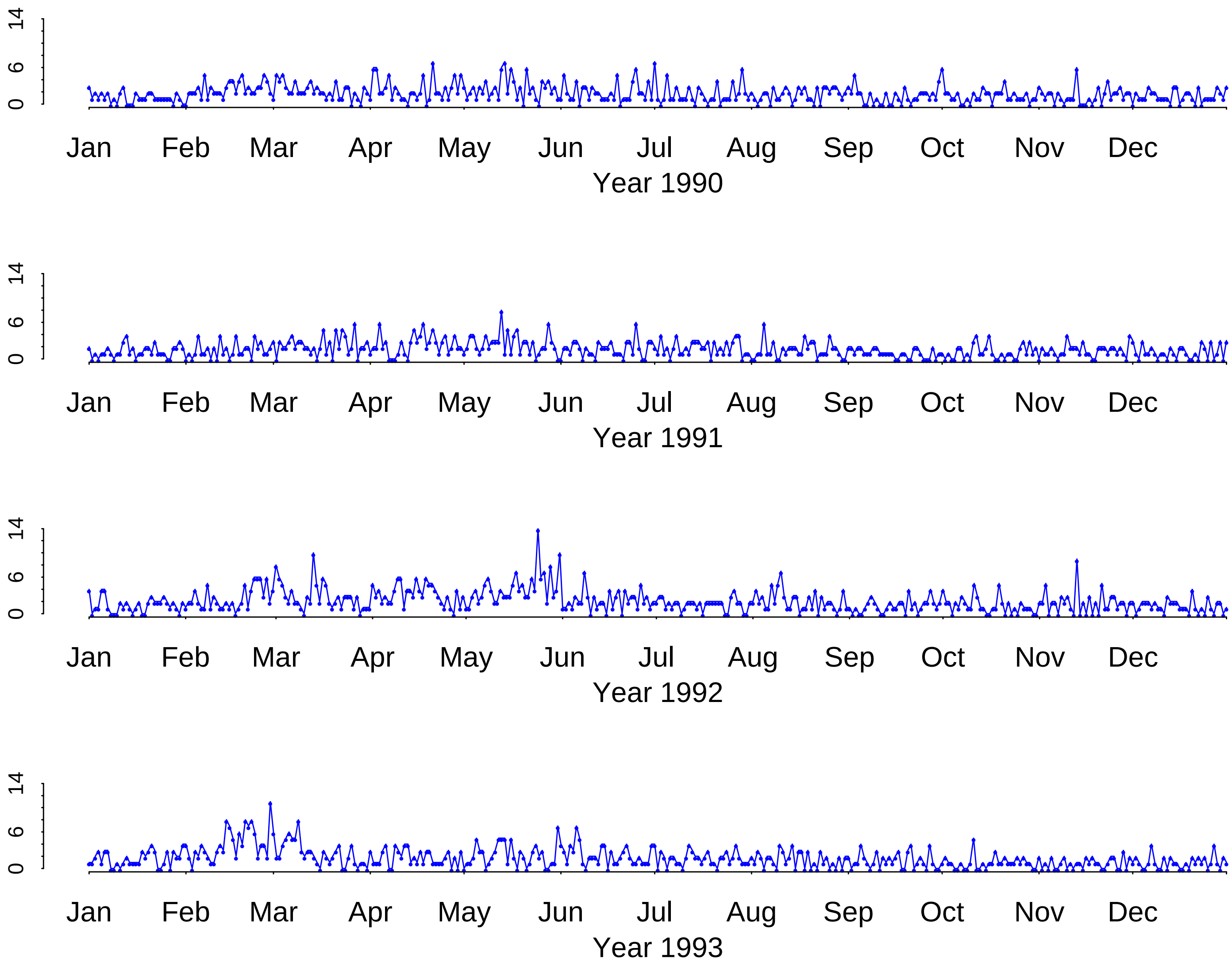
$$\mu_t = \exp(\mathbf{x}_t^T \beta + \alpha_t).$$

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

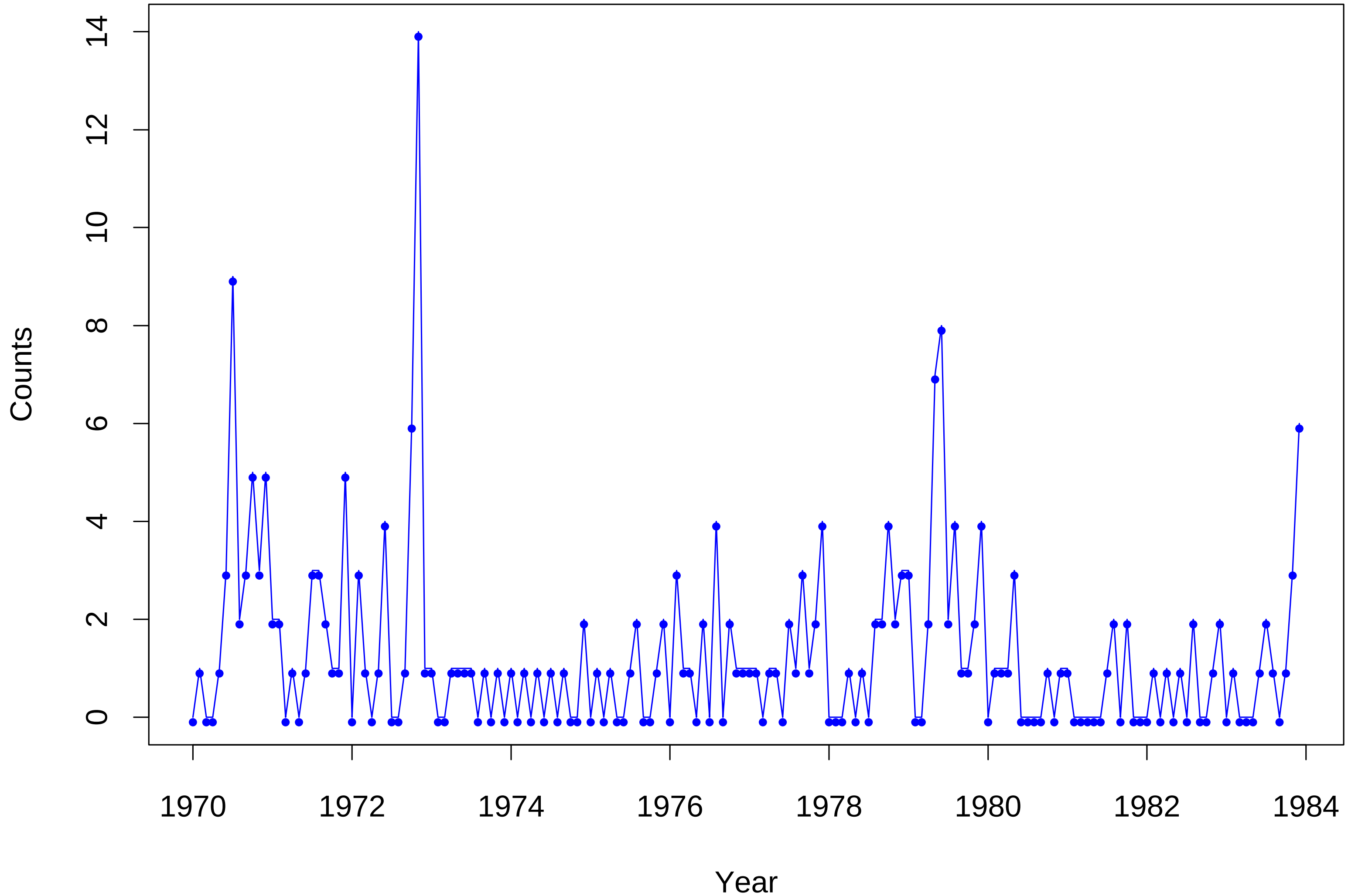
Note: $\alpha_t = 0$ corresponds to standard Poisson regression model.

Primary objective: Inference about β.

Example: Daily Asthma Presentations (1990:1993)



Example: Monthly Polio Counts in USA (Zeger 1988)



Parameter-Driven Model for the Mean Function μ_t

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

$$\log \mu_t = \mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t \ ,$$

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

$$E(Y_t) = \exp(\mathbf{x}_t^T \boldsymbol{\beta})E\exp(\nu_t) = \exp(\mathbf{x}_t^T \boldsymbol{\beta}), \ \text{if} \ E\exp(\alpha_t) = 1.$$

Disadvantages:

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

Remark: See Davis, Dunsmuir, and Wang (1999) for testing of the existence of a latent process and estimating its ACF.

Estimation Methods — GLM estimation

Model: $Y_t \mid \alpha_t\, ,\, \mathbf{x}_t \sim Pois(\exp(\mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t\,))$.

GLM log-likelihood:

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

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

Assumptions on regressors:

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

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

Theory of GLM Estimation in Presence of Latent Process

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

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

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 $\mathbf{x}_t$ can depend on the sample size n.

When Does CLT Apply?

Conditions on the regressors hold for:

1. Trend functions.

$$\mathbf{x}_{\mathbf{n}t} = \mathbf{f}(t/n)$$

where $\mathbf{f}$ is a continuous function on [0,1]. In this case,

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

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

Remark. $\mathbf{x}_{\mathbf{nt}}$ = (1, t/n) corresponds to linear regression and works. However $\mathbf{x}_{\mathbf{t}}$ = (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 Model for Polio Data

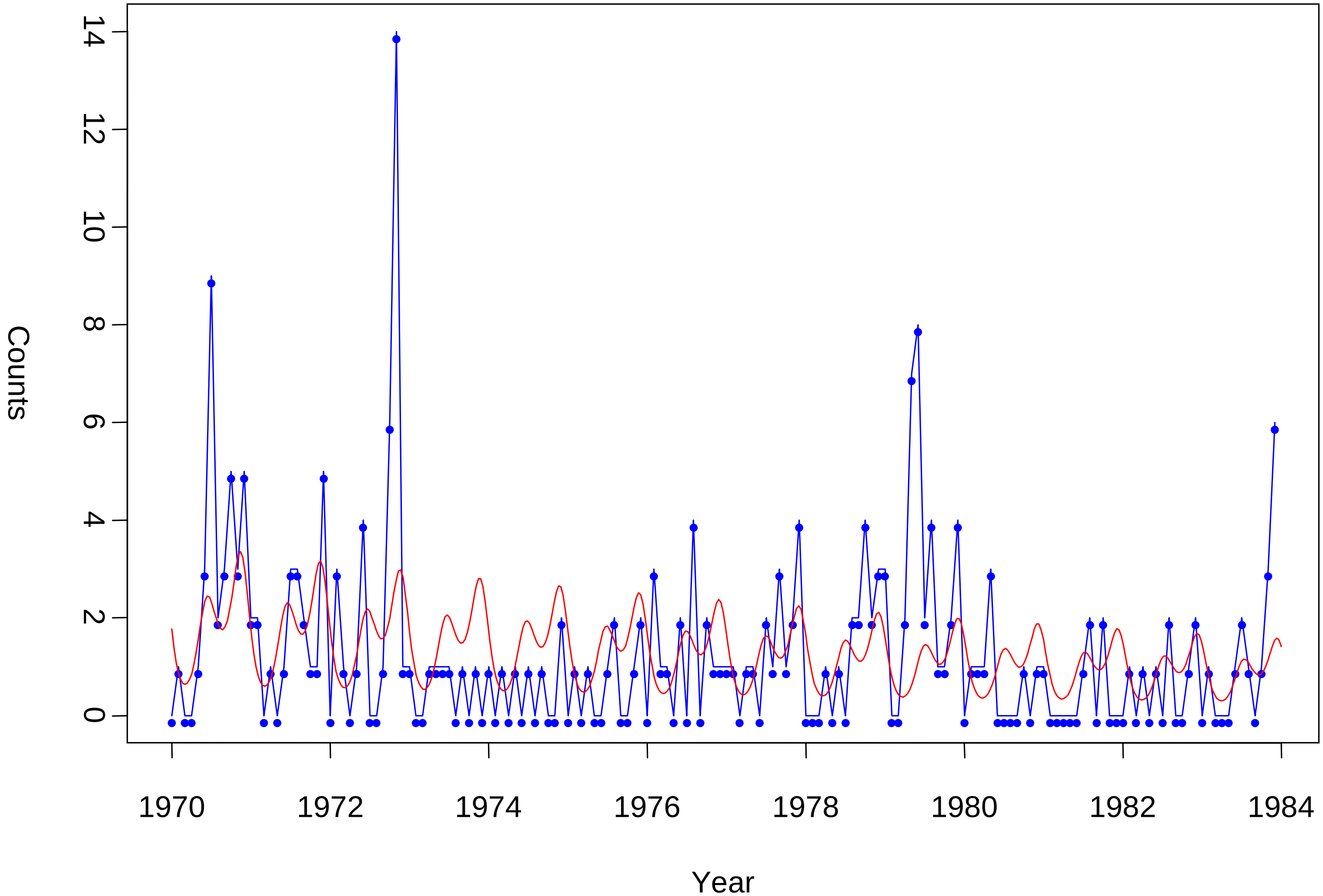
Assume the $\{\alpha_t\}$ follows a log-normal AR(1), where

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

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.d.
Intercept 0.17	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
$\cos(2\pi t/12)$	-0.11	0.16	-0.15	.097	.157	-.145	.144
$\sin(2\pi t/12)$	-.048	0.17	-0.53	.109	.168	-.531	.168
$\cos(2\pi t/6)$	0.20	0.14	.169	.098	.122	.167	.123
$\sin(2\pi t/6)$	-0.41	0.14	-.432	.101	.125	-.440	.125

Polio Data With Estimated Regression Function



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

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 $Q(\psi | \psi^{(k)}) = E(L(\psi; \mathbf{y}_n, \boldsymbol{\alpha}_n) \mid \mathbf{y}_n, \psi^{(k)})$,

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

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 $p(\boldsymbol{\alpha}_n \mid \mathbf{y}_n, \psi^{(k)})$ in the E-step. Chan and Ledolter use a Gibbs sampler for this.

Estimation Methods — Importance Sampling (Durbin and Koopman)

Model:

$$Y_t \mid \alpha_t , \mathbf{x}_t \sim Pois(\exp(\mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t))$$

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

Relative Likelihood: Let $\psi=(\boldsymbol{\beta}, \phi, \sigma^2)$ and suppose $g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0)$ is an approximating joint density for $\mathbf{Y}_n= (Y_1, . . . , Y_n)'$ and $\boldsymbol{\alpha}_n= (\alpha_1, . . . , \alpha_n)'$.

$$
\begin{aligned}
L(\psi) &= \int p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n) p(\boldsymbol{\alpha}_n) d\boldsymbol{\alpha}_n \\
&= \int \frac{p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n) p(\boldsymbol{\alpha}_n)}{g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0)} g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0) d\boldsymbol{\alpha}_n \\
&= \int \frac{p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n) p(\boldsymbol{\alpha}_n)}{g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0)} g(\boldsymbol{\alpha}_n \mid \mathbf{y}_n; \psi_0) g(\mathbf{y}_n; \psi_0) d\boldsymbol{\alpha}_n
\end{aligned}
$$

$$\frac{L(\psi)}{L_g(\psi_0)} = \int \frac{p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n) p(\boldsymbol{\alpha}_n)}{g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0)} g(\boldsymbol{\alpha}_n \mid \mathbf{y}_n; \psi_0) d\boldsymbol{\alpha}_n$$

Importance Sampling (cont)

$$\frac{L(\psi)}{L_g(\psi_0)} = \int \frac{p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n)\, p(\boldsymbol{\alpha}_n)}{g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0)}\, g(\boldsymbol{\alpha}_n \mid \mathbf{y}_n; \psi_0)\, d\boldsymbol{\alpha}_n$$

$$= E_g\left[\frac{p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n)\, p(\boldsymbol{\alpha}_n)}{g(\mathbf{y}_n, \boldsymbol{\alpha}_n; \psi_0)} \mid \mathbf{y}_n; \psi_0) \right]$$

$$\sim \frac{1}{N} \sum_{j=1}^{N} \frac{p(\mathbf{y}_n \mid \boldsymbol{\alpha}_n^{(j)})\, p(\boldsymbol{\alpha}_n^{(j)})}{g(\mathbf{y}_n, \boldsymbol{\alpha}_n^{(j)}; \psi_0)},$$

where $\{\boldsymbol{\alpha}_n^{(j)};\, j = 1, \ldots, N\} \sim \text{iid } \mathrm{g}(\boldsymbol{\alpha}_n \mid \mathbf{y}_n; \psi_0)$.

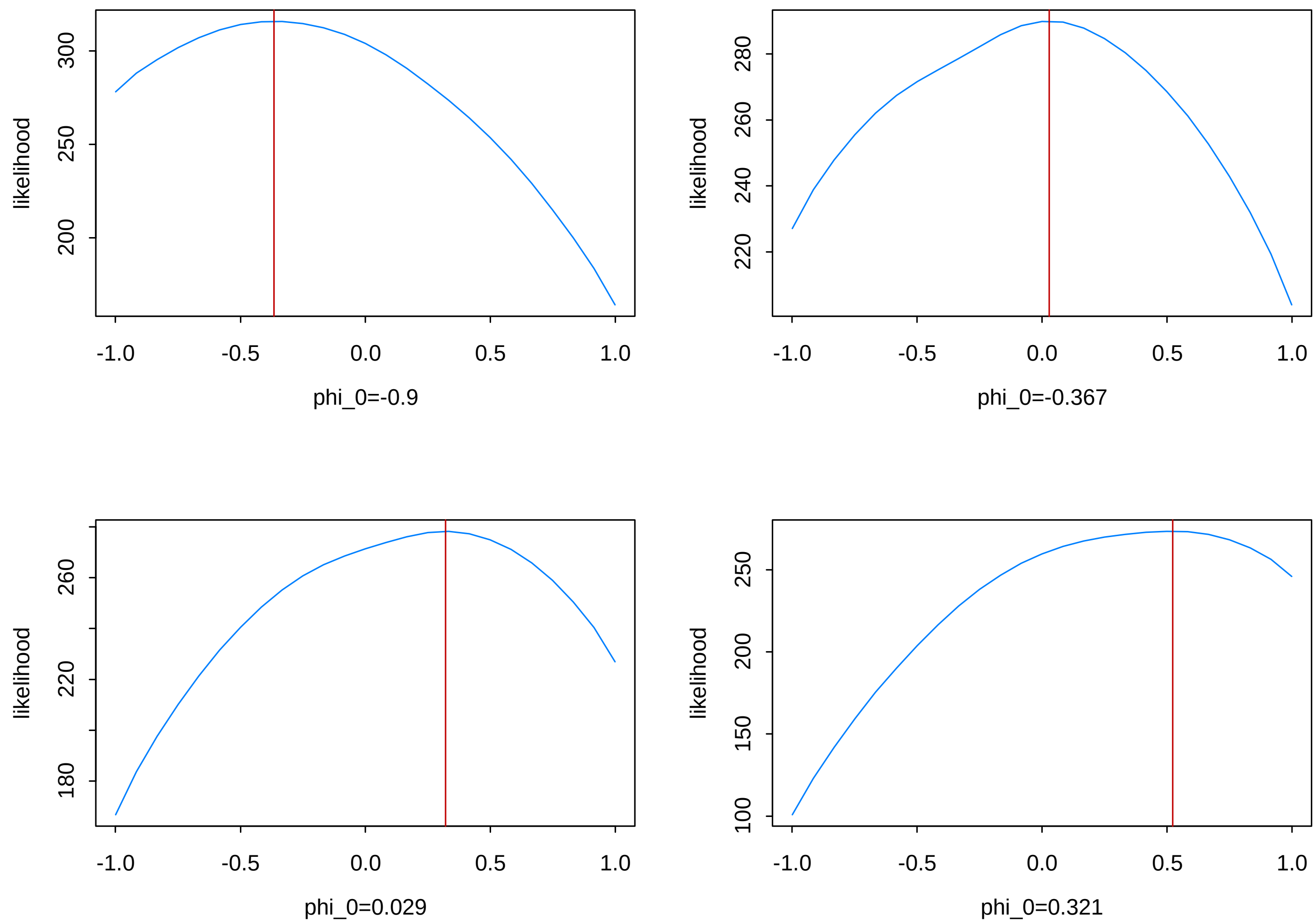
Notes:

- 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 $\hat{\psi}$.

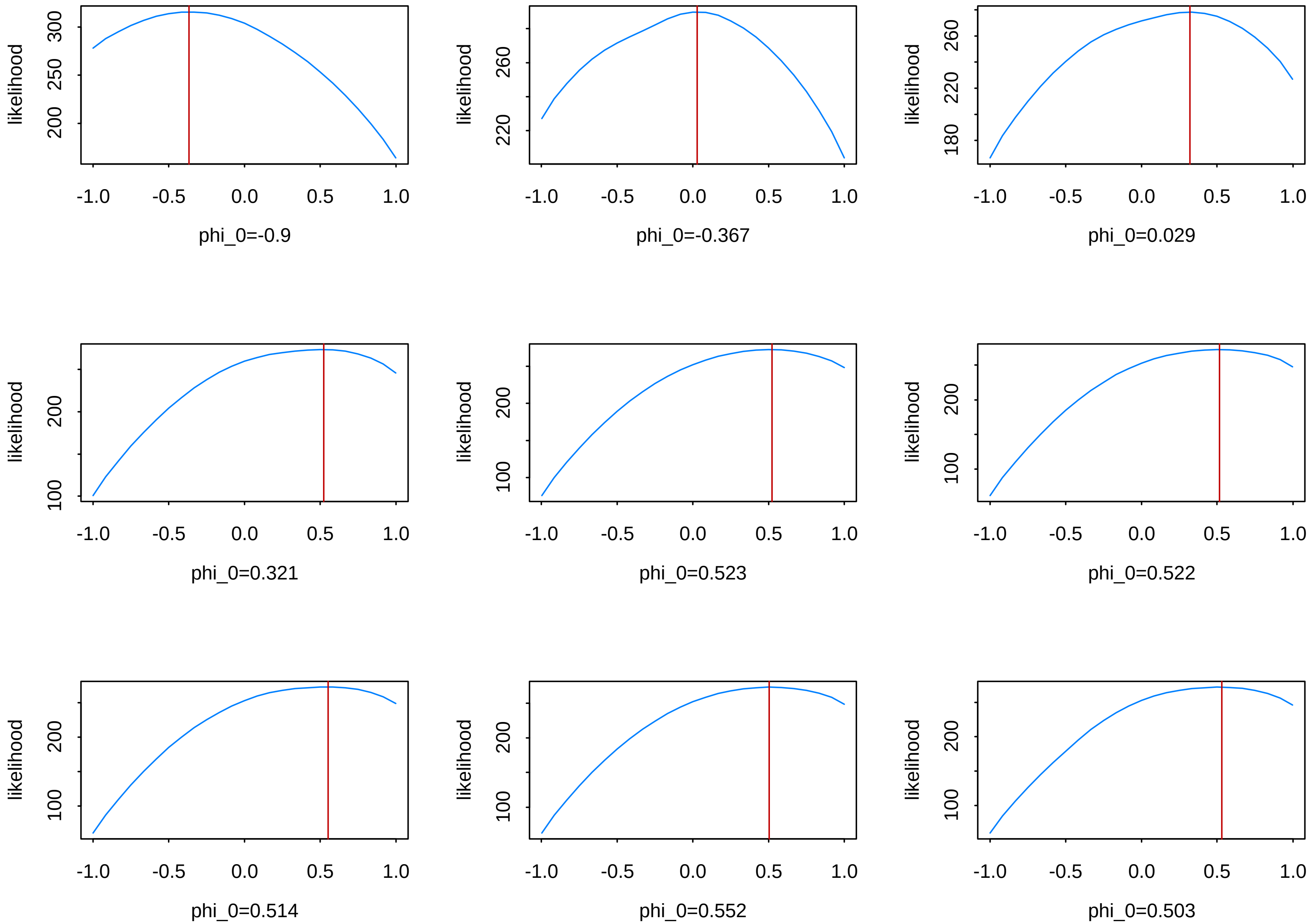
Importance Sampling — example

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

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



Simulation example: $Y_t \mid \alpha_t \sim Pois(\exp(.7 + \alpha_t)), \quad \phi = .5, \sigma^2 = .3, n = 200, N = 1000$



Importance Sampling (cont)

Choice of *importance density* g:

Durbin and Koopman suggest a linear state-space approximating model

$$Y_t = \mu_t + \mathbf{x}_t^T \boldsymbol{\beta} + \alpha_t + Z_t\,, \quad Z_t \sim N(0, H_t),$$

with

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

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

With this choice of approximating model, it turns out that

$$\mathrm{g}(\boldsymbol{\alpha}_n \mid \mathbf{y}_n; \psi_0) \sim N(\Gamma_n^{-1} \tilde{\mathbf{y}}_n, \Gamma_n^{-1}),$$

where

$$\tilde{\mathbf{y}}_n = \mathbf{y}_n - e^{\mathrm{X}\boldsymbol{\beta} + \hat{\boldsymbol{\alpha}}_n} + e^{\mathrm{X}\boldsymbol{\beta} + \hat{\boldsymbol{\alpha}}_n} \hat{\boldsymbol{\alpha}}_n,$$
$$\Gamma_n = \mathrm{diag}(e^{\mathrm{X}\boldsymbol{\beta} + \hat{\boldsymbol{\alpha}}_n}) + (E(\boldsymbol{\alpha}_n \boldsymbol{\alpha}'_n))^{-1}.$$

Importance Sampling (cont)

Components required in the calculation.

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

Importance Sampling (cont)

Details.
$$(E(\boldsymbol{\alpha}_n\boldsymbol{\alpha}'_n))^{-1} = \sigma^{-2}\begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & \cdots & 0 \\ 0 & -\phi & 1+\phi^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\phi^2 \end{pmatrix}$$

$$\Gamma_n = \operatorname{diag}(e^{\hat{\boldsymbol{\alpha}}+X\boldsymbol{\beta}}) + \sigma^{-2}\begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & \cdots & 0 \\ 0 & -\phi & 1+\phi^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\phi^2 \end{pmatrix}.$$

This is the covariance function of a 1-dependent sequence, so that $\Gamma_n = C_n D_n C'_n$, where

$$C_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{1,1} & 1 & 0 & \cdots & 0 \\ 0 & \theta_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Importance Sampling (cont)

It follows that

$$\tilde{\mathbf{y}}'_n \Gamma_n^{-1} \tilde{\mathbf{y}}_n = \sum_{t=1}^{n} (\tilde{y}_t - \hat{\tilde{y}}_t)^2 / v_{t-1}$$

and

$$\begin{aligned}\Gamma_n^{-1}\tilde{\mathbf{y}}_n &= C'^{-1}_n D_n^{-1} C_n^{-1} C_n (\tilde{\mathbf{y}}_n - \hat{\tilde{\mathbf{y}}}_n) \\ &= C'^{-1}_n (D_n^{-1}(\tilde{\mathbf{y}}_n - \hat{\tilde{\mathbf{y}}}_n))\end{aligned}$$

which can be solved for the vector $\Gamma_n^{-1}\tilde{\mathbf{y}}_n$ via the recursion

$$C'_n \Gamma_n^{-1}\tilde{\mathbf{y}}_n = D_n^{-1}(\tilde{\mathbf{y}}_n - \hat{\tilde{\mathbf{y}}}_n).$$

All of these calculations can be carried out quickly using the *innovations algorithm*.

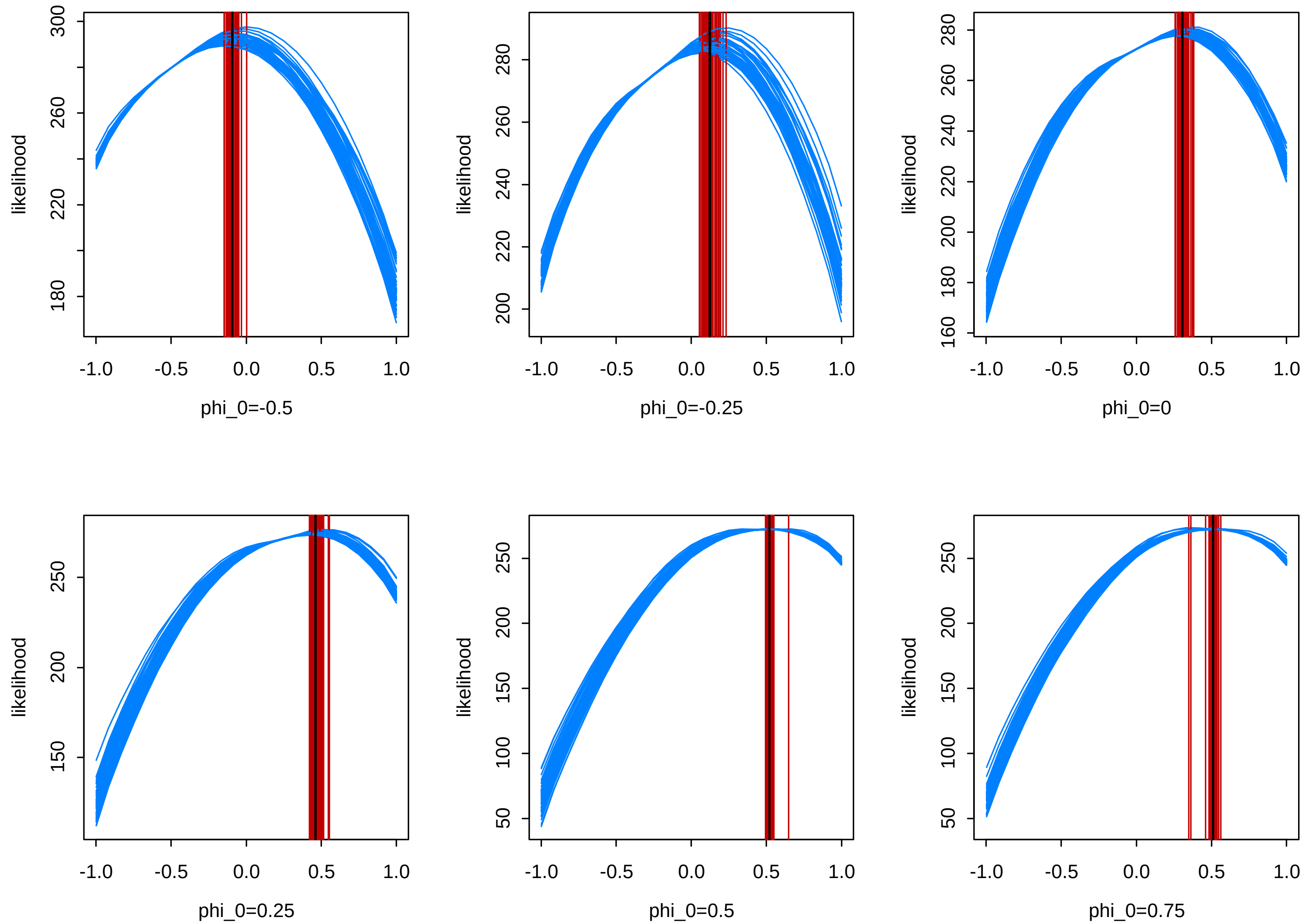
To simulate from $N(\mathbf{0}, \Gamma_n^{-1})$ note that

$$\mathbf{U}_n = C'^{-1}_n D_n^{-1} \mathbf{Z}_n,$$

where $\mathbf{Z}_n \sim \mathrm{N}(0,1)$, has covariance matrix Γ_n^{-1}.

Importance Sampling — example

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



Estimation Methods — Approximation to the likelihood

Joint density function:

$$p(\mathbf{y}_n, \boldsymbol{\alpha}_n) \propto \frac{\det(G)^{1/2}}{\prod_{t=1}^{n} y_t!} \exp\{-(\mathbf{y}_n^T(\boldsymbol{\alpha}_n + \mathrm{X}\boldsymbol{\beta}) - e^{\mathbf{1}^T(\boldsymbol{\alpha}_n + \mathrm{X}\boldsymbol{\beta})} - \boldsymbol{\alpha}_n^T G_n \boldsymbol{\alpha}_n / 2\},$$

where $G_n^{-1} = E(\boldsymbol{\alpha}_n^T \boldsymbol{\alpha}_n)$.

Conditional density function:

$$p(\boldsymbol{\alpha}_n \mid \mathbf{y}_n) \propto \exp\{-\mathbf{y}_n^T \boldsymbol{\alpha}_n - e^{\mathbf{1}^T(\boldsymbol{\alpha}_n + \mathrm{X}\boldsymbol{\beta})} - \boldsymbol{\alpha}_n^T G_n \boldsymbol{\alpha}_n / 2\},$$

which, by expanding the term, $e^{\mathbf{1}^T(\boldsymbol{\alpha}_n + X\boldsymbol{\beta})}$ in a neighborhood of α_n^*, and ignoring third-order + terms yields the approximation

$$p_a(\boldsymbol{\alpha}_n \mid \mathbf{y}_n) \propto \exp\{-(\mathbf{y}_n^T(\boldsymbol{\alpha}_n + \mathrm{X}\boldsymbol{\beta}) - e^{\mathbf{1}^T(\boldsymbol{\alpha}_n^* + \mathrm{X}\boldsymbol{\beta})} + (\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*_n)^T e^{\boldsymbol{\alpha}^*_n + X\boldsymbol{\beta}}$$
$$+ \frac{1}{2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^*)^T \operatorname{diag}(e^{\boldsymbol{\alpha}_n^* + \mathrm{X}\boldsymbol{\beta}})(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^*) - \boldsymbol{\alpha}_n^T G_n \boldsymbol{\alpha}_n / 2\}.$$

Estimation Methods — Approximation to the likelihood

After simplification, we find

$$p_{\mathrm{a}}(\boldsymbol{\alpha}_n \mid \mathbf{y}_n) \propto \exp\{-(\mathbf{y}_n^T(\boldsymbol{\alpha}_n + \mathrm{X}\boldsymbol{\beta}) - e^{\mathbf{1}^T(\boldsymbol{\alpha}_n^* + \mathrm{X}\boldsymbol{\beta})} + (\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^*)^T e^{\boldsymbol{\alpha}_n^* + \mathrm{X}\boldsymbol{\beta}}$$

$$+\frac{1}{2}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^*)^T \operatorname{diag}(e^{\boldsymbol{\alpha}_n^* + \mathrm{X}\boldsymbol{\beta}})(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n^*) - \boldsymbol{\alpha}_n^T G_n \boldsymbol{\alpha}_n / 2\}.$$

$$\sim N(\Gamma_n^{-1}\tilde{\mathbf{y}}_n, \Gamma_n^{-1})$$

Approximate likelihood:

$$p_a(\mathbf{y}_n;\psi) = \frac{p(\mathbf{y}_n, \boldsymbol{\alpha}_n)}{p_a(\boldsymbol{\alpha}_n \mid \mathbf{y}_n)} \propto \frac{\det(G_n)^{1/2}}{\det(\Gamma_n)^{1/2}} \exp\{\mathbf{y}_n^T \mathrm{X}\boldsymbol{\beta} + .5\tilde{\mathbf{y}}_n^T \Gamma_n^{-1} \tilde{\mathbf{y}}_n\},$$

$$\tilde{\mathbf{y}}_n = \mathbf{y}_n - \exp\{\mathrm{X}\boldsymbol{\beta}\}\exp\{\boldsymbol{\alpha}_n^*\} + \exp\{\boldsymbol{\alpha}_n^*\}\exp\{\mathrm{X}\boldsymbol{\beta}\}\boldsymbol{\alpha}_n^*$$

(component-wise multiplication for vectors)

Note: We actually expand the joint density for $\mathbf{Y}_{\mathrm{n}}$ and α_{n} in a neighborhood of α^*.

Estimation Methods — Approximation to the likelihood

Implementation:

1. Let $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*(\psi)$ be the converged value of $\boldsymbol{\alpha}^{(j)}(\psi)$, where

$$\boldsymbol{\alpha}^{(j+1)}(\psi) = \Gamma_n^{-1}\tilde{\mathbf{y}}_n(\psi)$$

2. Maximize $p_a(\mathbf{y}_n;\psi)$ with respect to ψ.

Simulation Results

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

Estimation methods:

- Importance sampling (N=1000, ψ_0 updated a maximum of 10 times)

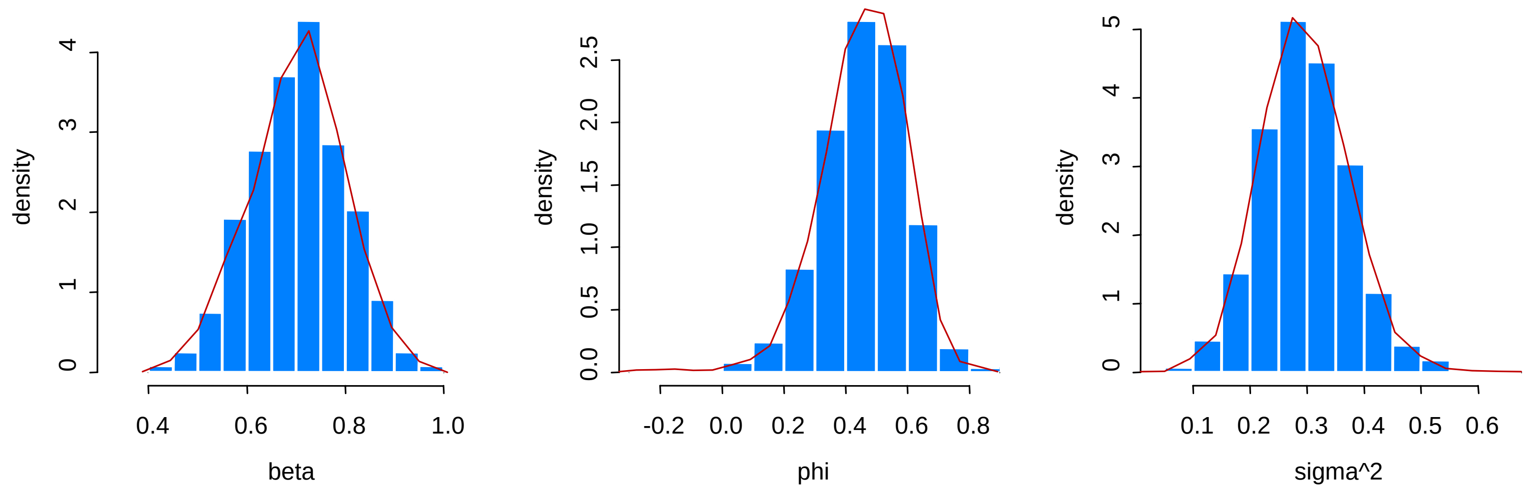
	beta	phi	sigma2
mean	0.6982	0.4718	0.3008
std	**0.1059**	**0.1476**	**0.0899**

- Approximation to likelihood

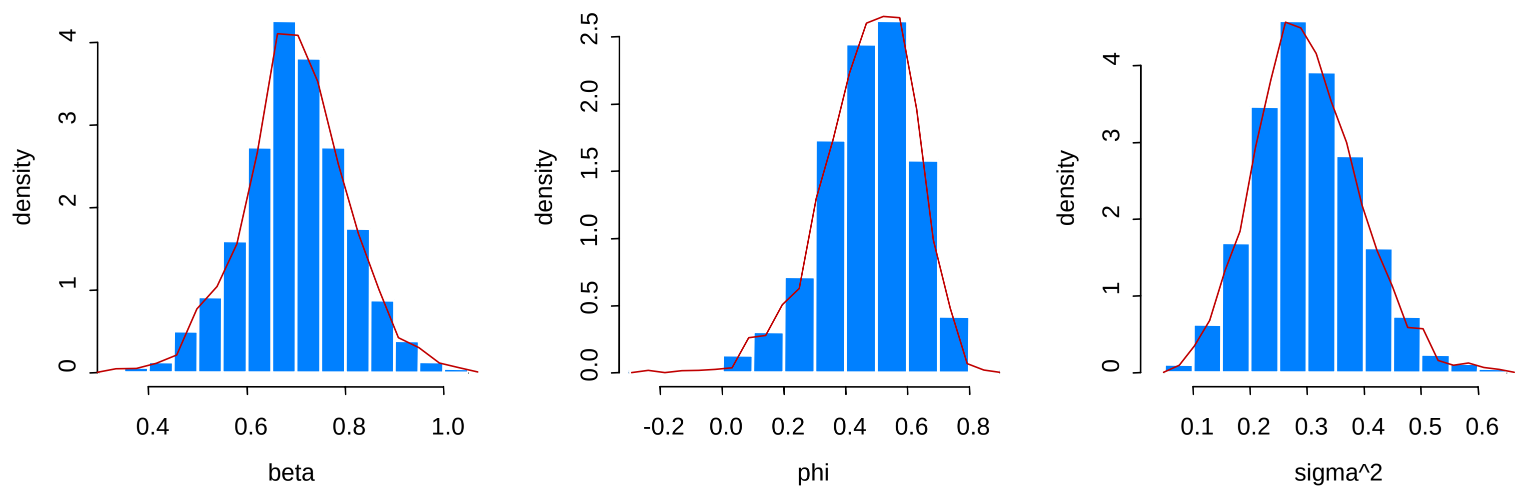
	beta	phi	sigma2
mean	0.7036	0.4579	0.2962
std	**0.0951**	**0.1365**	**0.0784**

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



Application to Model Fitting for the Polio Data

Model for $\{\alpha_t\}$:

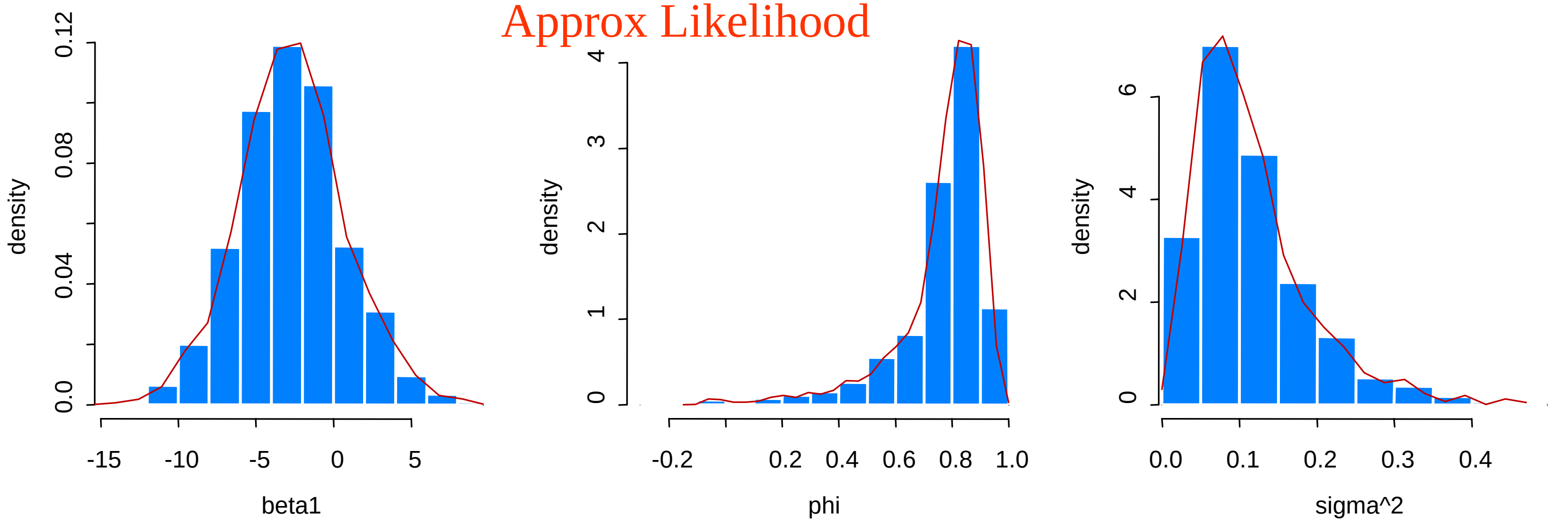
$$\alpha_t = \phi\alpha_{t-1}+\varepsilon_t\,, \quad \{\varepsilon_t\}\sim\text{IID } N(0, \sigma^2).$$

- Importance sampling (ψ_0 updated 5 times for each N=100, 500, 1000,)
- Simulation based on 1000 replications and the fitted AL model.

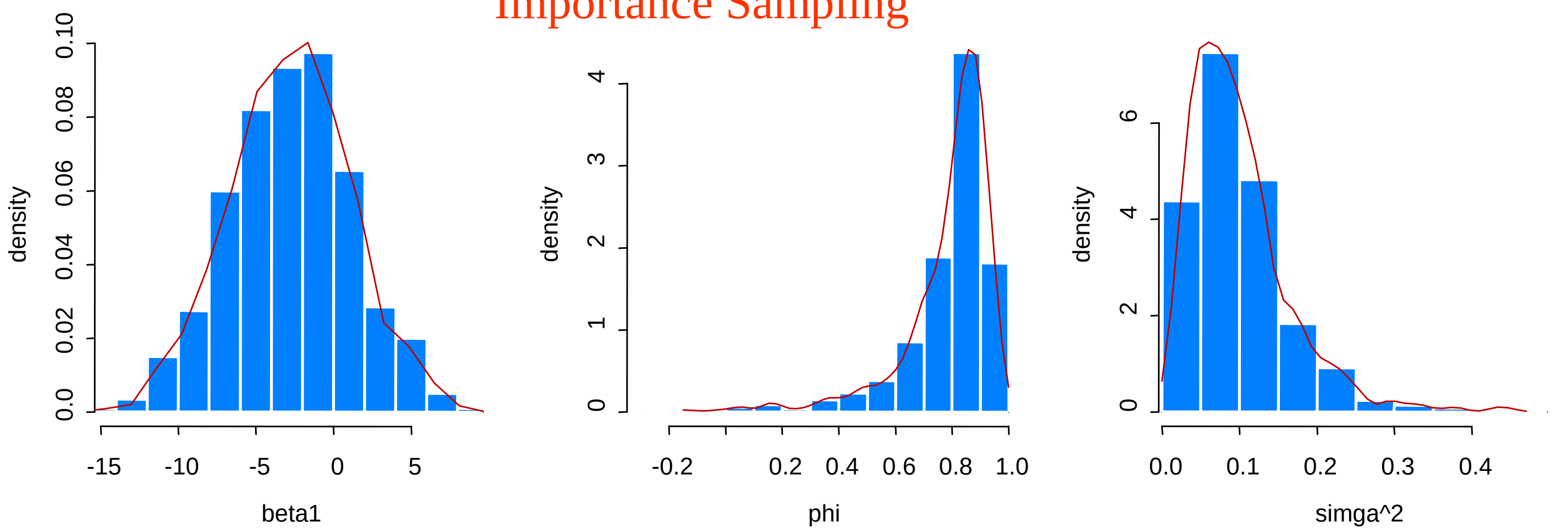
	Import Sampling			Approx Like			GLM	
		Simulation			Simulation			
	$\hat{\beta}_{IS}$	Mean	SD	$\hat{\beta}_{AL}$	Mean	SD	$\hat{\beta}_{GLM}$	SD
Intercept	**0.203**	0.223	0.381	**0.202**	0.210	0.343	**.207**	0.078
Trend($\times 10^{-3}$)	**-2.675**	-2.778	3.979	**-2.690**	-2.720	3.415	**-4.18**	1.400
$\cos(2\pi t/12)$	**0.110**	0.103	0.124	**0.113**	0.111	0.123	**-.152**	0.097
$\sin(2\pi t/12)$	**-0.456**	-0.456	0.151	**-0.454**	-0.454	0.143	**-.532**	0.109
$\cos(2\pi t/6)$	**0.399**	0.401	0.123	**0.396**	0.400	0.114	**.169**	0.098
$\sin(2\pi t/6)$	**0.015**	0.024	0.118	**0.016**	0.012	0.110	**-.432**	0.101
ϕ	**0.865**	0.777	0.198	**0.845**	0.764	0.165		
σ^2	**0.088**	0.100	0.068	**0.104**	0.114	0.075		

Application to Model Fitting for the Polio Data (cont)

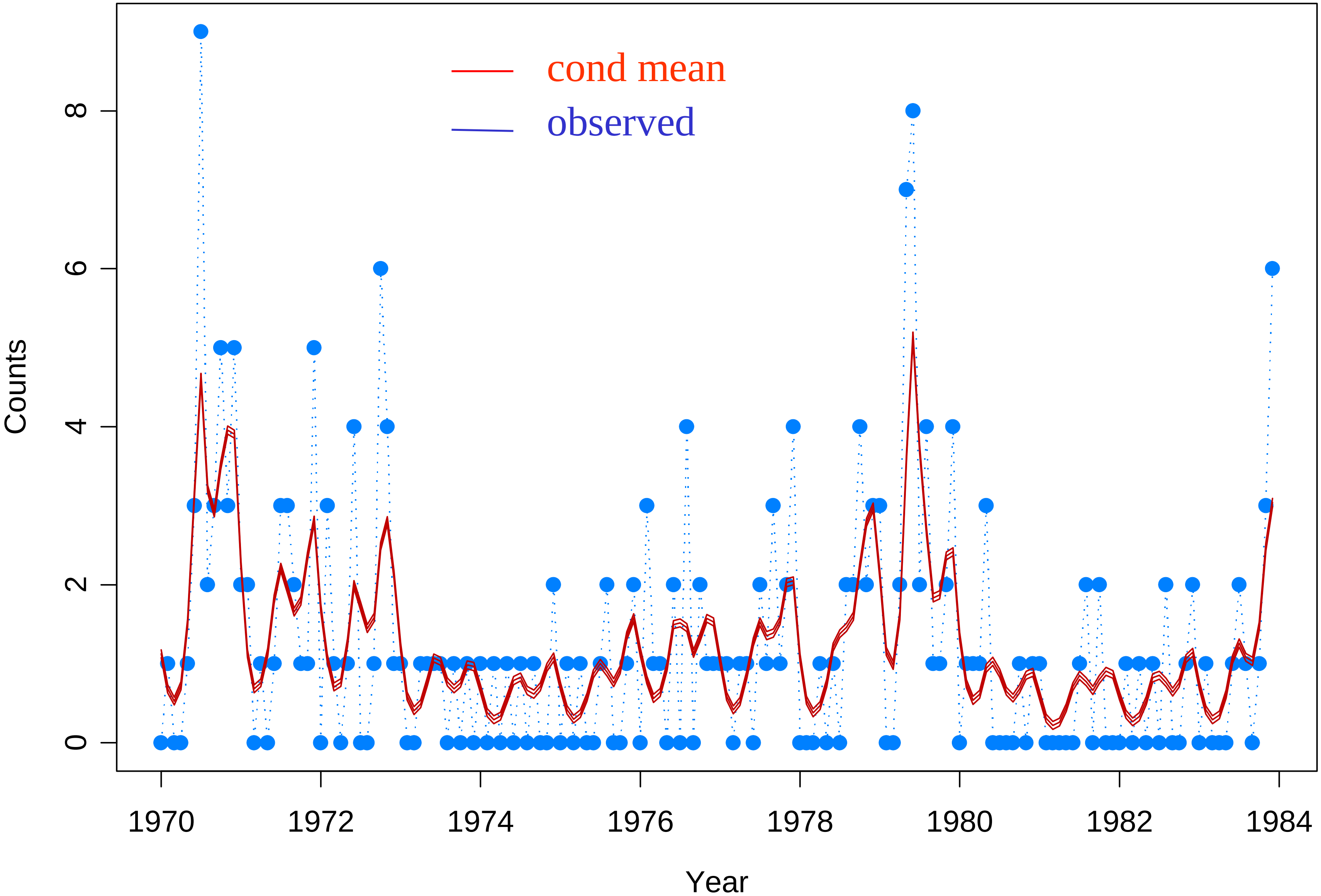
Approx Likelihood



Importance Sampling



Polio Data: observed and conditional mean (approx like)



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

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

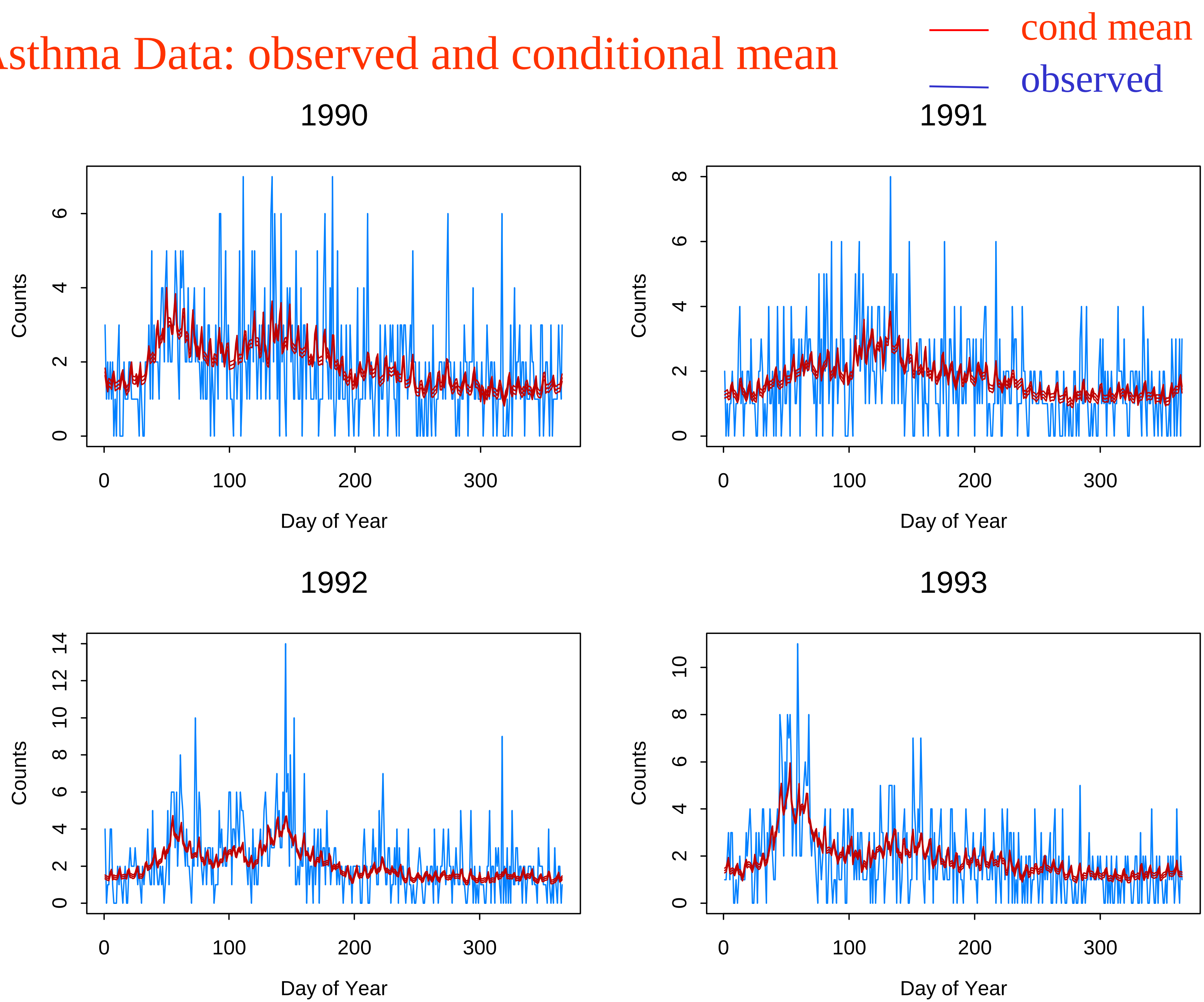
  where $B(2.5,5)$ is the beta function and T_{ij} is the start of the j^{th} school term in year i.

- 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—(IS & AL)

Term	IS	AL	Mean	SD
Intercept	0.590	0.591	0.593	.0658
Sunday effect	0.138	0.138	0.139	.0531
Monday effect	0.229	0.231	0.230	.0495
$\cos(2\pi t/365)$	-0.218	-0.218	-0.217	.0415
$\sin(2\pi t/365)$	0.200	0.179	0.181	.0437
Term 1, 1990	0.188	0.198	0.194	.0638
Term 2, 1990	0.183	0.130	0.129	.0664
Term 1, 1991	0.080	0.075	0.070	.0733
Term 2, 1991	0.177	0.164	0.157	.0665
Term 1, 1992	0.223	0.221	0.214	.0667
Term 2, 1992	0.243	0.239	0.237	.0620
Term 1, 1993	0.379	0.397	0.394	.0625
Term 2, 1993	0.127	0.111	0.108	.0682
Humidity $H_t/20$	0.009	0.010	0.007	.0032
NO_2 max	-0.125	-0.107	-0.108	.0347
AR(1), ϕ	0.385	0.788	0.468	.3790
σ^2	0.053	0.010	0.018	.0153

Asthma Data: observed and conditional mean



Summary Remarks

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

2. The innovations algorithm allows for quick implementation of importance sampling. Extends easily to higher-order AR structure.

3. Relative likelihood approach is a one-sample based procedure.

4. 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. .