



Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

Modeling The Variance of a Time Series

Peter Bloomfield

Department of Statistics
North Carolina State University

July 31, 2009 / Ben Kedem Symposium



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- 1 Introduction
- 2 Time Series Models
 - First Wave
 - Second Wave
- 3 Stochastic Volatility
- 4 Stochastic Volatility and GARCH
 - A Simple Tractable Model
 - An Application
- 5 Summary



Ben Kedem

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable
Model

An Application

Summary

- Ben has made many contributions to time series methodology.
- A common theme is that some unobserved (*latent*) series controls either:
 - the *values* of the observed data, or
 - the *distribution* the observed data.
- In a *stochastic volatility* model, a latent series controls specifically the *variance* of the observed data.
- We relate stochastic volatility models to other time series models.



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

1 Introduction

2 Time Series Models

■ First Wave

■ Second Wave

3 Stochastic Volatility

4 Stochastic Volatility and GARCH

■ A Simple Tractable Model

■ An Application

5 Summary



Correlation

Modeling The
Variance of a
Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

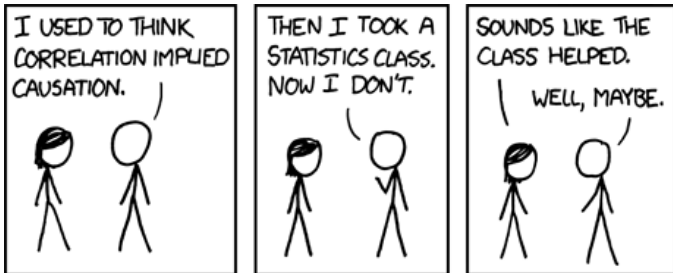
Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

Time series modeling is not just about correlation...



<http://xkcd.com/552/>



Time Domain Approach

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- The time domain approach to modeling a time series $\{Y_t\}$ focuses on the conditional distribution of $Y_t | Y_{t-1}, Y_{t-2}, \dots$.
- One reason for this focus is that the joint distribution of Y_1, Y_2, \dots, Y_n can be factorized as

$$\begin{aligned} f_{1:n}(y_1, y_2, \dots, y_n) \\ = f_1(y_1) f_{2|1}(y_2 | y_1) \cdots f_{n|n-1:1}(y_n | y_{n-1}, y_{n-2}, \dots, y_1). \end{aligned}$$

- So the likelihood function is determined by these conditional distributions.



Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

The conditional distribution may be defined by:

- the conditional mean,

$$\mu_t = E(Y_t | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \dots);$$

- the conditional variance,

$$h_t = \text{Var}(Y_t | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \dots);$$

- the *shape* of the conditional distribution.



Forecasting

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- The conditional distribution of $Y_t | Y_{t-1}, Y_{t-2}, \dots$ also gives the most complete solution to the forecasting problem:
 - We observe Y_{t-1}, Y_{t-2}, \dots ;
 - what statements can we make about Y_t ?
- The conditional mean is our best forecast, and the conditional standard deviation measures how far we believe the actual value might differ from the forecast.
- The conditional shape, usually a fixed distribution such as the normal, allows us to make probability statements about the actual value.



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- 1 Introduction
- 2 Time Series Models
 - First Wave
 - Second Wave
- 3 Stochastic Volatility
- 4 Stochastic Volatility and GARCH
 - A Simple Tractable Model
 - An Application
- 5 Summary



First Wave

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

- The first wave of time series methods focused on the conditional mean, μ_t .
 - The conditional variance was assumed to be constant.
 - The conditional shape was either normal or unspecified.
- Need only to specify the form of

$$\mu_t = \mu_t(y_{t-1}, y_{t-2}, \dots).$$

- Time-homogeneous:

$$\mu_t = \mu(y_{t-1}, y_{t-2}, \dots),$$

depends on t only through y_{t-1}, y_{t-2}, \dots



Autoregression

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Simplest form:

μ_t = a linear function of a small number of values:

$$\mu_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}.$$

- Equivalently, and more familiarly,

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t,$$

where $\epsilon_t = Y_t - \mu_t$ satisfies

$$E(\epsilon_t | Y_{t-1}, Y_{t-2}, \dots) = 0,$$

$$\text{Var}(\epsilon_t | Y_{t-1}, Y_{t-2}, \dots) = h.$$



Recursion

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Problem: some time series need large p .
- Solution: recursion; include also some past values of μ_t :

$$\mu_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \psi_1 \mu_{t-1} + \cdots + \psi_q \mu_{t-q}.$$

- Equivalently, and more familiarly,

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}.$$

- This is the ARMA (AutoRegressive Moving Average) model of order (p, q) .



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

1 Introduction

2 Time Series Models

■ First Wave

■ **Second Wave**

3 Stochastic Volatility

4 Stochastic Volatility and GARCH

■ A Simple Tractable Model

■ An Application

5 Summary



Two Years of S&P 500: Changing Variance

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

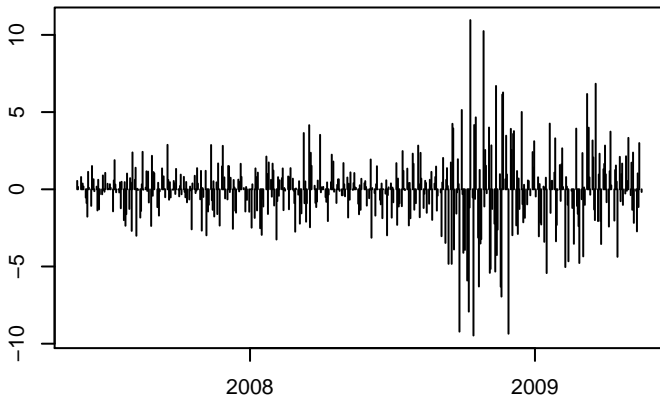
Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary





Second Wave

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

- The second wave of time series methods added a focus on the conditional variance, h_t .
- Now need to specify the form of

$$h_t = h_t(y_{t-1}, y_{t-2}, \dots).$$

- Time-homogeneous:

$$h_t = h(y_{t-1}, y_{t-2}, \dots),$$

depends on t only through y_{t-1}, y_{t-2}, \dots



ARCH

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Simplest form: h_t a linear function of a small number of squared ϵ s:

$$h_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2.$$

- Engle, ARCH (AutoRegressive Conditional Heteroscedasticity):
 - proposed in 1982;
 - Nobel Prize in Economics, 2003 (shared with the late Sir Clive Granger).



Recursion

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Problem: some time series need large q .
- Solution: recursion; include also some past values of h_t :

$$h_t = \omega + \alpha_1 \epsilon_{t-1}^2 + \cdots + \alpha_q \epsilon_{t-q}^2 + \beta_1 h_{t-1} + \cdots + \beta_p h_{t-p}.$$

- Bollerslev, 1987; GARCH (Generalized ARCH; no Nobel yet, nor yet a Knighthood).
- Warning! note the reversal of the roles of p and q from the convention of ARMA(p, q).



GARCH(1,1)

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- The simplest GARCH model has $p = 1, q = 1$:

$$h_t = \omega + \alpha \epsilon_{t-1}^2 + \beta h_{t-1}$$

- If $\alpha + \beta < 1$, there exists a stationary process with this structure.
- If $\alpha + \beta = 1$, the model is called *integrated*: IGARCH(1, 1).



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

1 Introduction

2 Time Series Models

■ First Wave

■ Second Wave

3 Stochastic Volatility

4 Stochastic Volatility and GARCH

■ A Simple Tractable Model

■ An Application

5 Summary



Stochastic Volatility

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- In a stochastic volatility model, an unobserved (*latent*) process $\{X_t\}$ affects the distribution of the observed process $\{Y_t\}$, specifically the variance of Y_t .
- Introducing a “second source of variability” is appealing from a modeling perspective.



Simple Example

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- For instance:
 - $\{X_t\}$ satisfies

$$X_t - \mu = \phi (X_{t-1} - \mu) + \xi_t,$$

where $\{\xi_t\}$ are i.i.d. $N(0, \sigma_\xi^2)$.

- If $|\phi| < 1$, this is a (stationary) autoregression, but if $\phi = 1$ it is a (non-stationary) random walk.
- $Y_t = \sigma_t \eta_t$, where $\sigma_t^2 = \sigma^2(X_t)$ is a non-negative function such as

$$\sigma^2(X_t) = \exp(X_t)$$

and $\{\eta_t\}$ are i.i.d. $(0, 1)$ -typically Gaussian, but also t .



Conditional Distributions

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- So the conditional distribution of Y_t given Y_{t-1}, Y_{t-2}, \dots and X_t, X_{t-1}, \dots is simple:

$$Y_t | Y_{t-1}, Y_{t-2}, \dots, X_t, X_{t-1}, \dots \sim N(0, \sigma^2(X_t)).$$

- But the conditional distribution of Y_t given only Y_{t-1}, Y_{t-2}, \dots is not analytically tractable.
- In particular,

$$h_t(y_{t-1}, y_{t-2}, \dots) = \text{Var}(Y_t | Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \dots)$$

is not a simple function.



Difficulties

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable
Model

An Application

Summary

- Analytic difficulties cause problems in:
 - estimation;
 - forecasting.
- Computationally intensive methods, e.g.:
 - Particle filtering;
 - Numerical quadrature.



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

1 Introduction

2 Time Series Models

■ First Wave

■ Second Wave

3 Stochastic Volatility

4 Stochastic Volatility and GARCH

■ A Simple Tractable Model

■ An Application

5 Summary



Stochastic Volatility and GARCH

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable
Model

An Application

Summary

- Stochastic volatility models have the attraction of an explicit model for the volatility, or variance.
- Is analytic difficulty the unavoidable cost of that advantage?



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

1 Introduction

2 Time Series Models

- First Wave
- Second Wave

3 Stochastic Volatility

4 Stochastic Volatility and GARCH

- A Simple Tractable Model
- An Application

5 Summary



The Latent Process

Modeling The
Variance of a
Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

We construct a latent process by:

$$X_0 \sim \Gamma\left(\frac{\nu}{2}, \frac{\tau^2}{2}\right),$$

and for $t > 0$

$$X_t = B_t X_{t-1},$$

where

$$\theta B_t \sim \beta\left(\frac{\nu}{2}, \frac{1}{2}\right)$$

and $\{B_t\}$ are i.i.d. and independent of X_0 .



The Observed Process

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- The observed process is defined for $t \geq 0$ by

$$Y_t = \sigma_t \eta_t$$

where

$$\sigma_t = \frac{1}{\sqrt{X_t}},$$

and $\{\eta_t\}$ are i.i.d. $N(0, 1)$ and independent of $\{X_t\}$.

- Equivalently: given $X_u = x_u, 0 \leq u$, and $Y_u = y_u, 0 \leq u < t$,

$$Y_t \sim N(0, \sigma_t^2)$$

with the same definition of σ_t .



Constraints

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Since

$$\text{Var}(Y_0) = E(X_0^{-1}),$$

we constrain $\nu > 2$ to ensure that

$$E(X_0^{-1}) < \infty.$$

- Requiring

$$E(X_t^{-1}) = E(X_0^{-1})$$

for all $t > 0$ is also convenient, and is met if

$$\theta = \frac{\nu - 2}{\nu - 1}.$$



Comparison with Earlier Example

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

This is quite similar to the earlier example, with $\phi = 1$:

■ Write $X_t^* = -\log(X_t)$.

■ Then

$$X_t^* = X_{t-1}^* + \xi_t^*,$$

where

$$\xi_t^* = -\log(B_t).$$

■ In terms of X_t^* ,

$$\sigma_t^2 = \exp(X_t^*).$$



Differences

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series
Models

First Wave

Second Wave

Stochastic
Volatility

Stochastic
Volatility and
GARCH

A Simple Tractable
Model

An Application

Summary

- A key constraint is that now $\phi = 1$, so $\{X_t^*\}$ is a (non-stationary) random walk, instead of a (stationary) auto-regression.
- Also $\{X_t^*\}$ is non-Gaussian, where in the earlier example, the latent process was Gaussian.
- Also $\{X_t^*\}$ has a drift, because

$$E(\xi_t^*) \neq 0.$$

- Of course, we could include a drift in the earlier example.



Matched Simulated Random Walks

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

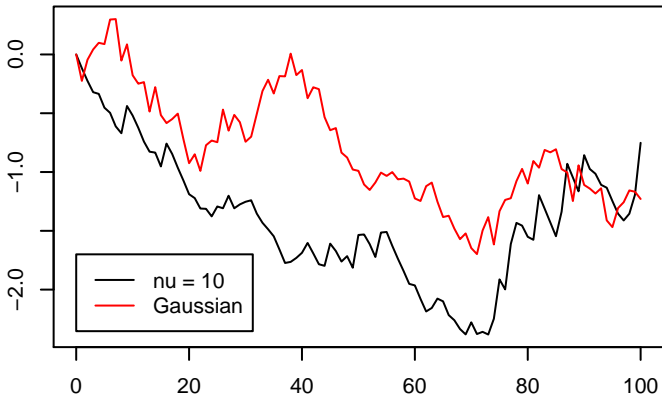
Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary





Matched Simulated Random Walks

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

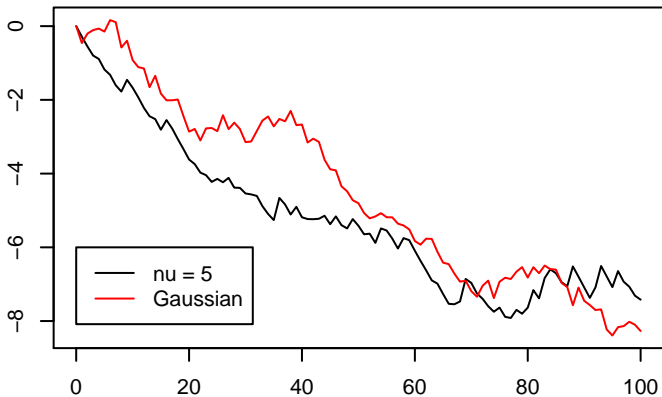
Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary





So What?

Modeling The Variance of a Time Series

Peter
Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable
Model

An Application

Summary

- So our model is not very different from (a carefully chosen instance of) the earlier example.
- So does it have any advantage?
- Note: the inverse Gamma distribution is the *conjugate prior* for the variance of the Gaussian distribution.



Marginal distribution of Y_0

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Marginal distribution of Y_0 :

$$Y_0 \sim \sqrt{h_0} t^*(\nu)$$

where

$$h_0 = \frac{\tau^2}{\nu - 2}$$

and $t^*(\nu)$ is the *standardized* t -distribution (i.e., scaled to have variance 1).



Conditional distributions of X_0 and $X_1|Y_0$

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Conjugate prior/posterior property: conditionally on $Y_0 = y_0$,

$$X_0 \sim \Gamma\left(\frac{\nu + 1}{2}, \frac{\tau^2 + y_0^2}{2}\right).$$

- Beta multiplier property: conditionally on $Y_0 = y_0$,

$$X_1 = B_1 X_0 \sim \Gamma\left[\frac{\nu}{2}, \theta \left(\frac{\tau^2 + y_0^2}{2}\right)\right].$$



Conditional distribution of $Y_1|Y_0$

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- The conditional distribution of $X_1|Y_0$ differs from the distribution of X_0 only in scale, so conditionally on $Y_0 = y_0$,

$$Y_1 \sim \sqrt{h_1} t^*(\nu),$$

where

$$h_1 = \frac{\theta}{\nu - 2} (\tau^2 + y_0^2) = \theta h_0 + (1 - \theta)y_0^2.$$

- Hmm...so the distribution of $Y_1|Y_0$ differs from the distribution of Y_0 only in scale...I smell a recursion!



The Recursion

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Write $\mathbf{Y}_{t-1} = (Y_{t-1}, Y_{t-2}, \dots, Y_0)$.
- For $t > 0$, conditionally on $\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}$,

$$Y_t \sim \sqrt{h_t} t^*(\nu),$$

where

$$h_t = \theta h_{t-1} + (1 - \theta) y_{t-1}^2.$$



The Structure

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- That is, $\{Y_t\}$ is IGARCH(1, 1) with $t(\nu)$ -distributed innovations.
- Constraints:
 - $\omega = 0$;
 - $\beta = 1 - \alpha = \frac{\nu-2}{\nu-1}$.
- So we can have a stochastic volatility structure, and still have (I)GARCH structure for the observed process $\{Y_t\}$.



The Structure

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- That is, $\{Y_t\}$ is IGARCH(1, 1) with $t(\nu)$ -distributed innovations.
- Constraints:
 - $\omega = 0$;
 - $\beta = 1 - \alpha = \frac{\nu-2}{\nu-1}$.
- So we can have a stochastic volatility structure, and still have (I)GARCH structure for the observed process $\{Y_t\}$.
- Details and some multivariate generalizations sketched out at <http://www4.stat.ncsu.edu/~bloomfld/talks/sv.pdf>



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

1 Introduction

2 Time Series Models

- First Wave
- Second Wave

3 Stochastic Volatility

4 Stochastic Volatility and GARCH

- A Simple Tractable Model
- An Application

5 Summary



Same Two Years of S&P 500

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

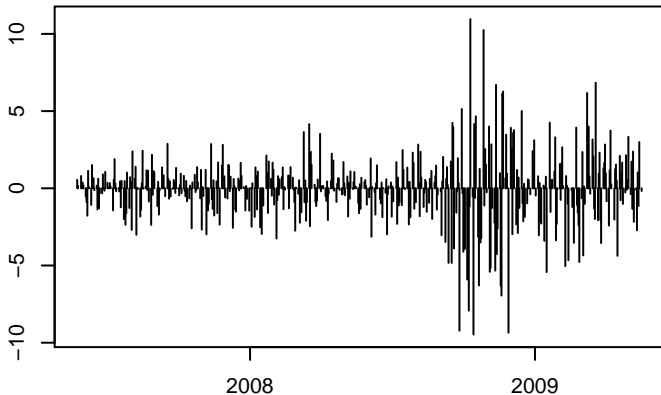
Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary





Two Years of S&P 500

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Data: 500 log-returns for the S&P 500 index, from 05/24/2007 to 05/19/2009.
- Maximum likelihood estimates:

$$\hat{\tau}^2 = 4.37$$

$$\hat{\theta} = 0.914$$

$$\Rightarrow \hat{\nu} = 12.6.$$

- With ν unconstrained:

$$\hat{\tau}^2 = 3.37$$

$$\hat{\theta} = 0.918$$

$$\hat{\nu} = 9.93.$$



Comparison

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Constrained result has less heavy tails and less memory than unconstrained result.
- Likelihood ratio test:

$$-2 \log(\text{likelihood ratio}) = 0.412$$

assuming $\sim \chi^2(1)$, $P = 0.521$,

so differences are not significant.

- With more data, difference becomes significant.



Outline

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- 1 Introduction
- 2 Time Series Models
 - First Wave
 - Second Wave
- 3 Stochastic Volatility
- 4 Stochastic Volatility and GARCH
 - A Simple Tractable Model
 - An Application
- 5 Summary



Summary

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Latent processes are useful in time series modeling.
- GARCH and Stochastic Volatility are both valuable tools for modeling time series with changing variance.
- GARCH fits naturally into the time domain approach.
- Stochastic Volatility is appealing but typically intractable.
- Exploiting conjugate distributions may bridge the gap.



Summary

Modeling The Variance of a Time Series

Peter Bloomfield

Introduction

Time Series Models

First Wave

Second Wave

Stochastic Volatility

Stochastic Volatility and GARCH

A Simple Tractable Model

An Application

Summary

- Latent processes are useful in time series modeling.
- GARCH and Stochastic Volatility are both valuable tools for modeling time series with changing variance.
- GARCH fits naturally into the time domain approach.
- Stochastic Volatility is appealing but typically intractable.
- Exploiting conjugate distributions may bridge the gap.

Thank you!

ON OPTIMAL PREDICTIVE INFERENCE
IN LOG-GAUSSIAN RANDOM FIELDS

Victor De Oliveira
Department of Management Sciences
and Statistics
University of Texas at San Antonio
`victor.deoliveira@utsa.edu`

`http://faculty.business.utsa.edu/vdeolive`

LOG-GAUSSIAN RANDOM FIELDS

The random field $\{Z(\mathbf{s}), \mathbf{s} \in D\}$, $D \subset \mathbb{R}^d$ and $d \geq 1$, is log-Gaussian if $\{Y(\mathbf{s}), \mathbf{s} \in D\}$, with $Y(\mathbf{s}) = \log(Z(\mathbf{s}))$, is Gaussian.

Here we assume that the mean and covariance functions of $Y(\cdot)$ are given by

$$\begin{aligned} E\{Y(\mathbf{s})\} &= \mu_Y \\ \text{cov}\{Y(\mathbf{s}), Y(\mathbf{u})\} &= C_Y(\mathbf{s}, \mathbf{u}) \end{aligned}$$

where $\mu_Y \in \mathbb{R}$ unknown and $C_Y(\mathbf{s}, \mathbf{u})$ a known covariance function in \mathbb{R}^d , satisfying that for all $\mathbf{s} \in D$, $C_Y(\mathbf{s}, \mathbf{s}) = \sigma_Y^2$.

It follows the mean and covariance functions of $Z(\cdot)$ are given by

$$\begin{aligned} E\{Z(\mathbf{s})\} &= \exp\left\{\mu_Y + \frac{\sigma_Y^2}{2}\right\} =: \mu_Z \\ \text{cov}\{Z(\mathbf{s}), Z(\mathbf{u})\} &= \mu_Z^2 (\exp\{C_Y(\mathbf{s}, \mathbf{u})\} - 1). \end{aligned}$$

PREDICTIVE INFERENCE

DATA: $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))$ measured at sampling locations $s_1, \dots, s_n \in D$. Let

$s_0 \in D$ an unmeasured location in D

$B \subset D$ a subregion of interest

GOALS:

- Obtain predictor of $Z(s_0)$
(point prediction).
- Obtain prediction interval for $Z(s_0)$
(interval prediction)
- Obtain predictor of $Z(B) = \frac{1}{|B|} \int_B Z(s) ds$
(block prediction).

POINT PREDICTION

Let \hat{Z}_0 be a predictor for Z_0 (within a family) and $L(Z_0, \hat{Z}_0)$ a loss function.

The optimal predictor of Z_0 is the predictor that minimizes the risk function

$$r(\hat{Z}_0) = E\{L(Z_0, \hat{Z}_0)\}.$$

For the squared error loss the risk function becomes the mean squared prediction error

$$\text{MSPE}(\hat{Z}_0) = E\{(\hat{Z}_0 - Z_0)^2\}.$$

Notation: All quantities that depend on the prediction location s_0 would be written with the subscript '0'.

If μ_Y were known, the optimal predictor and its MSPE are given by

$$\begin{aligned}\hat{Z}_0^* &= E\{Z_0 | \mathbf{Z}\} \\ &= \exp\{\hat{Y}_0^* + \frac{\hat{\sigma}_{0Y}^*}{2}\} \\ \text{MSPE}(\hat{Z}_0^*) &= \text{var}(Z_0) - \text{var}(\hat{Z}_0^*) \\ &= \mu_Z^2 (\exp\{\sigma_Y^2\} - \exp\{\mathbf{c}'_{0Y} \Sigma_Y^{-1} \mathbf{c}_{0Y}\}),\end{aligned}$$

where

$$\begin{aligned}\hat{Y}_0^* &= E\{Y_0 | \mathbf{Y}\} \\ &= \mu_Y + \mathbf{c}'_{0Y} \Sigma_Y^{-1} (\mathbf{Y} - \mu_Y \mathbf{1}) \\ \hat{\sigma}_{0Y}^* &= \text{var}(Y_0 | \mathbf{Y}) \\ &= \sigma_Y^2 - \mathbf{c}'_{0Y} \Sigma_Y^{-1} \mathbf{c}_{0Y};\end{aligned}$$

$$\begin{aligned}\Sigma_{Y,ij} &= C_Y(\mathbf{s}_i, \mathbf{s}_j) \\ \mathbf{c}_{0Y,i} &= C_Y(\mathbf{s}_0, \mathbf{s}_i).\end{aligned}$$

These are known in the geostatistical literature as the simple kriging predictors and simple kriging variances of Z_0 and Y_0 .

Unbiased Prediction:

In practice the most used predictor is the lognormal kriging predictor

$$\hat{Z}_0^{LK} = \exp\left\{\hat{Y}_0^{OK} + \frac{1}{2}(\sigma_Y^2 - \lambda'_{0Y}\Sigma_Y\lambda_{0Y})\right\},$$

$$\text{MSPE}(\hat{Z}_0^{LK}) = \mu_Z^2(\exp\{\sigma_Y^2\} + \exp\{\lambda'_{0Y}\Sigma_Y\lambda_{0Y}\} - 2\exp\{\lambda'_{0Y}\Sigma_Y\lambda_{0Y} - m_{0Y}\});$$

where

$$\begin{aligned}\hat{Y}_0^{OK} &= \lambda'_{0Y}\mathbf{Y} \quad (\text{the BLUP of } Y_0) \\ \lambda'_{0Y} &= (\mathbf{c}_{0Y} + \frac{1 - \mathbf{1}'\Sigma_Y^{-1}\mathbf{c}_{0Y}}{\mathbf{1}'\Sigma_Y^{-1}\mathbf{1}}\mathbf{1})'\Sigma_Y^{-1}.\end{aligned}$$

By construction \hat{Z}_0^{LK} satisfies the following optimality property:

Proposition 1. The predictor \hat{Z}_0^{LK} minimizes $E\{(\log(\hat{Z}_0) - \log(Z_0))^2\}$ over the class of predictors of the form $\hat{Z}_0 = \exp\{\lambda_0' \log(\mathbf{Z}) + k_0\}$, where $\lambda_0 \in \mathbb{R}^n$ and $k_0 \in \mathbb{R}$ are constrained such that $E\{\hat{Z}_0\} = E\{Z_0\}$ for every $\mu_Y \in \mathbb{R}$.

Recently, Cox (2004) noted a stronger optimality property:

Proposition 2. The predictor \hat{Z}_0^{LK} minimizes $E\left\{\frac{(\hat{Z}_0 - Z_0)^2}{\exp\{c'_{0Y} \Sigma_Y^{-1} \log(\mathbf{Z})\}}\right\}$ over the class of all unbiased predictors of Z_0 .

These optimality properties are somewhat unsatisfactory.

The former holds in the transformed log-scale rather than in the original scale of measurement.

The latter, although holds in the original scale, is with respect to a weighted squared error loss function with little intuitive appeal.

Optimal Point Prediction:

Consider the family of predictors

$$\mathcal{P}_0 = \{\hat{Z}_0 = \exp\{\mathbf{a}'_0 \mathbf{Y} + k_0\} : k_0 \in \mathbb{R}, \mathbf{a}_0 \in \mathbb{R}^n, \mathbf{a}'_0 \mathbf{1} = 1\}$$

which includes many special cases:

$$\hat{Z}_0^{LK}, \hat{Z}_0^N = \exp\{\hat{Y}_0^{OK}\} \text{ and}$$

$$\hat{Z}_0^{ML} = \exp\{\hat{Y}_0^{OK} + \frac{1}{2}(\sigma_Y^2 - \mathbf{c}'_{0Y} \Sigma_Y^{-1} \mathbf{c}_{0Y})\},$$

$$\hat{Z}_0^B = \exp\{\hat{Y}_0^{OK} + \frac{1}{2}(\sigma_Y^2 + \boldsymbol{\lambda}'_{0Y} \Sigma_Y \boldsymbol{\lambda}_{0Y} - 2\boldsymbol{\lambda}'_{0Y} \mathbf{c}_{0Y})\},$$

Theorem 1. The predictor in the family \mathcal{P}_0 that minimizes $E\{(\hat{Z}_0 - Z_0)^2\}$ is given by

$$\hat{Z}_0^{ME} = \exp\{\hat{Y}_0^{OK} + \frac{1}{2}(\sigma_Y^2 - \boldsymbol{\lambda}'_{0Y} \Sigma_Y \boldsymbol{\lambda}_{0Y} - 2m_{0Y})\},$$

and its MSPE is given by

$$\text{MSPE}(\hat{Z}_0^{ME}) = \mu_Z^2 (\exp\{\sigma_Y^2\} - \exp\{\boldsymbol{\lambda}'_{0Y} \Sigma_Y \boldsymbol{\lambda}_{0Y} - 2m_{0Y}\}).$$

$$\text{where } m_{0Y} = \frac{1 - \mathbf{1}' \Sigma_Y^{-1} \mathbf{c}_{0Y}}{\mathbf{1}' \Sigma_Y^{-1} \mathbf{1}}.$$

BLOCK PREDICTION

A related problem is the prediction of:

$$Z(B) = \frac{1}{|B|} \int_B Z(s) ds, \quad B \subset D,$$

based on (point) data \mathbf{Z} .

Examples where inference about this arises:

- Environmental assessment
- Precision farming

Two predictors have been proposed in the geostatistical literature:

The lognormal kriging block predictor

$$\hat{Z}(B)^{LK} = \frac{1}{|B|} \int_B \hat{Z}^{LK}(s) ds.$$

A block predictor motivated by the assumption of “preservation of lognormality”

$$\hat{Z}(B)^{PL} = \exp \left\{ \hat{Y}(B)^{OK} + \frac{1}{2}(\sigma_Y^2 - \lambda'_Y(B)\Sigma_Y\lambda_Y(B)) \right\},$$

where $\hat{Y}(B)^{OK} = \lambda'_Y(B)\mathbf{Y}$ is the BLUP of $Y(B) = \int_B Y(s) ds / |B|$ based on \mathbf{Y} .

Optimal Block Prediction:

Consider the family of block predictors

$$\mathcal{P}_B = \left\{ \hat{Z}(B) = \frac{1}{|B|} \int_B \exp\{\hat{Y}^{OK}(s) + k(s)\} ds : k(s) \in \mathcal{C}(B) \right\},$$

where $\mathcal{C}(B)$ is the space of bounded and Lebesgue measurable functions on B .

Theorem 2. The predictor in the family of predictors \mathcal{P}_B that minimizes $E\{(\hat{Z}(B) - Z(B))^2\}$ is given by

$$\hat{Z}(B)^{ME} = \frac{1}{|B|} \int_B \hat{Z}^{ME}(s) ds,$$

where $\hat{Z}^{ME}(s)$ is the optimal point predictor given before, and $\text{MSPE}(\hat{Z}(B)^{ME})$ is given by

$$\frac{\mu_Z^2}{|B|^2} \int_B \int_B \left(\exp\{C_Y(s, \mathbf{u})\} - \exp\{\lambda'_Y(s) \Sigma_Y \lambda_Y(\mathbf{u}) - m_Y(s) - m_Y(\mathbf{u})\} \right) ds du.$$

Considering now the family of block predictors

$$\tilde{\mathcal{P}}_B = \{\hat{Z}(B) = \exp\{\hat{Y}(B)^{OK} + k_B\} : k_B \in \mathbb{R}\}.$$

Theorem 3. The predictor in the family of predictors $\tilde{\mathcal{P}}_B$ that minimizes $E\{(\hat{Z}(B) - Z(B))^2\}$ is given by

$$\begin{aligned} \hat{Z}(B)^{MP} = \exp\{ & \hat{Y}(B)^{OK} + \frac{1}{2}(\sigma_Y^2 - 3\boldsymbol{\lambda}'_Y(B)\boldsymbol{\Sigma}_Y\boldsymbol{\lambda}_Y(B)) \\ & + \log\left(\frac{1}{|B|} \int_B e^{\boldsymbol{\lambda}'_Y(B)\mathbf{c}_Y(s)} ds\right)\} \end{aligned}$$

and $\text{MSPE}(\hat{Z}(B)^{MP})$ is given by

$$\begin{aligned} & \frac{\mu_Z^2}{|B|^2} \left(\int_B \int_B \exp\{C_Y(\mathbf{s}, \mathbf{u})\} ds du \right. \\ & \left. - |B|^2 \exp\left\{2 \log\left(\frac{1}{|B|} \int_B e^{\boldsymbol{\lambda}'_Y(B)\mathbf{c}_Y(s)} ds\right) - \boldsymbol{\lambda}'_Y(B)\boldsymbol{\Sigma}_Y\boldsymbol{\lambda}_Y(B)\right\}\right). \end{aligned}$$

Remarks.

From the above results follow that:

- \hat{Z}_0^{LK} is *inadmissible*, in the sense that $\text{MSPE}(\hat{Z}_0^{ME}) \leq \text{MSPE}(\hat{Z}_0^{LK})$ for all $\mu_Y \in \mathbb{R}$.
- $\hat{Z}(B)^{LK}$ and $\hat{Z}(B)^{PL}$ are both *inadmissible*.
- $\hat{Z}(B)^{ME}$ and $\hat{Z}(B)^{MP}$, and their MSPEs can not be compared analytically; they are compared numerically.

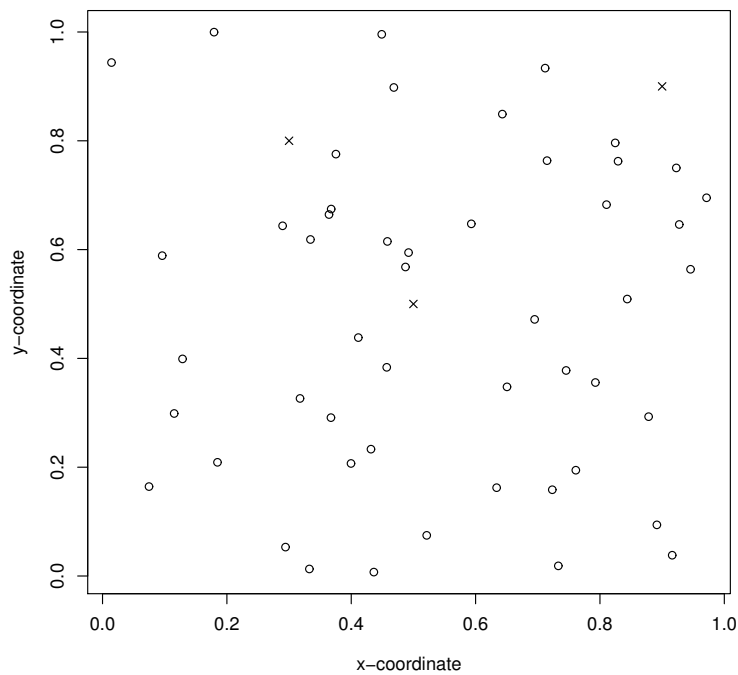
COMPARISON OF PREDICTORS

Let the region $D = [0, 1] \times [0, 1]$ and random field $Z(\mathbf{s}) = \exp\{Y(\mathbf{s})\}$, where $\{Y(\mathbf{s}), \mathbf{s} \in D\}$ is Gaussian with

$$E\{Y(\mathbf{s})\} = \mu_Y \quad , \quad C_Y(\mathbf{s}, \mathbf{u}) = \sigma_Y^2 \exp\left\{-\frac{l}{\theta_Y}\right\};$$

$l = \|\mathbf{s} - \mathbf{u}\|$ is Euclidean distance, $\mu_Y \in \mathbb{R}$ and $\sigma_Y^2, \theta_Y > 0$.

Data on $Z(\cdot)$ is observed at $n = 50$ sampling locations chosen at random.



Point Predictors:

We compare the values of \hat{Z}_0^{ME} and \hat{Z}_0^{LK} by predicting $Z(s_0)$ for locations:

$s_0 = (0.5, 0.5), (0.3, 0.8)$ and $(0.9, 0.9)$.

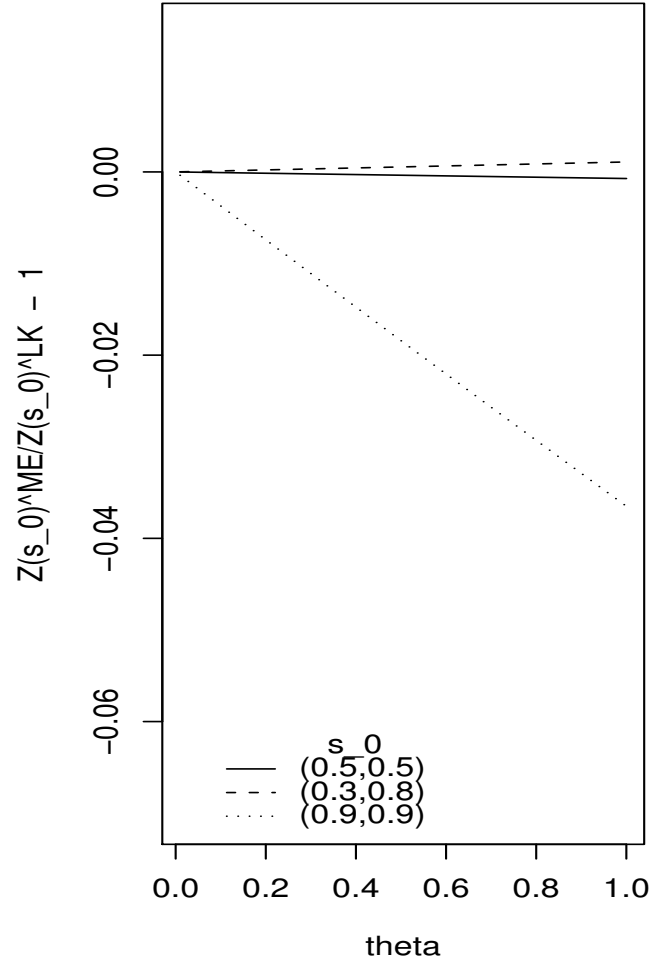
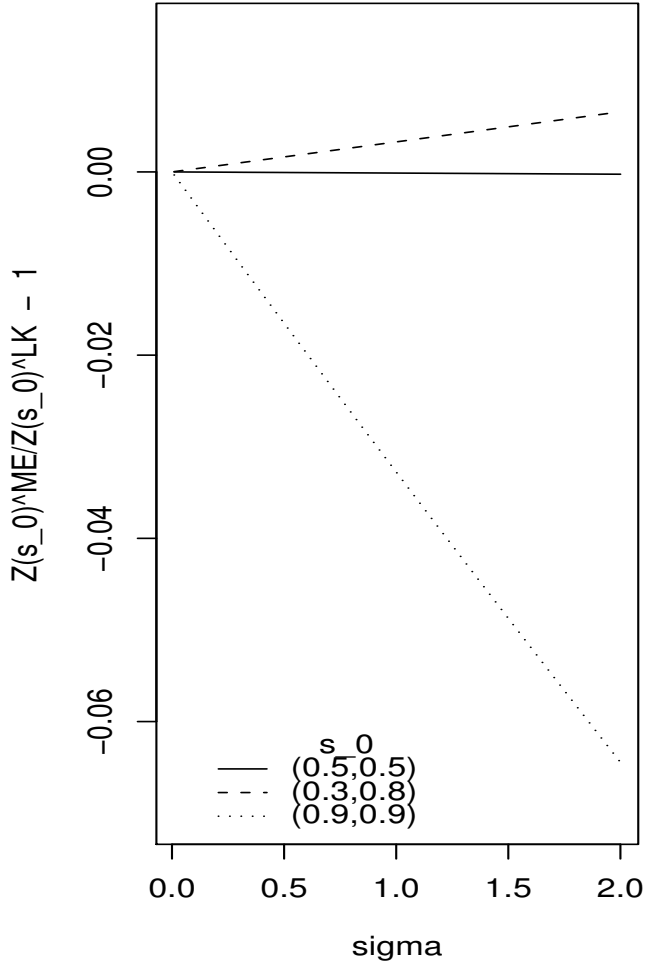
For that note

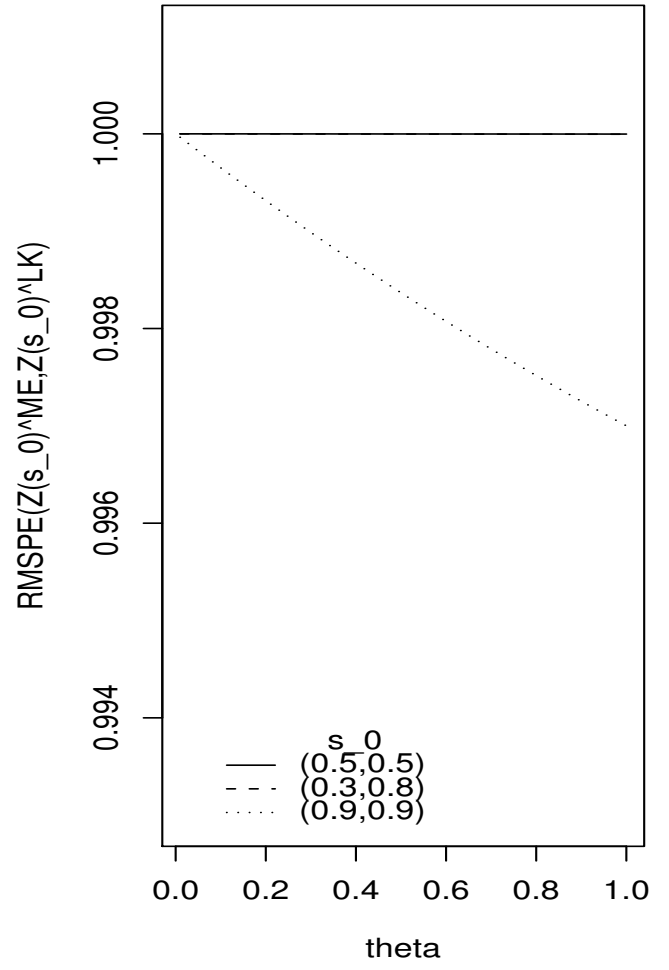
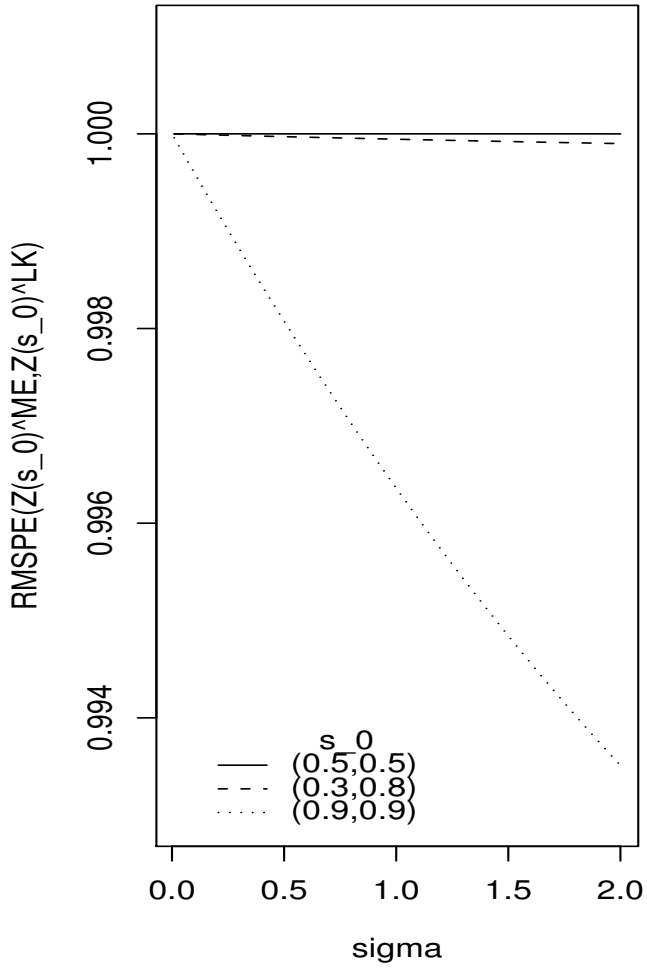
$$\frac{\hat{Z}_0^{ME}}{\hat{Z}_0^{LK}} = \exp\{-m_{0Y}\} = \frac{E\{\hat{Z}_0^{ME}\}}{E\{Z_0\}},$$

which does not depend on the observed data.

To compare the predictors in terms of their MSPE's we use the *predictive efficiency* of \hat{Z}_0^{ME} relative to \hat{Z}_0^{LK}

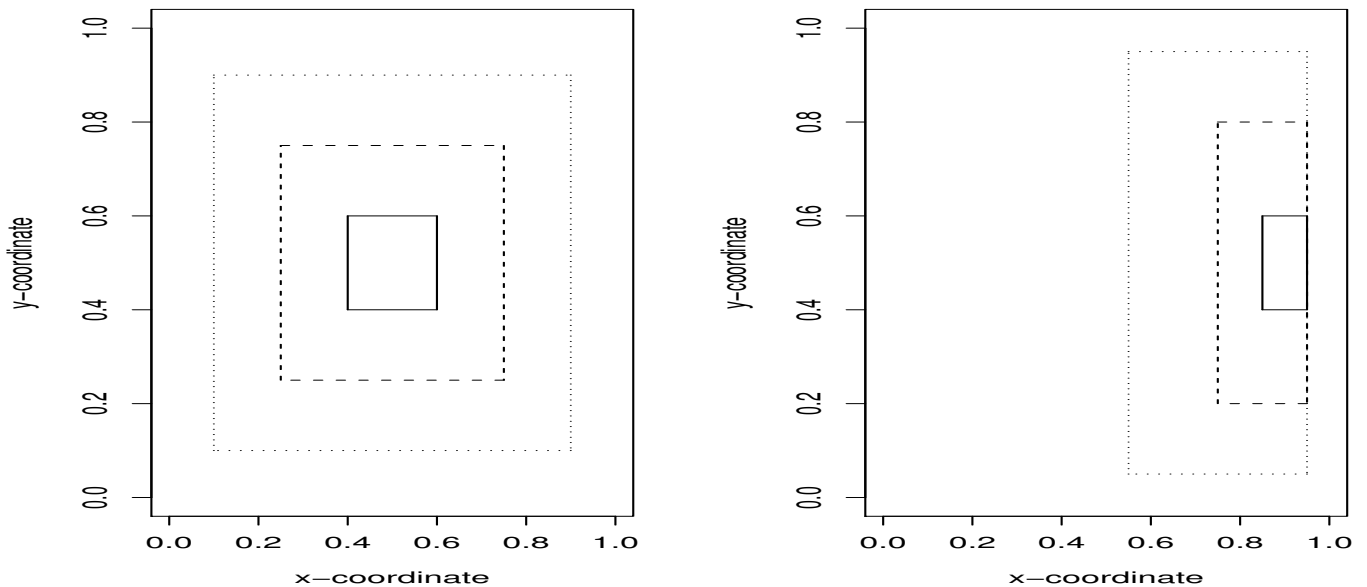
$$\text{RMSPE}(\hat{Z}_0^{ME}, \hat{Z}_0^{LK}) = \frac{\text{MSPE}(\hat{Z}_0^{ME})}{\text{MSPE}(\hat{Z}_0^{LK})}.$$





Block Predictors:

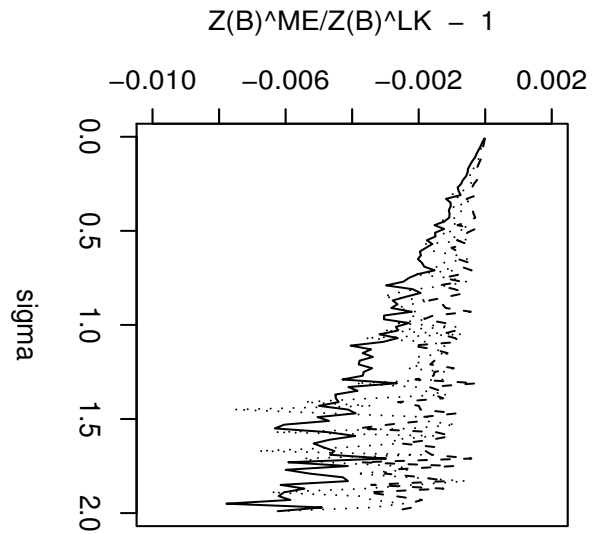
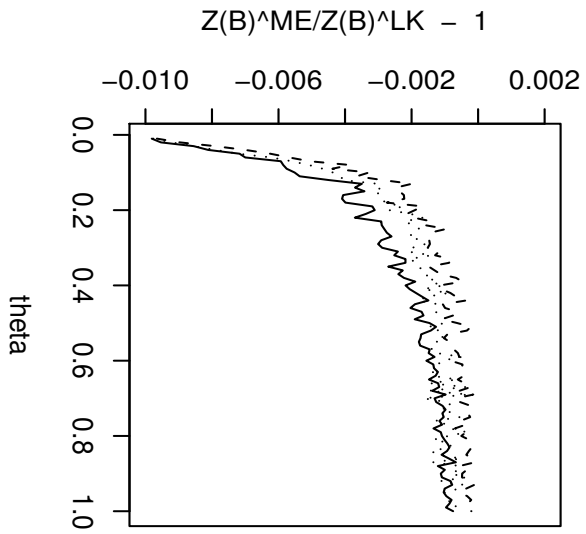
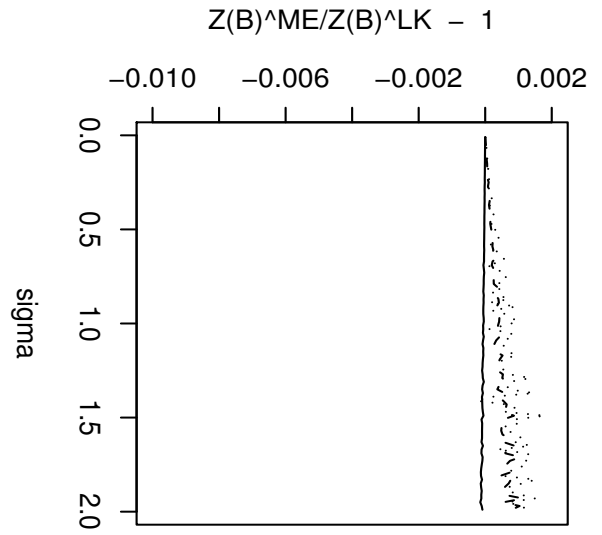
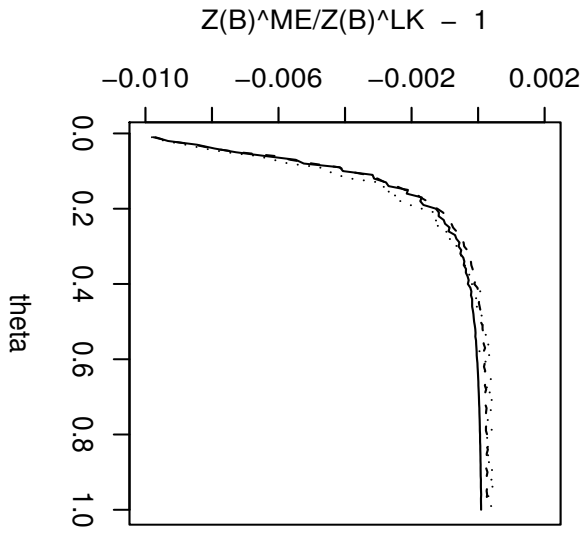
We compare the block predictors of $Z(B)$ for the sub-regions B shown below:

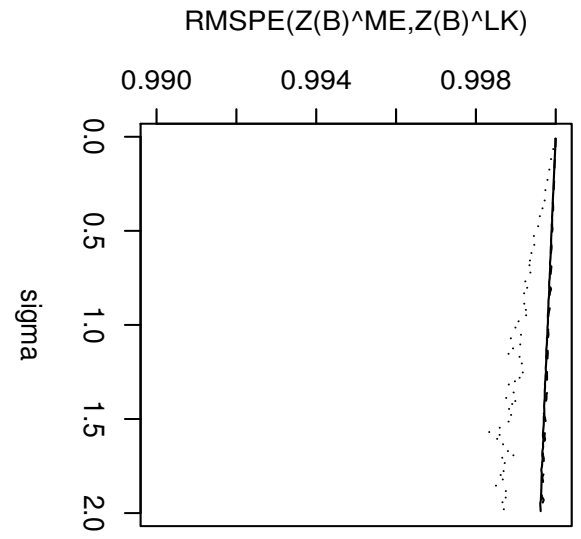
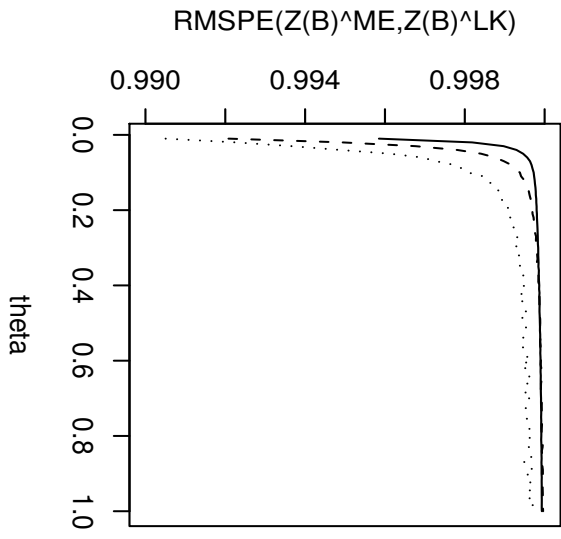
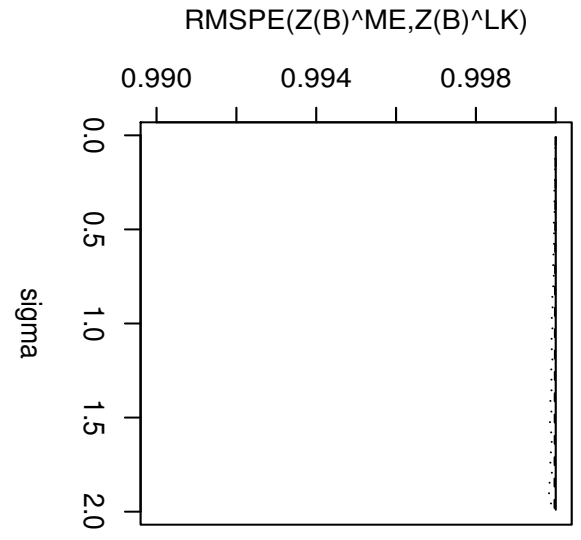
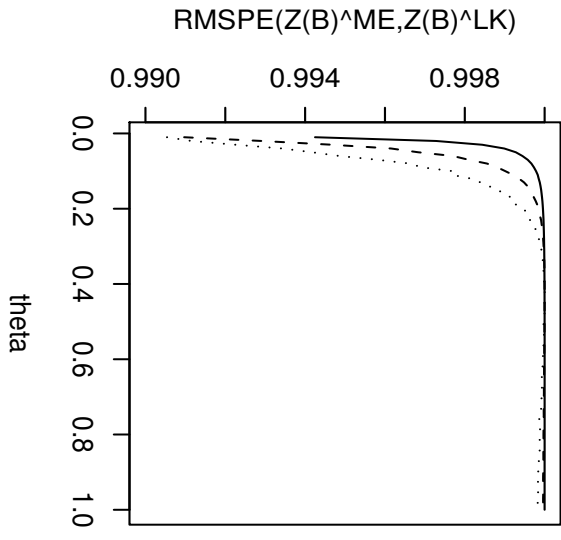


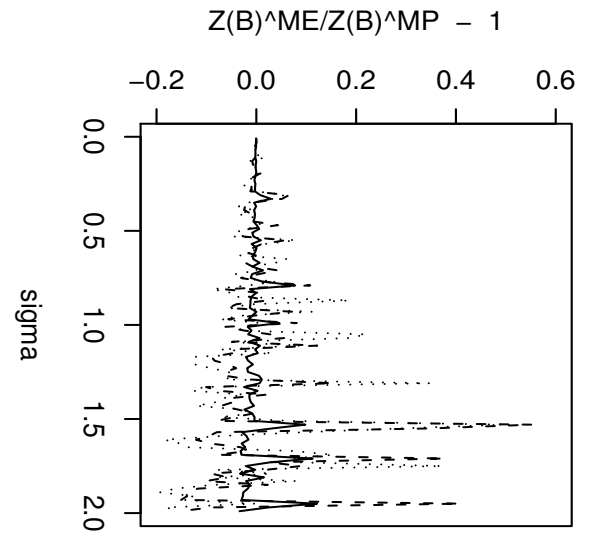
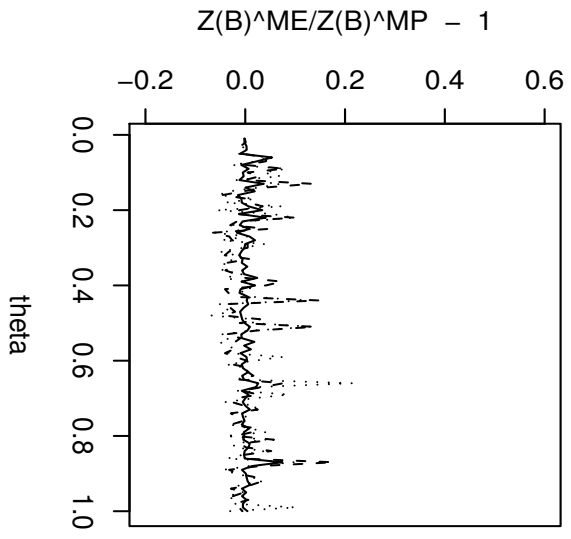
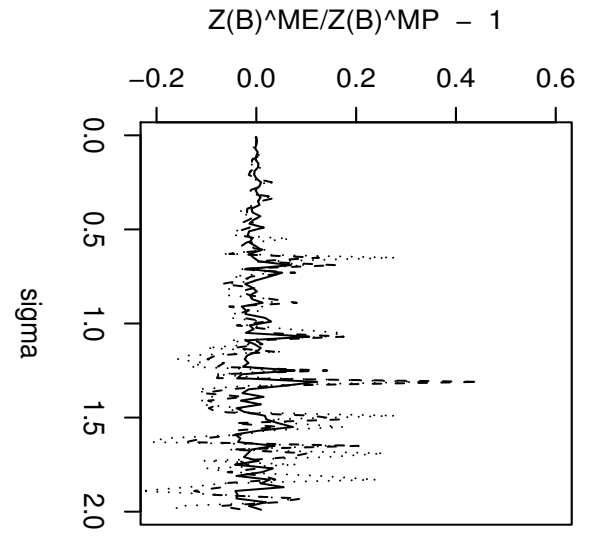
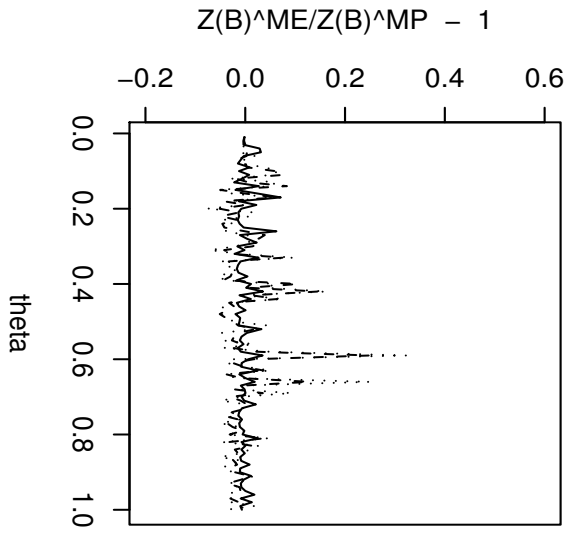
The predictors are approximated by noting that

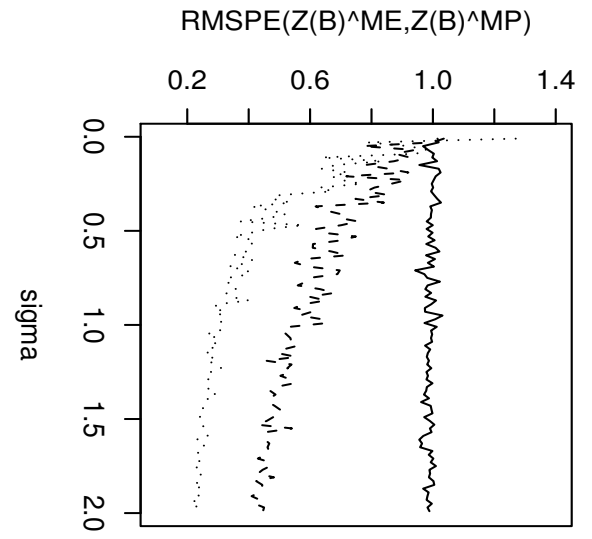
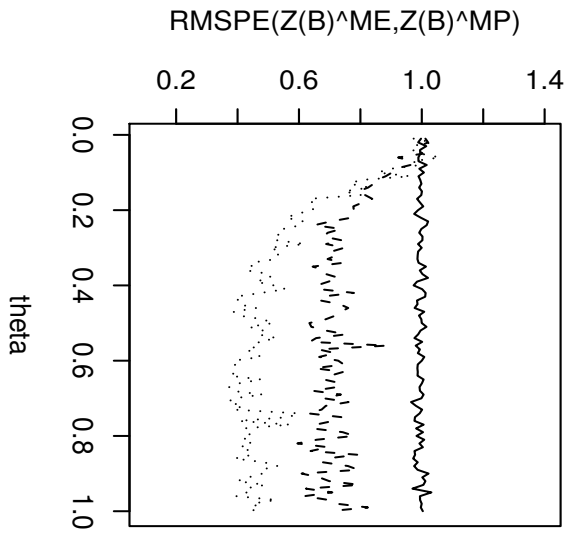
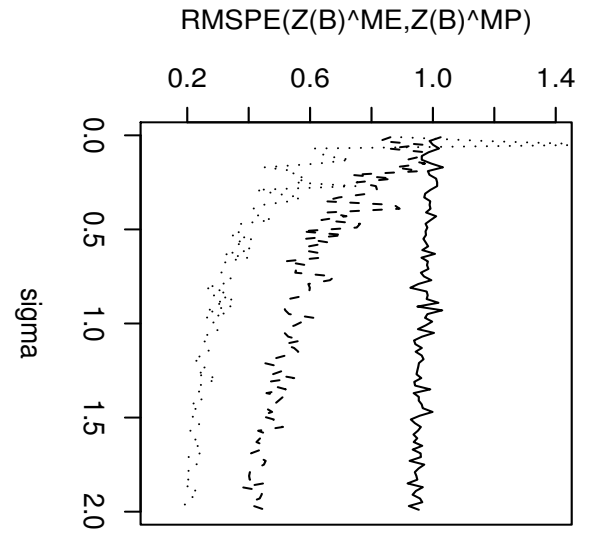
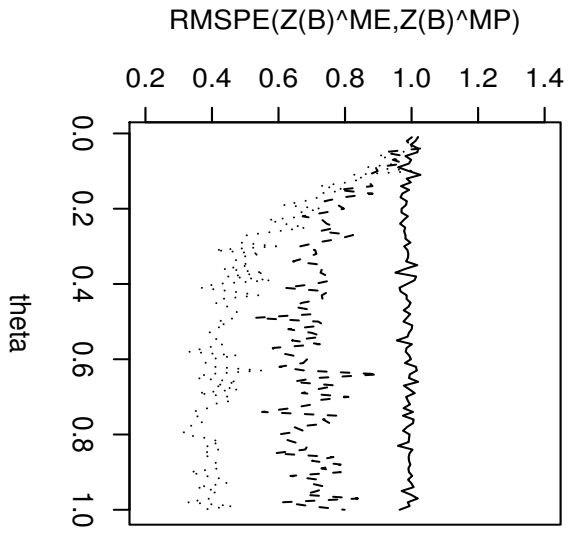
$$\hat{Z}(B)^{LK} = E_S\{\hat{Z}^{LK}(S)\} \quad , \quad \hat{Z}(B)^{ME} = E_S\{\hat{Z}^{ME}(S)\},$$

where $\hat{Z}^{LK}(\cdot)$ and $\hat{Z}^{ME}(\cdot)$ are point predictors and expectation is taken with respect to $S \sim \text{unif}(B)$.









The Non-constant Mean Case:

Suppose now that

$$\mu_Y(\mathbf{s}) = \sum_{j=1}^p \beta_j f_j(\mathbf{s}),$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ are *unknown* regression parameters, $(f_1(\mathbf{s}), \dots, f_p(\mathbf{s}))'$ are known location-dependent covariates.

In this case we have:

- The result on optimal point prediction (Theorem 1) can be easily extended.
- The results on optimal block prediction (Theorems 2 and 3) *cannot* be extended.

FIRST CONCLUSIONS

- New point and block predictors for log-Gaussian processes have been proposed that improve upon existing ones.
- The lognormal kriging point and block predictors have (near) optimality properties in the original scale.
- The lognormal kriging block predictor is substantially better than the block predictor motivated by “permanence of lognormality”. Also, the best predictor in \mathcal{P}_B is substantially better than the the best predictor in $\tilde{\mathcal{P}}_B$.
- For random fields with non-constant mean the optimal results also hold for point prediction, but not for block prediction.

INTERVAL PREDICTION

Here we assume that the mean and covariance functions of $Y(\cdot)$ are given by

$$\begin{aligned} E\{Y(\mathbf{s})\} &= \sum_{j=1}^p \beta_j f_j(\mathbf{s}) \\ \text{cov}\{Y(\mathbf{s}), Y(\mathbf{u})\} &= C(\mathbf{s}, \mathbf{u}) \end{aligned}$$

$f_1(\mathbf{s}), \dots, f_p(\mathbf{s})$ *known* covariates

$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ *unknown* parameters

$C(\mathbf{s}, \mathbf{u})$ parametric covariance function in \mathbb{R}^d
satisfying $C(\mathbf{s}, \mathbf{s}) = \sigma^2 > 0$

OBSERVED DATA:

Noisy measurements of the random field $Z(\cdot)$ at known sampling locations $\mathbf{s}_1, \dots, \mathbf{s}_n \in D$:

$$Z_{i,\text{obs}} = Z(\mathbf{s}_i)\epsilon_i, \quad i = 1, \dots, n,$$

$\{\log(\epsilon_i)\} \stackrel{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$ are measurement errors distributed independently of $Z(\cdot)$, and $\sigma_\epsilon^2 \geq 0$.

GOAL:

obtain prediction interval for $Z_0 = Z(\mathbf{s}_0)$, the unobserved value of the process at $\mathbf{s}_0 \in D$, based on $\mathbf{Z} = \{Z_{i,\text{obs}}\}_{i=1}^n$.

Model parameters are $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^q$ include $\sigma_\epsilon^2, \sigma^2$ and other parameters in $C(\mathbf{s}, \mathbf{u})$.

Common approach to construct prediction intervals for $Z(\cdot)$ is to transform prediction intervals for $Y(\cdot)$. Let

$$\mathbf{Y} = \log(\mathbf{Z}) \quad \text{and} \quad Y_0 = \log(Z_0)$$

The BLUP of Y_0 based on \mathbf{Y} and its mean squared prediction error are

$$\hat{Y}_0(\boldsymbol{\vartheta}) = \boldsymbol{\lambda}'_0(\boldsymbol{\vartheta})\mathbf{Y} \quad , \quad \hat{\sigma}_0^2(\boldsymbol{\vartheta}) = \sigma^2 - 2\boldsymbol{\lambda}'_0(\boldsymbol{\vartheta})\mathbf{c}_0(\boldsymbol{\vartheta}) + \boldsymbol{\lambda}'_0(\boldsymbol{\vartheta})\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}\boldsymbol{\lambda}_0(\boldsymbol{\vartheta})$$

with

$$\boldsymbol{\lambda}'_0(\boldsymbol{\vartheta}) = (\mathbf{c}_0(\boldsymbol{\vartheta}) + X(X'\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{-1}X)^{-1}(\mathbf{x}_0 - X'\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{-1}\mathbf{c}_0(\boldsymbol{\vartheta})))'\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}^{-1}$$

$$X = (f_j(\mathbf{s}_i))_{n \times p}, \quad \mathbf{x}_0 = (f_1(\mathbf{s}_0), \dots, f_p(\mathbf{s}_0))'$$

$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}$, $\mathbf{c}_0(\boldsymbol{\vartheta})$ are the $n \times n$ and $n \times 1$ matrices:

$$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta},ij} = C(\mathbf{s}_i, \mathbf{s}_j) + \sigma_{\epsilon}^2 \mathbf{1}\{i = j\}, \quad \mathbf{c}_0(\boldsymbol{\vartheta})_i = C(\mathbf{s}_0, \mathbf{s}_i).$$

$\boldsymbol{\Sigma}_{\boldsymbol{\vartheta}}$ is positive definite for any $\boldsymbol{\vartheta} \in \Theta$.

We start by assuming that $\boldsymbol{\vartheta}$ is known.

It follows

$$\begin{pmatrix} Y_0 \\ \hat{Y}_0(\boldsymbol{\vartheta}) \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mathbf{x}'_0 \boldsymbol{\beta} \\ \mathbf{x}'_0 \boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \boldsymbol{\lambda}'_0(\boldsymbol{\vartheta}) \mathbf{c}_0(\boldsymbol{\vartheta}) \\ \boldsymbol{\lambda}'_0(\boldsymbol{\vartheta}) \mathbf{c}_0(\boldsymbol{\vartheta}) & \boldsymbol{\lambda}'_0(\boldsymbol{\vartheta}) \boldsymbol{\Sigma}_{\boldsymbol{\vartheta}} \boldsymbol{\lambda}_0(\boldsymbol{\vartheta}) \end{pmatrix} \right)$$

so $T = Y_0 - \hat{Y}_0(\boldsymbol{\vartheta}) \sim N(0, \hat{\sigma}_0^2(\boldsymbol{\vartheta}))$ is a pivot for the prediction of Y_0 .

Then a $1 - \alpha$ prediction interval for Y_0 is

$$\hat{Y}_0(\boldsymbol{\vartheta}) \pm \Phi^{-1}(1 - \alpha/2) \hat{\sigma}_0(\boldsymbol{\vartheta})$$

and a $1 - \alpha$ prediction interval for Z_0 is

$$\exp\{\hat{Y}_0(\boldsymbol{\vartheta}) \pm \Phi^{-1}(1 - \alpha/2) \hat{\sigma}_0(\boldsymbol{\vartheta})\}$$

We denote this PI as $I_0^N(\alpha, \boldsymbol{\vartheta})$ and call it the *standard* $1 - \alpha$ prediction interval for Z_0 .

SHORTEST PREDICTION INTERVALS: KNOWN COVARIANCE CASE

Consider the family of $1 - \alpha$ prediction intervals for Z_0

$$\mathcal{F}_0 = \{ (\exp\{\hat{Y}_0(\boldsymbol{\vartheta}) - \Phi^{-1}(1 - \gamma)\hat{\sigma}_0(\boldsymbol{\vartheta})\}, \exp\{\hat{Y}_0(\boldsymbol{\vartheta}) - \Phi^{-1}(1 - \alpha + \gamma)\hat{\sigma}_0(\boldsymbol{\vartheta})\}) : \gamma \in [0, \alpha) \}$$

which includes the standard prediction interval (obtained for $\gamma = \alpha/2$).

THEOREM. Let $\alpha \in (0, 1)$, $\boldsymbol{\vartheta} \in \Theta$ and $s_0 \in D$. Then the shortest prediction interval in \mathcal{F}_0 is the one corresponding to the value

$$\gamma = \gamma_0^{\text{opt}} = \gamma_0^{\text{opt}}(\alpha, \boldsymbol{\vartheta}) \in (0, \alpha/2)$$

which is the (unique) solution to the equation

$$\Phi^{-1}(1 - \gamma) - \Phi^{-1}(1 - \alpha + \gamma) = 2\hat{\sigma}_0(\boldsymbol{\vartheta})$$

Hence the shortest $1 - \alpha$ PI for Z_0 in \mathcal{F}_0 is

$$I_0^S(\alpha, \boldsymbol{\vartheta}) = (\exp\{\hat{Y}_0(\boldsymbol{\vartheta}) - \Phi^{-1}(1 - \gamma_0^{\text{opt}})\hat{\sigma}_0(\boldsymbol{\vartheta})\}, \exp\{\hat{Y}_0(\boldsymbol{\vartheta}) - \Phi^{-1}(1 - \alpha + \gamma_0^{\text{opt}})\hat{\sigma}_0(\boldsymbol{\vartheta})\})$$

COMPARISON

Let $RL(I_0^N(\alpha, \boldsymbol{\vartheta}), I_0^S(\alpha, \boldsymbol{\vartheta}))$ be defined as

$$\frac{\text{len}(I_0^S(\alpha, \boldsymbol{\vartheta}))}{\text{len}(I_0^N(\alpha, \boldsymbol{\vartheta}))} = \frac{\exp\{\Phi^{-1}(1 - \alpha + \gamma_0^{\text{opt}})\hat{\sigma}_0(\boldsymbol{\vartheta})\} - \exp\{-\Phi^{-1}(1 - \gamma_0^{\text{opt}})\hat{\sigma}_0(\boldsymbol{\vartheta})\}}{\exp\{\Phi^{-1}(1 - \alpha/2)\hat{\sigma}_0(\boldsymbol{\vartheta})\} - \exp\{-\Phi^{-1}(1 - \alpha/2)\hat{\sigma}_0(\boldsymbol{\vartheta})\}}$$

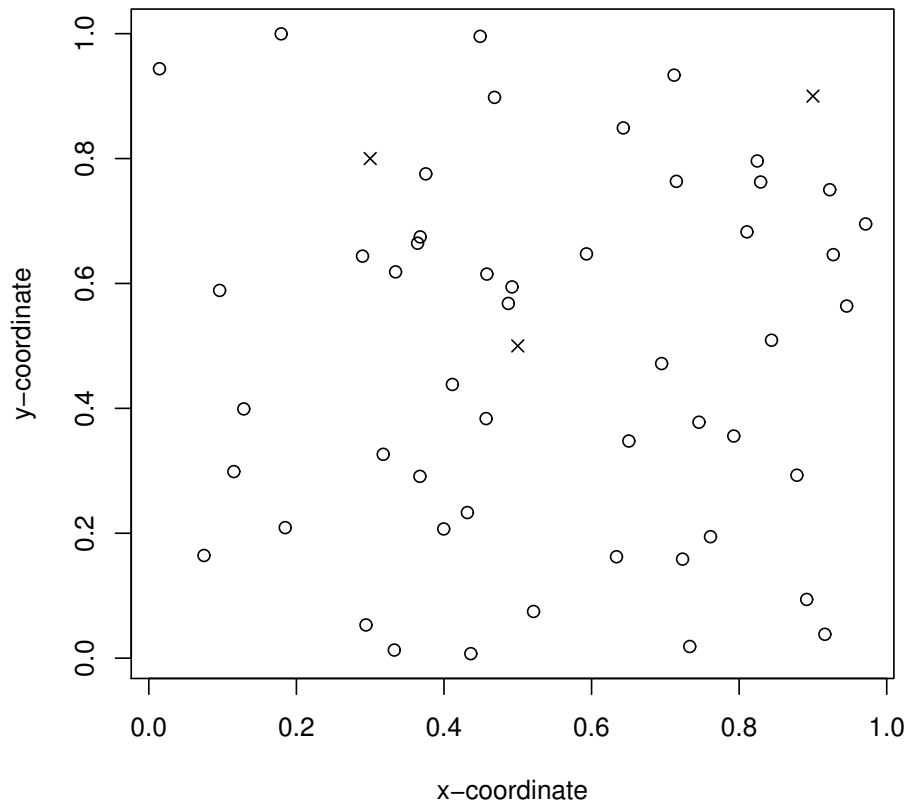
Consider $D = [0, 1] \times [0, 1]$ and random field $Z(\mathbf{s}) = \exp\{Y(\mathbf{s})\}$, where $\{Y(\mathbf{s}), \mathbf{s} \in D\}$ is Gaussian with constant mean and Matérn covariance function

$$C(\mathbf{s}, \mathbf{u}) = \frac{\sigma^2}{2^{\theta_2-1}\Gamma(\theta_2)} \left(\frac{l}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{l}{\theta_1}\right),$$

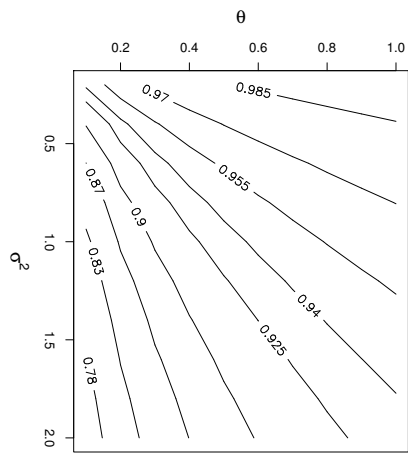
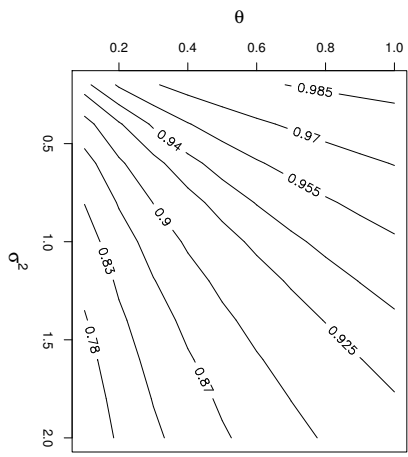
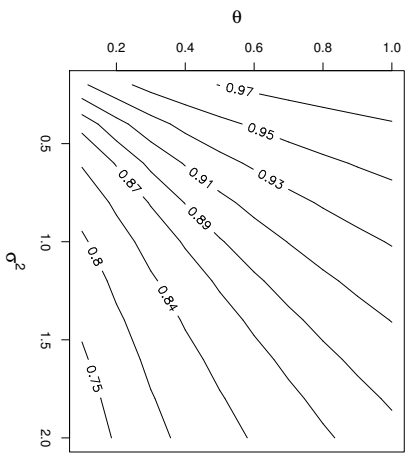
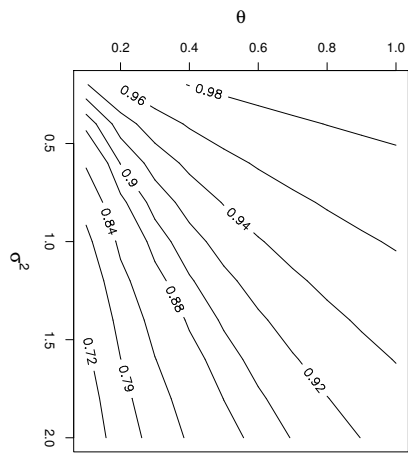
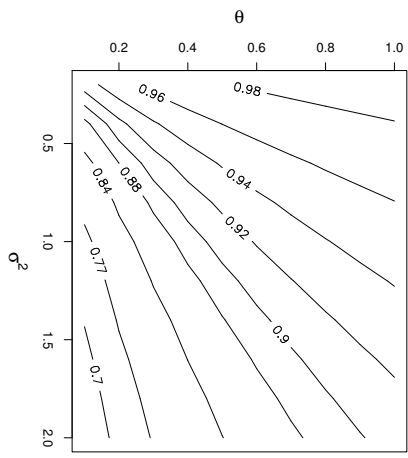
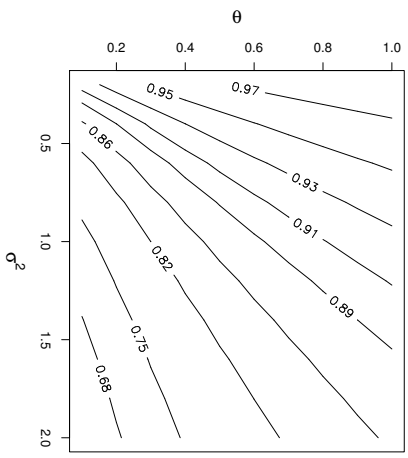
$l = \|\mathbf{s} - \mathbf{u}\|$ is Euclidean distance

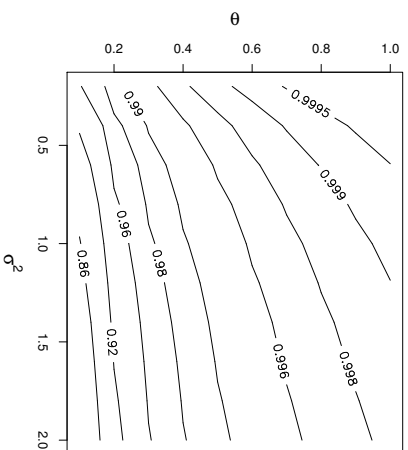
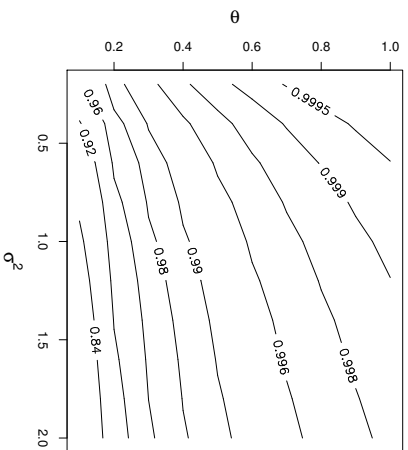
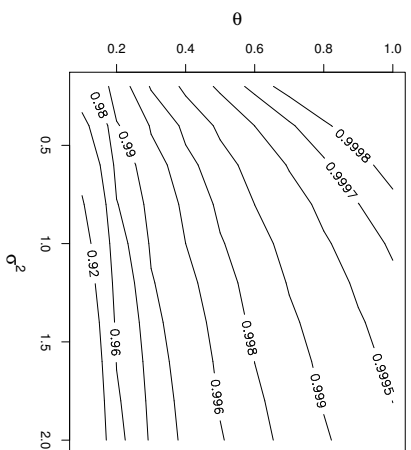
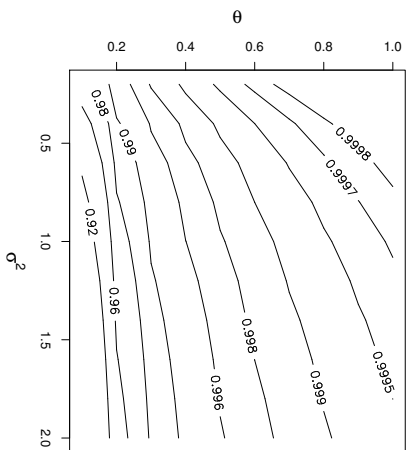
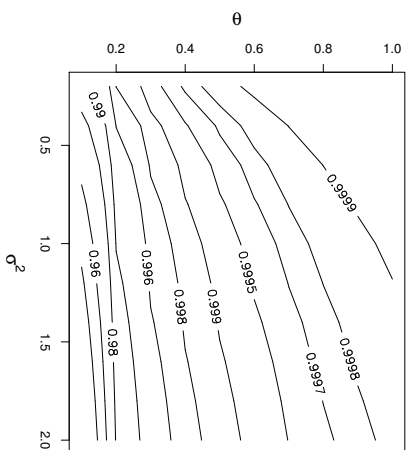
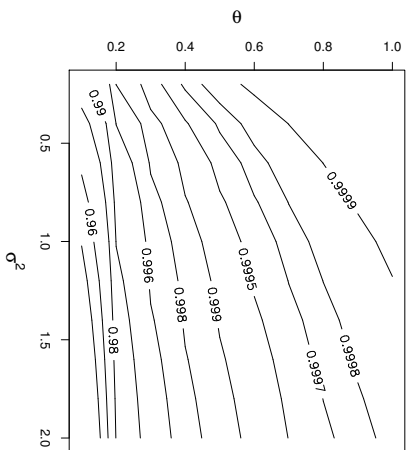
$\boldsymbol{\vartheta} = (\sigma^2, \theta_1, \theta_2)$ are covariance parameters.

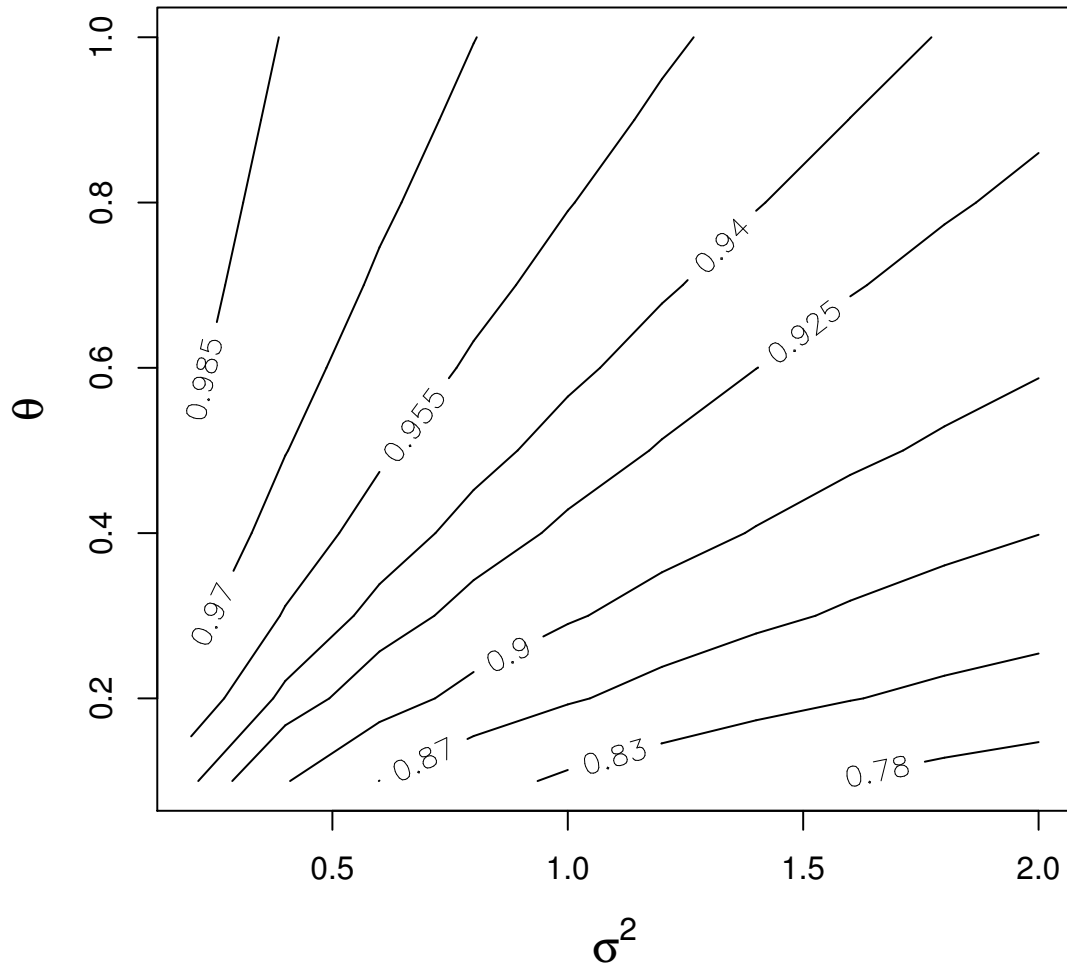
We consider the cases $\theta_2 = 0.5$ and $\theta_2 = 1.5$
Also assume $\sigma_\epsilon^2 = 0$ (no measurement error).



Fifty sampling locations (o) chosen at random within the region D and prediction locations $s_0 = (0.5, 0.5), (0.3, 0.8)$ and $(0.9, 0.9)$ (x).







$$\theta_2 = 0.5$$

Findings:

- Length reductions in the range 1–35%
- Length reductions decrease when confidence level increases
- Length reductions decrease when smoothness of the process increases
- The largest reductions are obtained in models with highly asymmetric marginals (σ^2 large) and moderate to weak dependence (θ small)

SHORTEST PREDICTION INTERVALS: UNKNOWN COVARIANCE CASE

The previous prediction intervals depend on ϑ . The most immediate fix to this problem is to use $I_0^N(\alpha, \hat{\vartheta})$ and $I_0^S(\alpha, \hat{\vartheta})$, where $\hat{\vartheta} = \hat{\vartheta}(\mathbf{z})$ is an estimate of ϑ .

These are called plug-in prediction intervals.

The drawback is that plug-in PIs have coverage properties that differ from the nominal coverage properties, usually having smaller coverage than the desired coverage since these PIs intervals do not take into account the sampling variability of the parameter estimates.

A solution is to calibrate these plug-in PIs: Cox (1975) and Beran (1990).

Calibrated Prediction Intervals

The *coverage probability* function of $I_0^S(\alpha, \hat{\vartheta})$ is defined as

$$\pi_0(\alpha, \vartheta) = P_{\vartheta}\{Z_0 \in I_0^S(\alpha, \hat{\vartheta}(\mathbf{Z}))\}$$

We start by estimating $\pi_0(\cdot, \vartheta)$ with $\pi_0(\cdot, \hat{\vartheta})$. The basic idea of calibrating plug-in prediction intervals is to find $\alpha_c \in (0, 1)$ for which it holds, exactly or approximately, that

$$\pi_0(\alpha_c, \hat{\vartheta}) = 1 - \alpha,$$

The calibrated prediction interval is $I_0^S(\alpha_c, \hat{\vartheta})$, which by construction has coverage probability close to $1 - \alpha$.

$\pi_0(\cdot, \hat{\vartheta})$ is usually not available in closed form so it needs to be approximated.

Bootstrap Calibration

Let $\pi_0^*(\cdot, \hat{\boldsymbol{\vartheta}})$ be a Monte Carlo estimate of $\pi_0(\cdot, \hat{\boldsymbol{\vartheta}})$.

One way is simulate $(\mathbf{Z}_j^*, Z_{0j}^*)$, say B times, from the lognormal model with parameters $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{z})$ and $\hat{\boldsymbol{\vartheta}} = \hat{\boldsymbol{\vartheta}}(\mathbf{z})$, and estimate $\pi_0(x, \hat{\boldsymbol{\vartheta}})$ with

$$\frac{1}{B} \sum_{j=1}^B \mathbf{1}\{\hat{Y}_0(\boldsymbol{\vartheta}_j^*) - \Phi^{-1}(1 - \gamma_{0j}^{\text{opt}*}) \hat{\sigma}_0(\boldsymbol{\vartheta}_j^*) < Y_{0j}^* < \hat{Y}_0(\boldsymbol{\vartheta}_j^*) + \Phi^{-1}(1 - x + \gamma_{0j}^{\text{opt}*}) \hat{\sigma}_0(\boldsymbol{\vartheta}_j^*)\}$$

$$Y_{0j}^* = \log(Z_{0j}^*), \boldsymbol{\vartheta}_j^* = \hat{\boldsymbol{\vartheta}}(\mathbf{Z}_j^*) \text{ and } \gamma_{0j}^{\text{opt}*} = \gamma_0^{\text{opt}}(x, \boldsymbol{\vartheta}_j^*).$$

A better way is to use ‘Rao-Blackwellization’ based on the identity

$$\pi_0(\alpha, \boldsymbol{\vartheta}) = E_{\boldsymbol{\vartheta}} \left\{ \Phi\left(\frac{U_0(\alpha, \hat{\boldsymbol{\vartheta}}, \mathbf{Y}) - \eta_0(\mathbf{0}, \boldsymbol{\vartheta}, \mathbf{Y})}{\tau_0(\boldsymbol{\vartheta})}\right) - \Phi\left(\frac{L_0(\alpha, \hat{\boldsymbol{\vartheta}}, \mathbf{Y}) - \eta_0(\mathbf{0}, \boldsymbol{\vartheta}, \mathbf{Y})}{\tau_0(\boldsymbol{\vartheta})}\right) \right\}$$

$$L_0(x, \boldsymbol{\vartheta}, \mathbf{Y}) = \hat{Y}_0(\boldsymbol{\vartheta}) - \Phi^{-1}(1 - \gamma_0^{\text{opt}}) \hat{\sigma}_0(\boldsymbol{\vartheta})$$

$$U_0(x, \boldsymbol{\vartheta}, \mathbf{Y}) = \hat{Y}_0(\boldsymbol{\vartheta}) + \Phi^{-1}(1 - x + \gamma_0^{\text{opt}}) \hat{\sigma}_0(\boldsymbol{\vartheta})$$

Expectation is wrt \mathbf{Y} when $\boldsymbol{\beta} = \mathbf{0}$.

Algorithm:

- Step 1.* Compute the ML (or REML) estimate $\hat{\boldsymbol{\vartheta}} = \hat{\boldsymbol{\vartheta}}(\mathbf{z})$ from the observed data \mathbf{z} .
- Step 2.* Simulate B independent and identically distributed bootstrap samples $\{\mathbf{Y}_j^* : 1 \leq j \leq B\}$ from the Gaussian random field $\{Y(\mathbf{s}), \mathbf{s} \in D\}$ with $\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}$.
- Step 3.* For each $j = 1, \dots, B$, compute the estimate $\boldsymbol{\vartheta}_j^* = \hat{\boldsymbol{\vartheta}}(\exp(\mathbf{Y}_j^*))$ based on the bootstrap sample \mathbf{Y}_j^* .
- Step 4.* For each $s_0 \in D$ where a PI is sought, compute $L_{0j}^* = L_0(x, \boldsymbol{\vartheta}_j^*, \mathbf{Y}_j^*)$, $U_{0j}^* = U_0(x, \boldsymbol{\vartheta}_j^*, \mathbf{Y}_j^*)$ and for $x \in (0, 1)$ estimate $\pi_0(x, \hat{\boldsymbol{\vartheta}})$ by

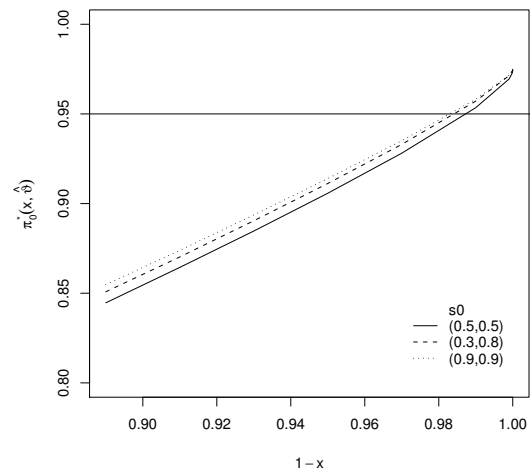
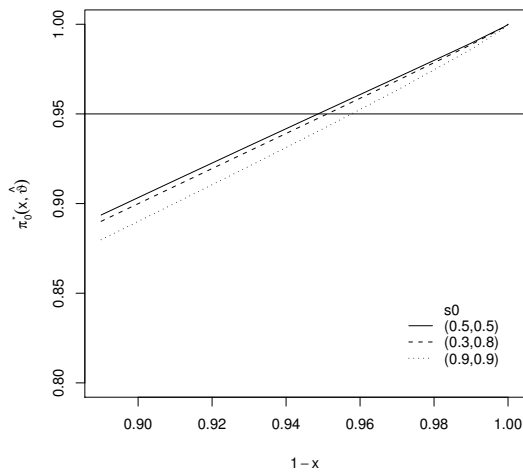
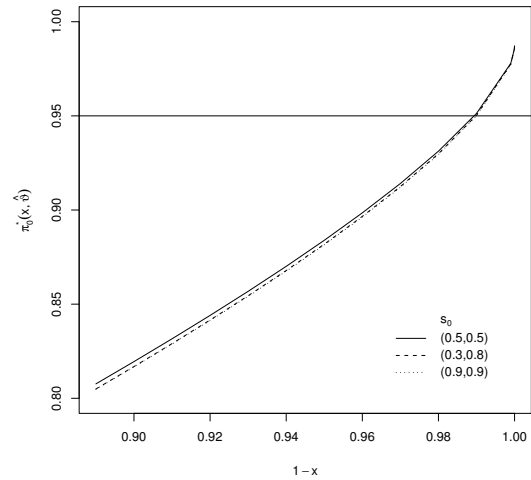
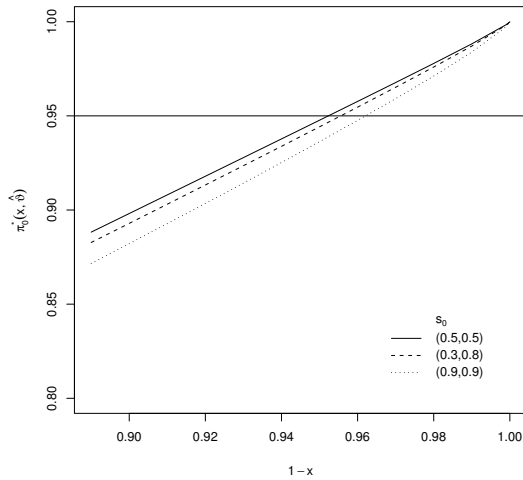
$$\pi_0^*(x, \hat{\boldsymbol{\vartheta}}) = \frac{1}{B} \sum_{j=1}^B \left[\Phi\left(\frac{U_{0j}^* - \hat{\eta}_{0j}^*}{\hat{\tau}_0}\right) - \Phi\left(\frac{L_{0j}^* - \hat{\eta}_{0j}^*}{\hat{\tau}_0}\right) \right]$$

where $\hat{\eta}_{0j}^* = \eta_0(\mathbf{0}, \hat{\boldsymbol{\vartheta}}, \mathbf{Y}_j^*)$ and $\hat{\tau}_0 = \tau_0(\hat{\boldsymbol{\vartheta}})$.

Finally, α_c is found as the solution (in x) of

$$\pi_0^*(x, \hat{\boldsymbol{\vartheta}}) = (1 - \alpha)$$

Illustration:



When the data have no nugget the effect of calibration is often minor.

But when the data contain measurement error the effect of calibration tends to be substantial.

Comparison of plug-in and calibrated PIs

| θ | σ^2 | 0.1 | | | 1.0 | |
|---------------------|-----------------|-----------------|------------------|--|-----------------|--|
| | 0.1 | 0.5 | 0.1 | | 0.5 | |
| | | | | $\sigma_\epsilon^2 = 0$ | | |
| plug-in standard | 1.068 [.939] | 0.551 [.951] | 5.620 [.940] | | 3.004 [.948] | |
| plug-in shortest | 1.033 [.939] | 0.545 [.950] | 4.450 [.942] | | 2.779 [.948] | |
| calibrated standard | 1.124 [.949] | 0.557 [.953] | 6.111 [.951] | | 3.050 [.951] | |
| calibrated shortest | 1.068 [.946] | 0.548 [.951] | 4.694 [.949] | | 2.795 [.949] | |
| | | | | $\sigma_\epsilon^2 = \frac{\sigma^2}{4}$ | | |
| plug-in standard | 1.011 [.899] | 0.522 [.895] | 6.131 [.899] | | 2.897 [.896] | |
| plug-in shortest | 0.981 [.900] | 0.518 [.895] | 4.908 [.909] | | 2.686 [.902] | |
| calibrated standard | 1.164 [.934] | 0.544 [.908] | 8.597 [.936] | | 3.041 [.910] | |
| calibrated shortest | 1.098 [.928] | 0.531 [.904] | 6.750 [.937] | | 2.775 [.911] | |
| | | | | $\sigma_\epsilon^2 = 0$ | | |
| plug-in standard | 1.225 [.933] | 0.727 [.936] | 6.507 [.934] | | 4.025 [.933] | |
| plug-in shortest | 1.174 [.934] | 0.716 [.937] | 4.945 [.937] | | 3.537 [.936] | |
| calibrated standard | 1.313 [.948] | 0.779 [.952] | 7.367 [.949] | | 4.422 [.949] | |
| calibrated shortest | 1.232 [.944] | 0.750 [.948] | 5.400 [.947] | | 3.776 [.947] | |
| | | | | $\sigma_\epsilon^2 = \frac{\sigma^2}{4}$ | | |
| plug-in standard | 1.175 [.891] | 0.697 [.896] | 6.960 [.890] | | 3.907 [.896] | |
| plug-in shortest | 1.129 [.894] | 0.686 [.897] | 5.345 [.903] | | 3.445 [.901] | |
| calibrated standard | 1.377 [.932] | 0.761 [.923] | 10.168 [.933] | | 4.390 [.923] | |
| calibrated shortest | 1.285 [.927] | 0.731 [.917] | 7.636 [.935] | | 3.755 [.921] | |

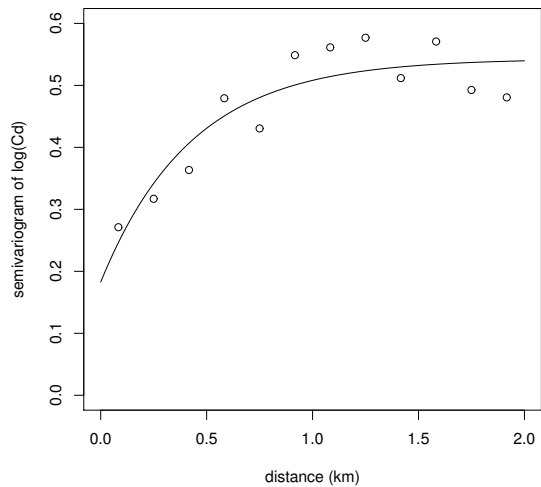
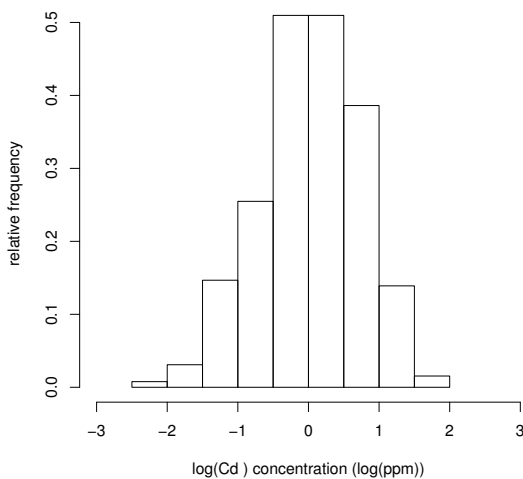
EXAMPLE:

Data on cadmium (Cd) concentrations (in ppm) measured at 259 locations in a region of about 15 km² in the Switzerland, collected in 1992.

Measurements at 100 locations are used for validation.

Exploratory analysis suggests the “best” model is the log-Gaussian random field associated with constant mean, nugget and exponential covariance function:

$$\hat{\beta}_1 = 0.084, \hat{\sigma}^2 = 0.4, \hat{\theta}_1 = 0.177 \text{ and } \hat{\sigma}_\epsilon^2 = 0.073.$$



| | plug-in | | | calibrated | | |
|----|----------|----------|-------|------------|----------|-------|
| | standard | shortest | RL | standard | shortest | RL |
| 1 | 1.169 | 1.053 | 0.900 | 1.176 | 1.053 | 0.890 |
| 2 | 4.763 | 4.201 | 0.882 | 4.853 | 4.230 | 0.871 |
| 3 | 4.527 | 3.908 | 0.863 | 4.663 | 3.999 | 0.858 |
| 4 | 3.550 | 3.094 | 0.871 | 3.636 | 3.130 | 0.861 |
| 5 | 3.337 | 2.881 | 0.863 | 3.433 | 2.940 | 0.856 |
| 6 | 2.999 | 2.615 | 0.872 | 3.100 | 2.680 | 0.864 |
| 7 | 3.984 | 3.494 | 0.877 | 4.075 | 3.568 | 0.875 |
| 8 | 2.738 | 2.363 | 0.863 | 2.832 | 2.401 | 0.847 |
| 9 | 2.813 | 2.479 | 0.881 | 2.873 | 2.502 | 0.871 |
| 10 | 3.817 | 3.326 | 0.871 | 3.911 | 3.393 | 0.867 |

The relative lengths of the shortest prediction intervals with respect to the standard prediction intervals vary from 0.84 to 0.89, with an average of 0.86, so on average the shortest prediction intervals are about 14% shorter than the standard prediction intervals.

For each of the 100 validation locations computed each type of 95% prediction interval for the Cd true value, and determined whether or not each Cd observed value falls into the corresponding prediction interval.

The proportion of 95% plug-in standard, plug-in shortest, calibrated standard and calibrated shortest prediction intervals covering the corresponding Cd observed values were, respectively, 0.93, 0.93, 0.93 and 0.95. The calibrated shortest prediction intervals seem to have coverage close to nominal.

A Tale about Hockey Sticks



g-north@tamu.edu

Before the Tale, A few Chance Encounters



Gerald R. North
Atmos. Sci.
Texas A&M

Sampling Errors were a concern for the proposed Tropical Rainfall Measuring Mission.

- 1985

TRMM



I was trying to learn statistics.
I read engineering books,
economics books, Bulmer, etc.
Then . . .



One of the key papers demonstrating the plausibility of TRMM

Estimation of Mean Rain Rate: Application to Satellite Observations

BENJAMIN KEDEM¹

Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park

LONG S. CHIU²

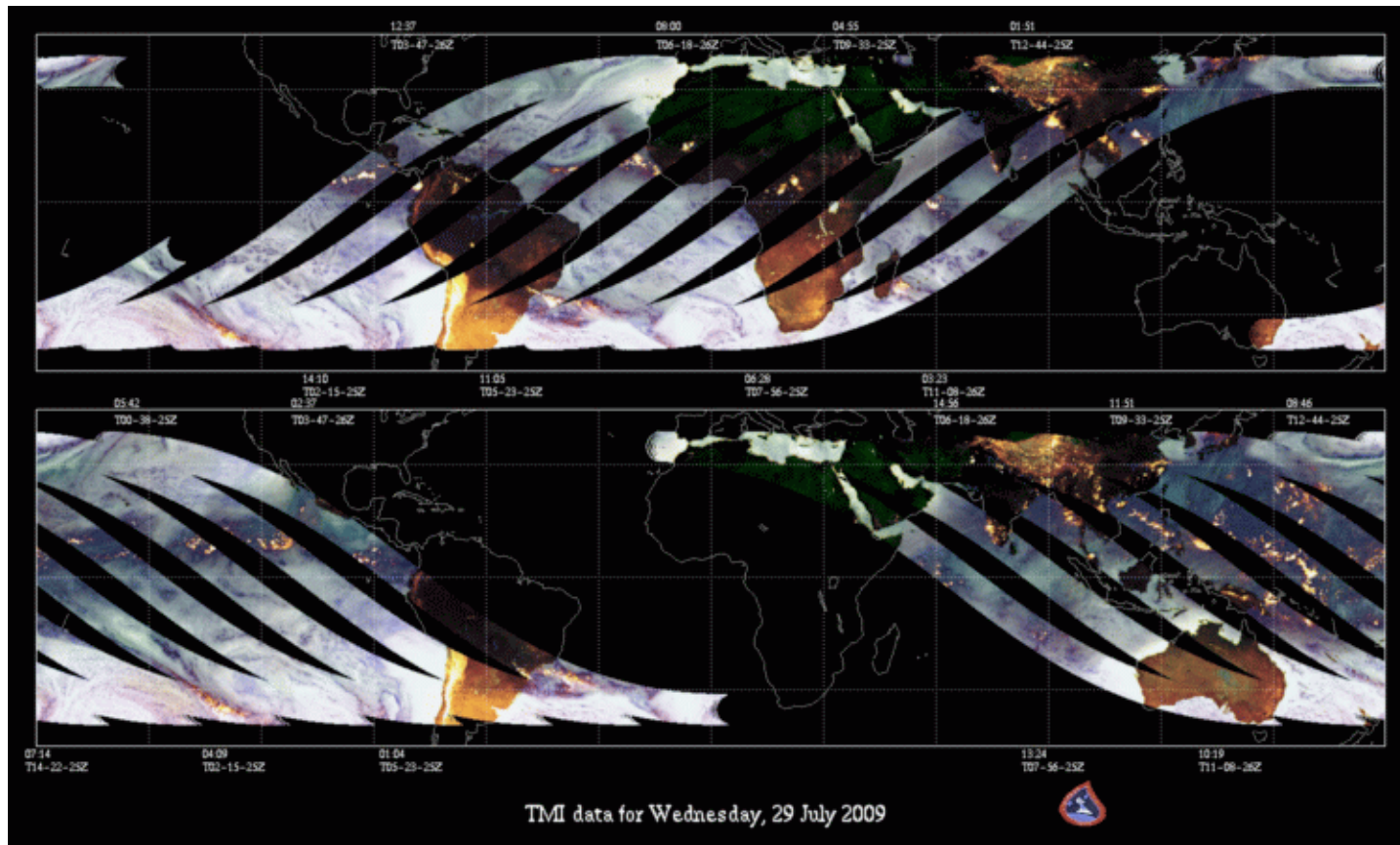
Applied Research Corporation, Landover, Maryland

GERALD R. NORTH³

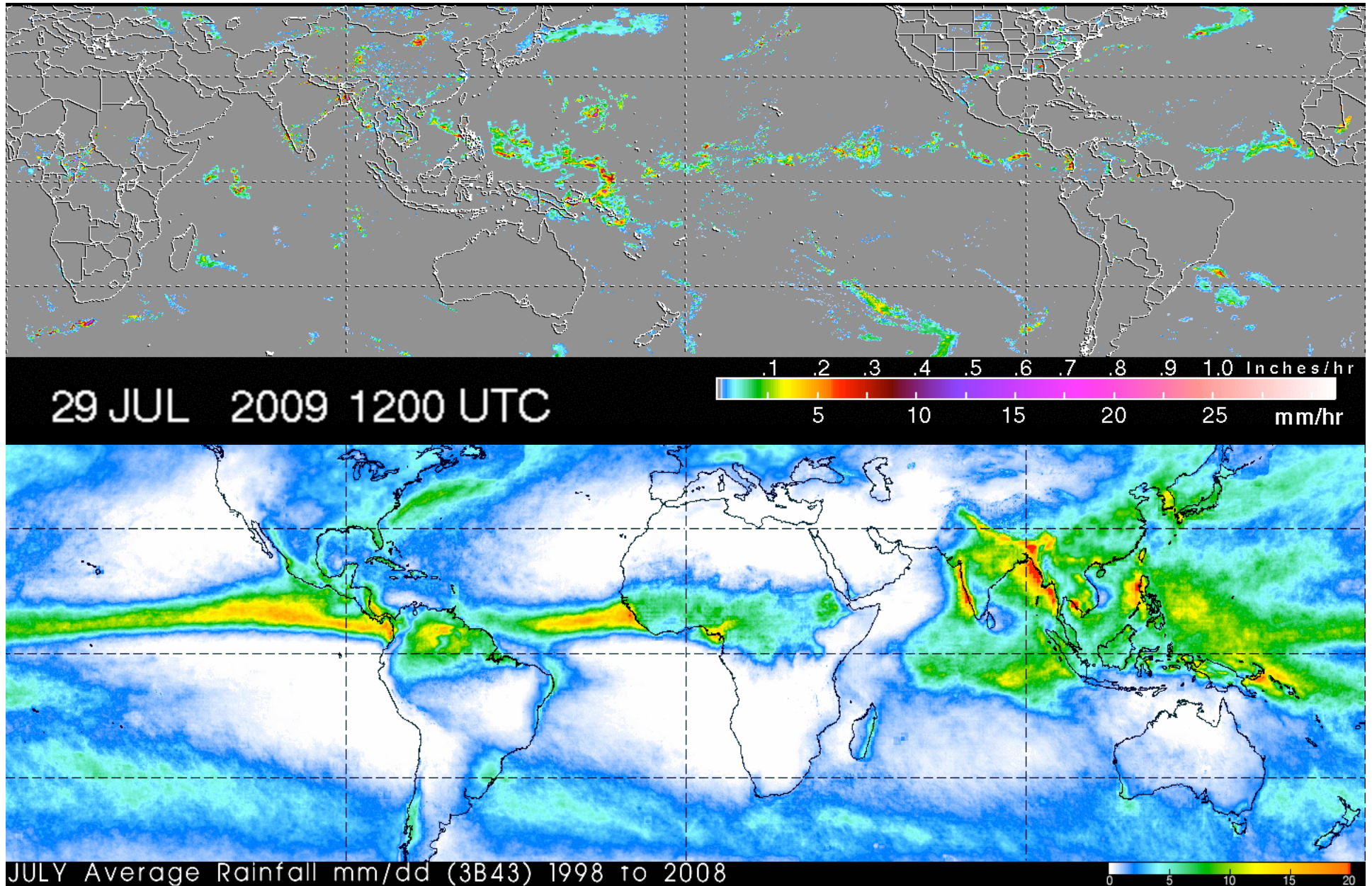
NASA Goddard Space Flight Center, Greenbelt, Maryland

A method for the estimation of the mean area average rain rate from dependent data is developed and applied to the GARP Atlantic Tropical Experiment GATE data. The method consists of fitting a mixed distribution, containing an atom at zero, by minimum chi-square in combination with certain time-space sampling designs. In modeling the continuous component of the mixed distribution it is shown that the lognormal distribution provides a very close fit for the nonzero area average rainrates. A comparison with the gamma distribution shows that the lognormal distribution is a better choice as expressed by the minimum chi-square criterion. Some of the time-space sampling designs correspond to satellite sampling. The results indicate that a satellite visiting an area of about $350 \times 350 \text{ km}^2$ in the tropics approximately every 10 hours over a period can provide a rather close estimate for the mean area average rain rate.

TRMM Orbits



Launched in 1997, TRMM is still flying



Not all of my encounters
were uncorrelated



Ann. Inst. Statist. Math.

Vol. 49, No. 2, 341–354 (1997)

ALIASING EFFECTS AND SAMPLING THEOREMS OF
SPHERICAL RANDOM FIELDS WHEN SAMPLED
ON A FINITE GRID

TA-HSIN LI¹ AND GERALD R. NORTH²

¹*Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.*

²*Climate System Research Program, Department of Meteorology,
Texas A&M University, College Station, TX 77843, U.S.A.*

Just a sample:

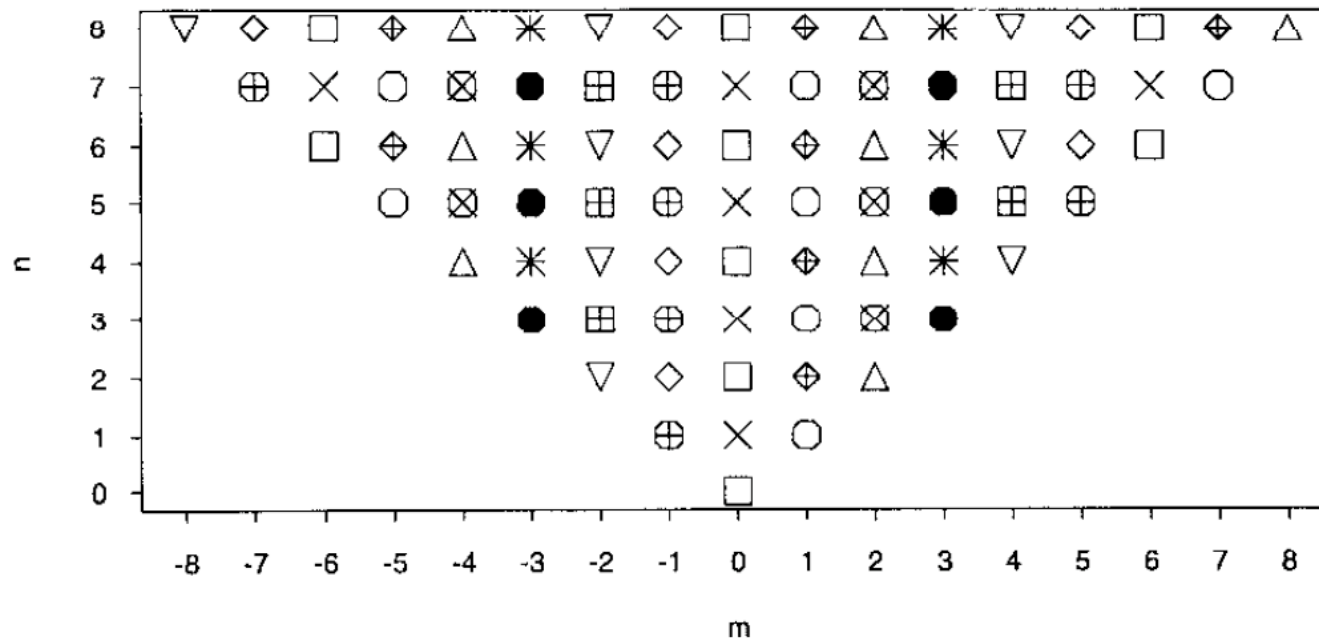
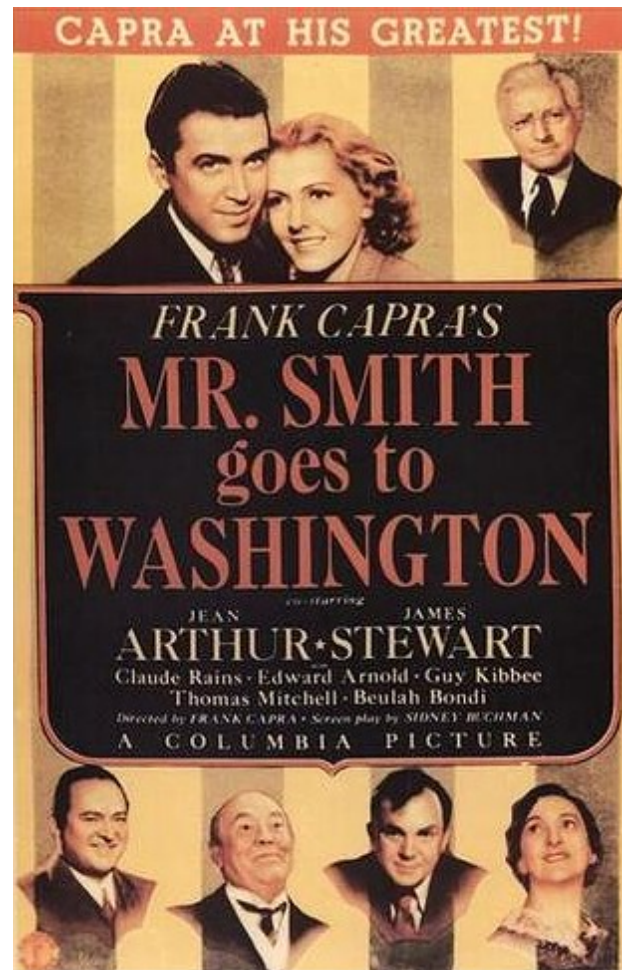


Fig. 1. Aliasing effects in \tilde{T}_{mn} when $M = 3$ (N arbitrary): Coordinates of possible mutual aliases are represented with identical symbols.

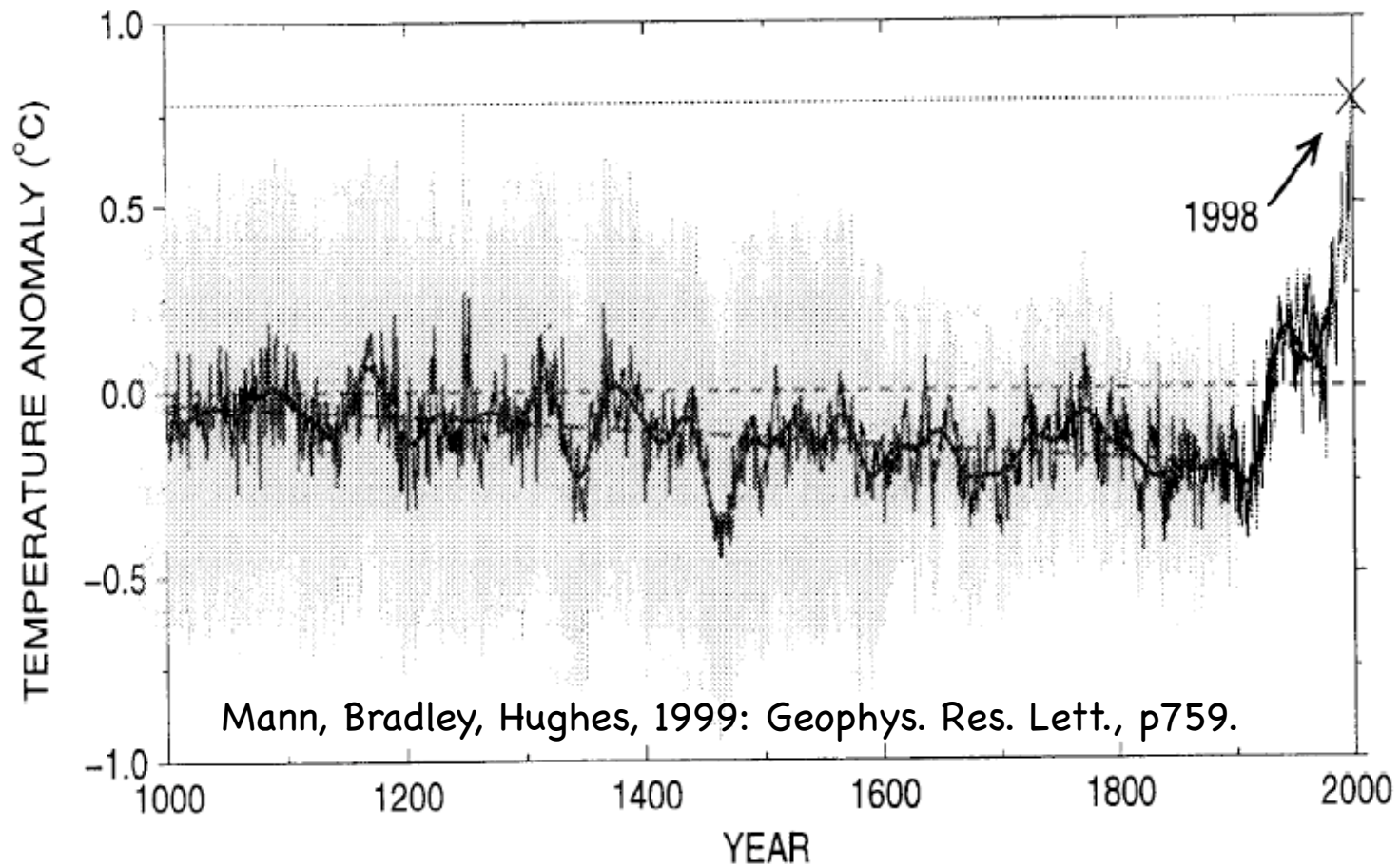
Fast Forward to the Present Century



Global Warming Goes to Washington (2006)



Did Earth cool gradually, then heat up fast?



It's the Notorious Hockey Stick!



↑
1000AD

↑
2000AD

http://en.wikipedia.org/wiki/Hockey_stick_controversy



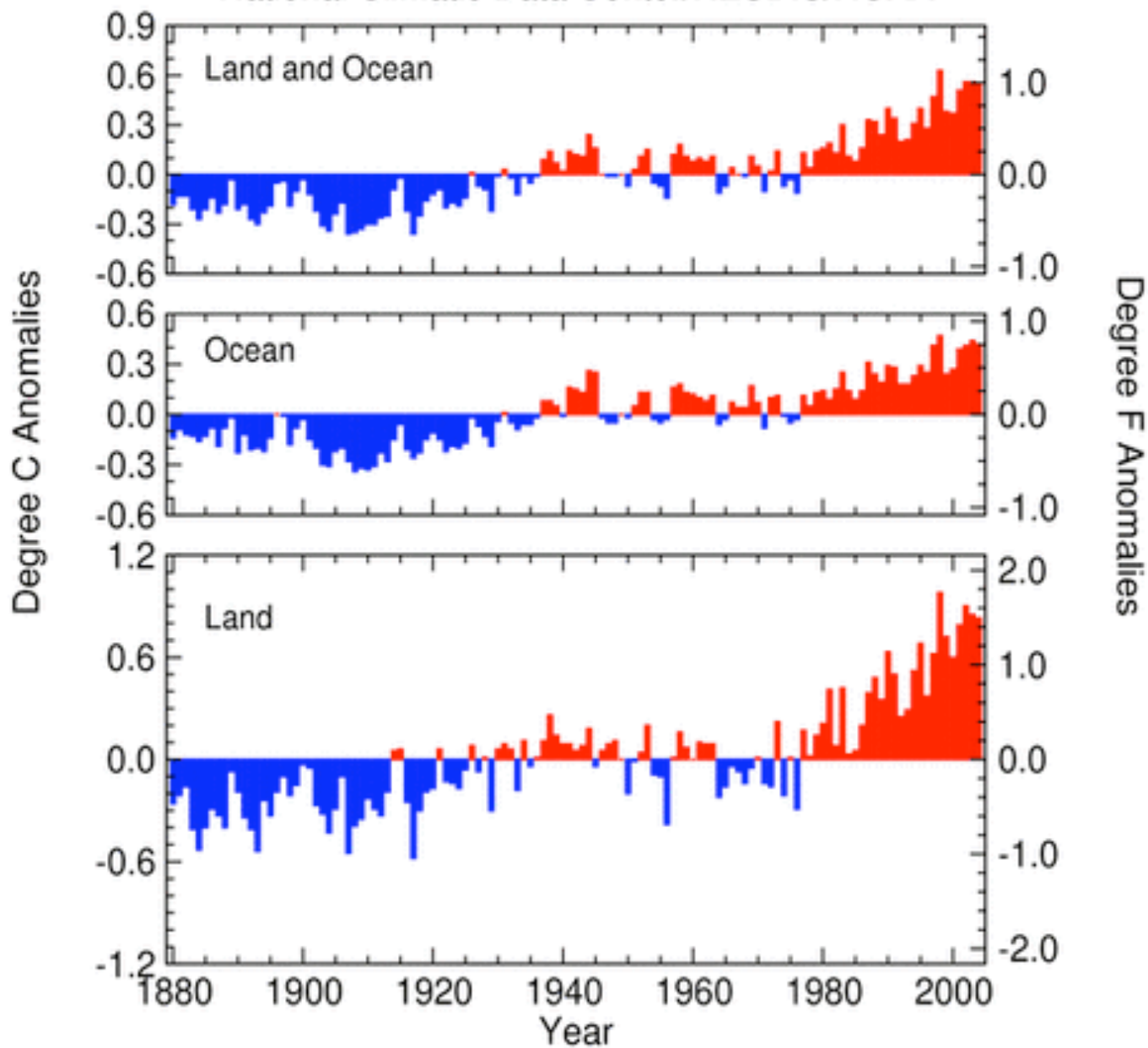
Past Climates can be Estimated from Proxy Data.

Instrument records go back over a century.

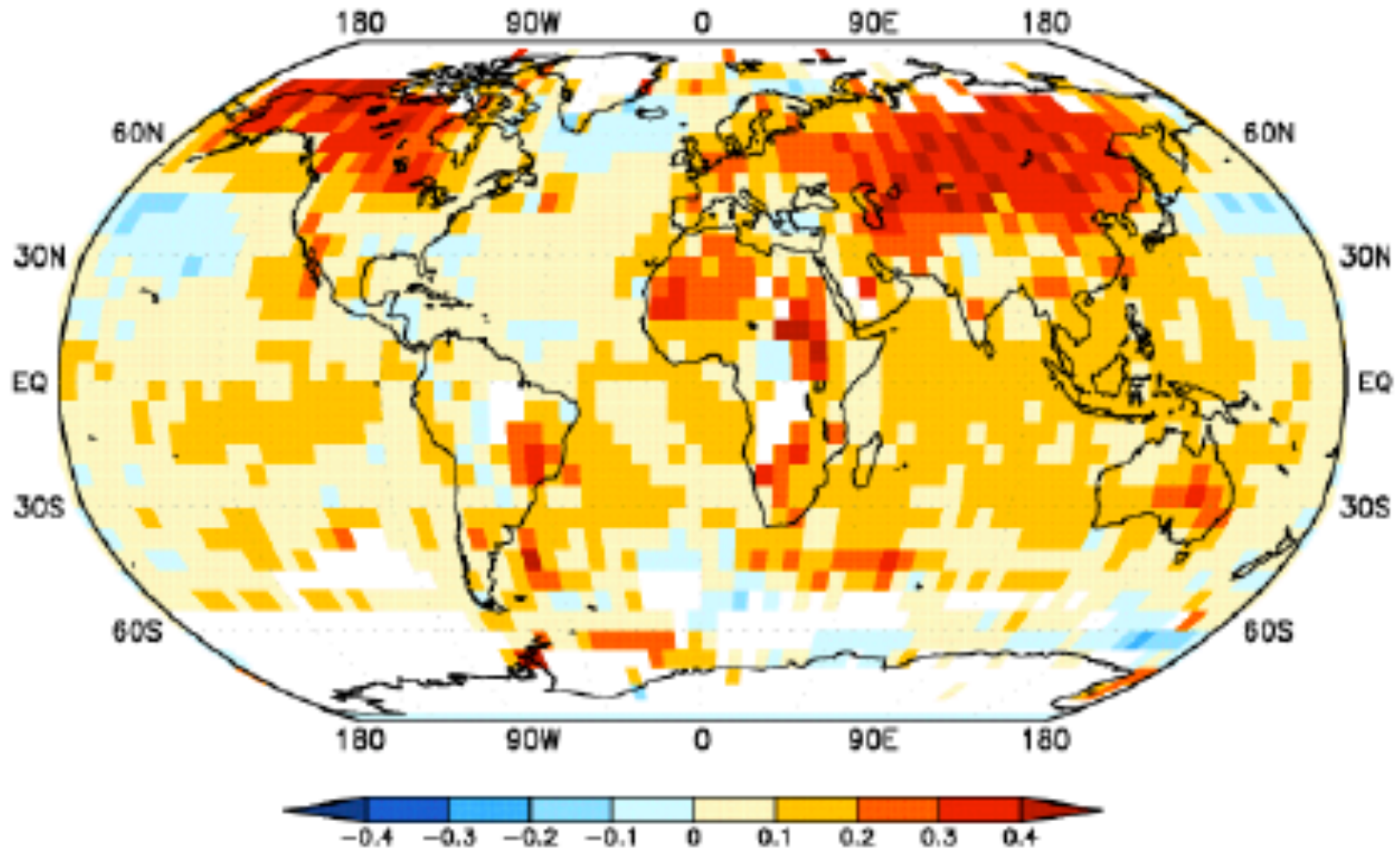


Jan - Dec Global Surface Mean Temp Anomalies

National Climatic Data Center/NESDIS/NOAA



Trend in deg C/decade since 1950



NRC Report 2006

Hockey Stick Time Line

- Mann, Hughes, Bradley (1998, 1999)
- Intergovernmental Panel on Climate Change (IPCC 2001)
- Enter the Amateurs (M&M, 2005)
- Enter Congressman Barton, then Boehlert (06)
- Enter the National Academy of Sciences
- Battling Banjos on the Hill

2001 IPCC features the Hockey Stick: it becomes an Icon

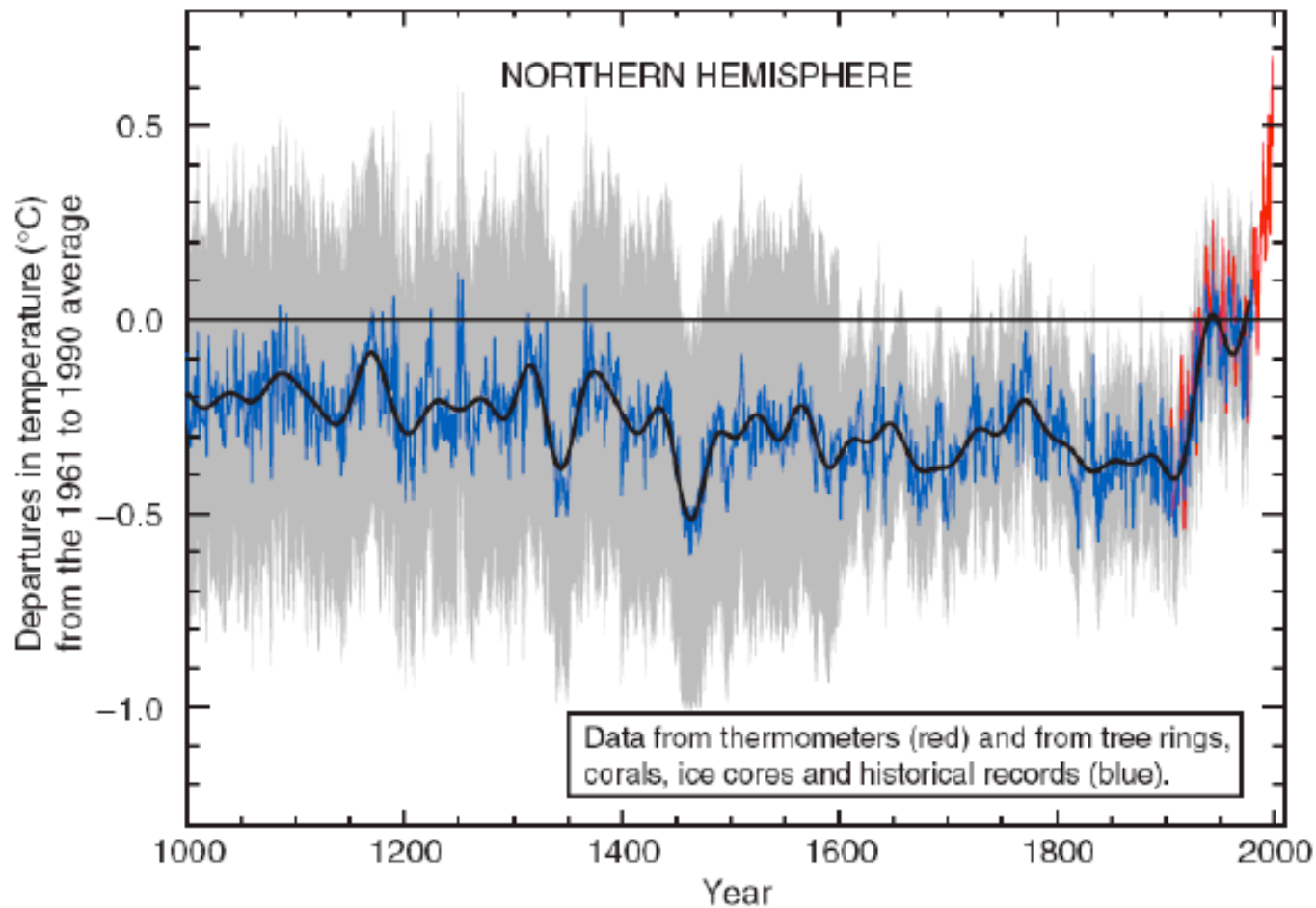


FIGURE O-4 Multiproxy reconstruction of Northern Hemisphere surface temperature variations over the past millennium (blue), along with 50-year average (black), a measure of the statistical uncertainty associated with the reconstruction (grey), and instrumental surface temperature data for the last 150 years (red), based on the work by Mann et al. (1999). This figure has sometimes been referred to as the “hockey stick.” SOURCE: IPCC (2001).

Enter the Amateurs (M&M, 2005)

Steve
McIntyre



A very crude estimate of global temps was featured in the 2000 IPCC Report.

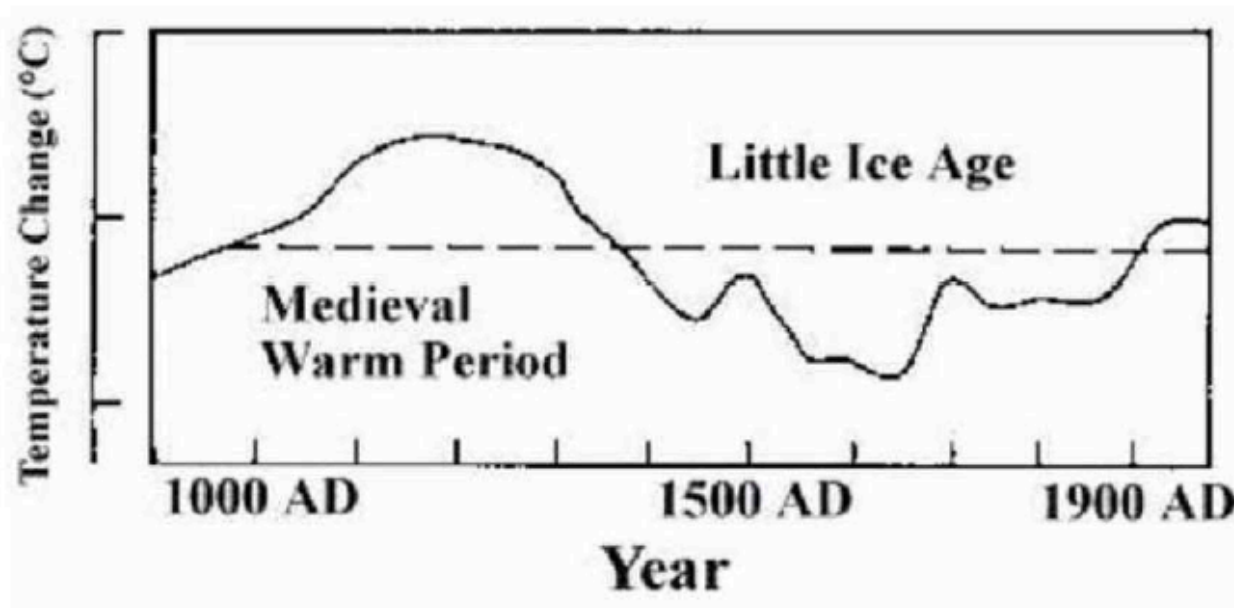


FIGURE O-3 Schematic description of global temperature variations in degrees Centigrade for the last 1 100 years published more than 15 years ago. SOURCE: IPCC (1990)

Enter Congressman Joe Barton



Chairman, Energy &
Commerce Committee
(2006)

- Wegman Report



Wegman,
Scott,
Said

SURFACE TEMPERATURE RECONSTRUCTIONS FOR THE LAST 2,000 YEARS

*Committee on Surface Temperature Reconstructions for the
Last 2,000 Years*

*Board on Atmospheric Sciences and Climate Division on
Earth and Life Studies*

**NATIONAL RESEARCH COUNCIL OF THE NATIONAL
ACADEMIES THE NATIONAL ACADEMIES PRESS**

Washington, D.C.

North, Wallace, Christy,
Cuffey, Turekian,
Druffle, Otto-Bliesner,
Nychka, Bloomfield,
Biondi, Roberts, Dickinson

12 Anonymous Referees, 2
Monitors



NRC Report

Jerry & Ed
with 1990
graphic.

From
Geotimes,
Sept 2006



Climatologist Gerald North (foreground) and statistician Edward Wegman testified in front of the House Energy and Commerce Subcommittee on Oversight and Investigations in July about the famed hockey stick climate analysis. Photograph is by Christine McCarty, House Committee on Energy and Commerce.

Kinds of Information Available

Thermometer Records for the Last 150 years

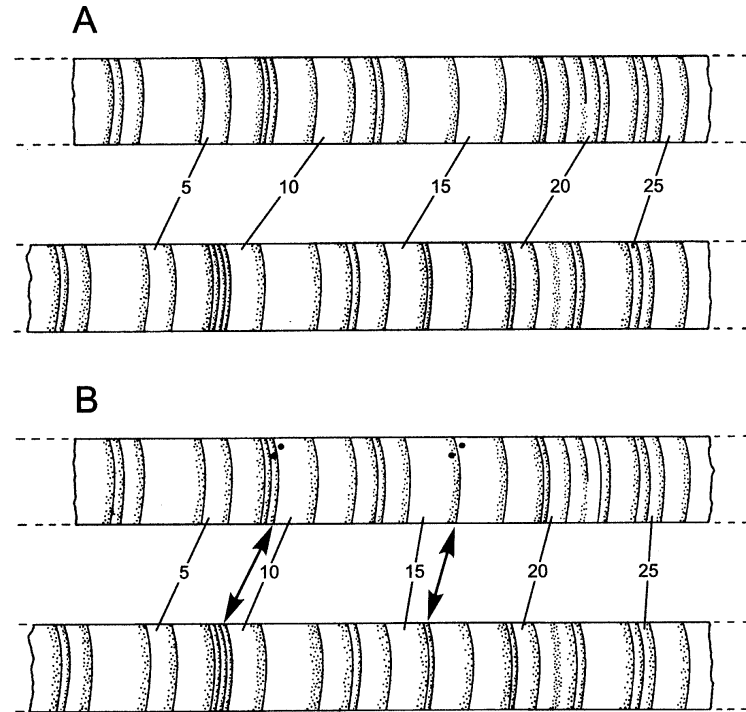
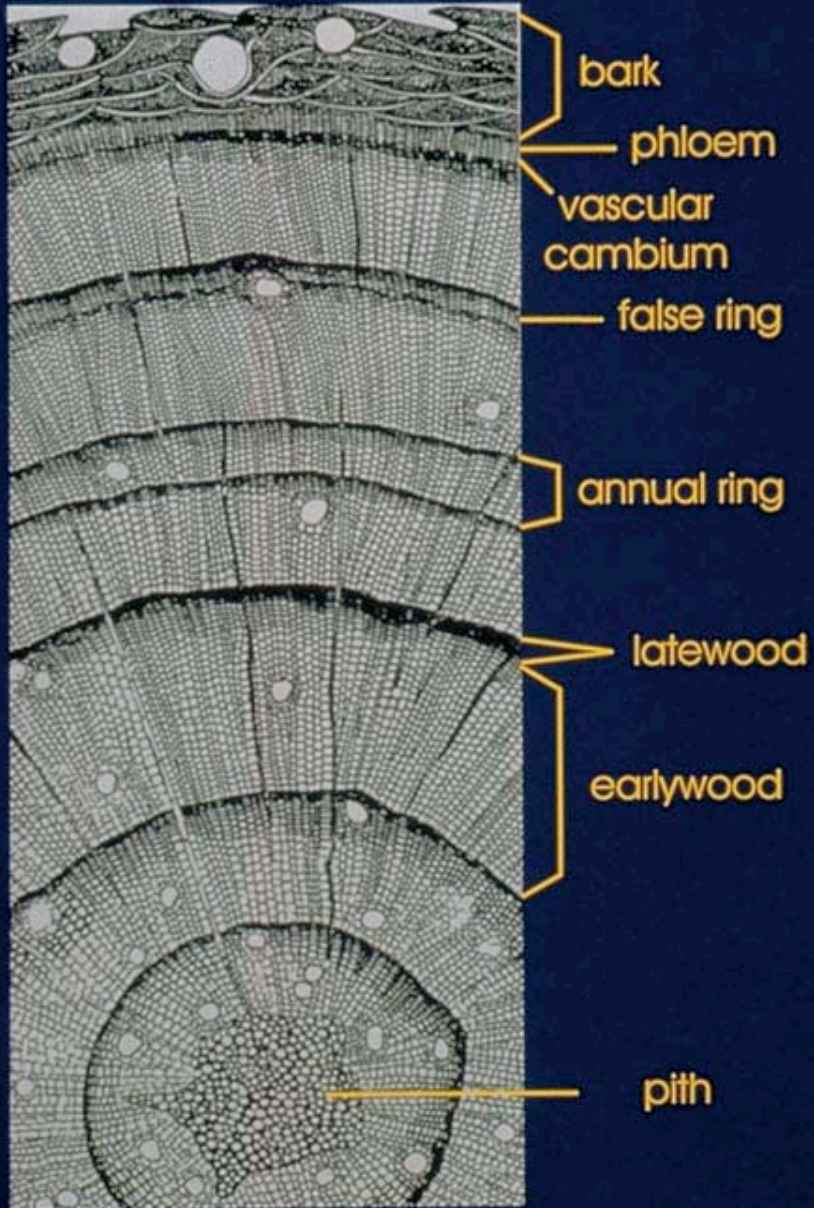
Proxies for Extrapolation Back in Time:

- Tree Rings
- Ice Cores and Isotopes
- Bore Holes (ice and ground)
- Glacial Lengths
- Historical Records
- Sediments (lake and oceanic)
- Corals, Pollen, Caves, Entomology

How to Reconstruct the Field (Tree Ring Example)

- Ring Widths are Correlated to Environmental Conditions
- Use the Instrumental Period to Set the Temp Scale on the Ring Widths
- Statistical Issues:
 - Stationarity, Confounding Variables
 - Range Match
 - Verification Period

CROSS SECTION of a CONIFER



Making a Record
from many Trees

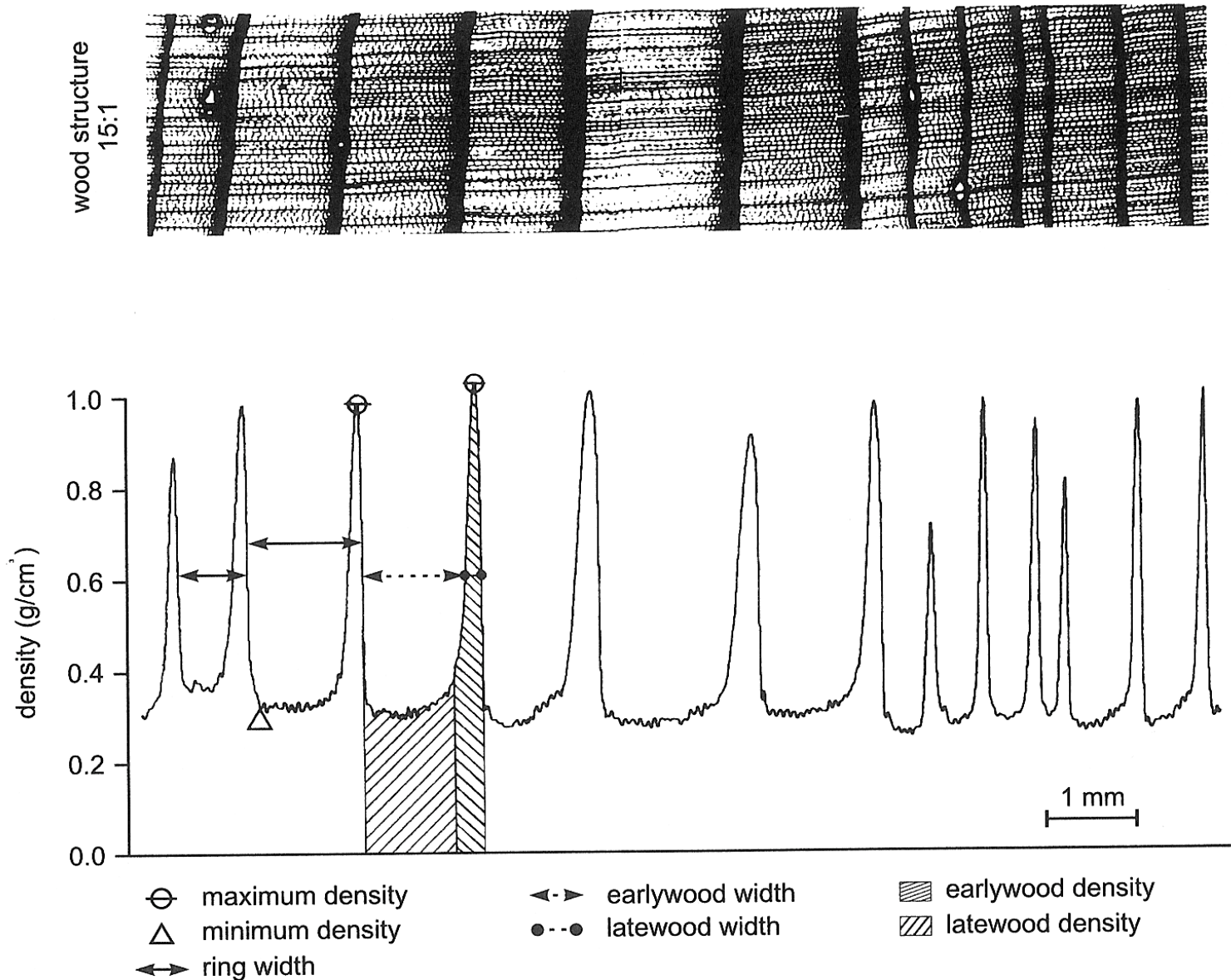
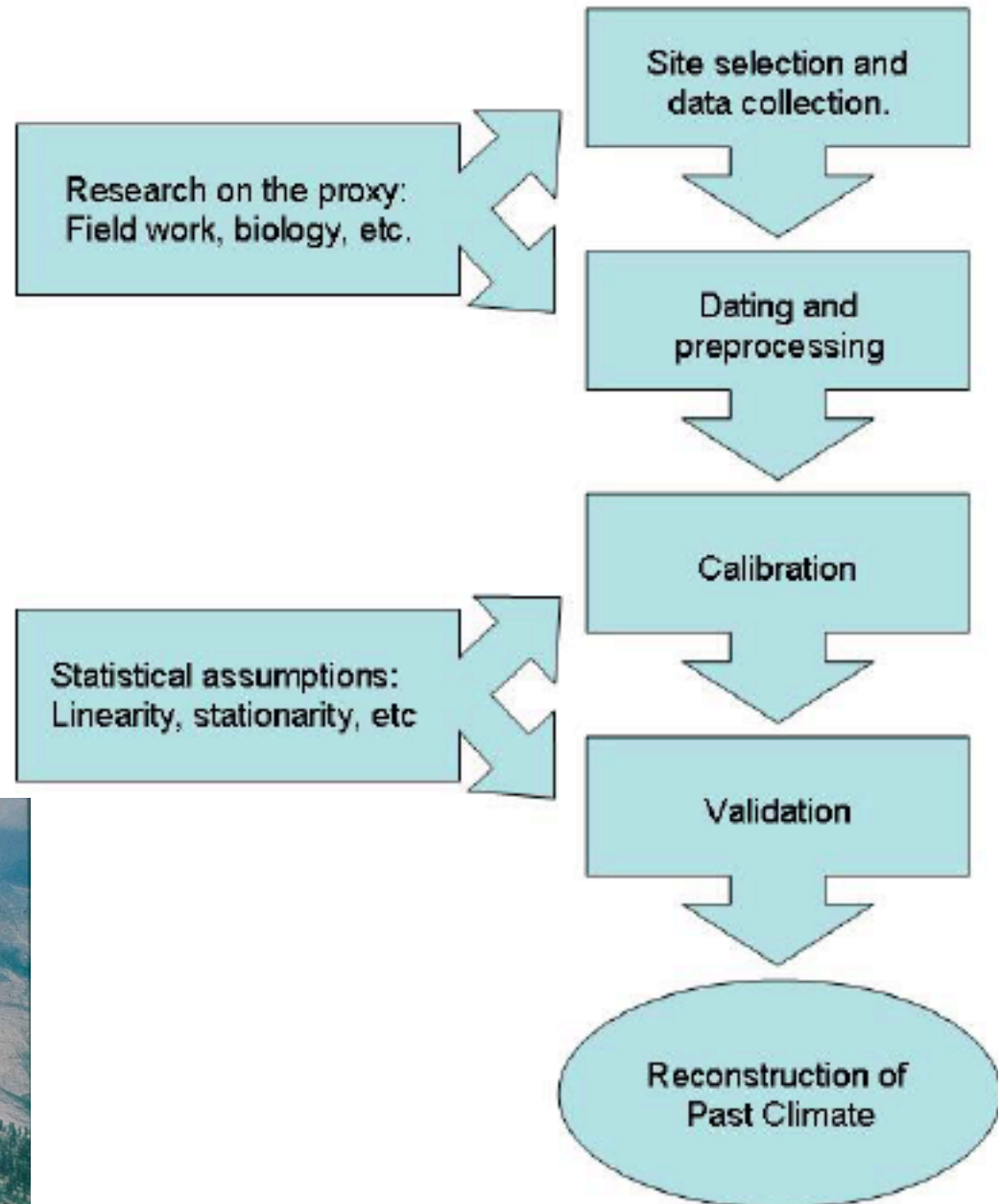


FIGURE 10.2 Example of a tree-ring density plot based on an x-ray negative of a section of wood (top of figure). Minimum and maximum densities in each annual ring are clearly seen, enabling the annual ring width to be measured as well as the width of both the earlywood and latewood (courtesy of F. Schweingruber).

Reconstructing a Temperature Field using proxies, e.g., tree rings



Tree Ring Sites

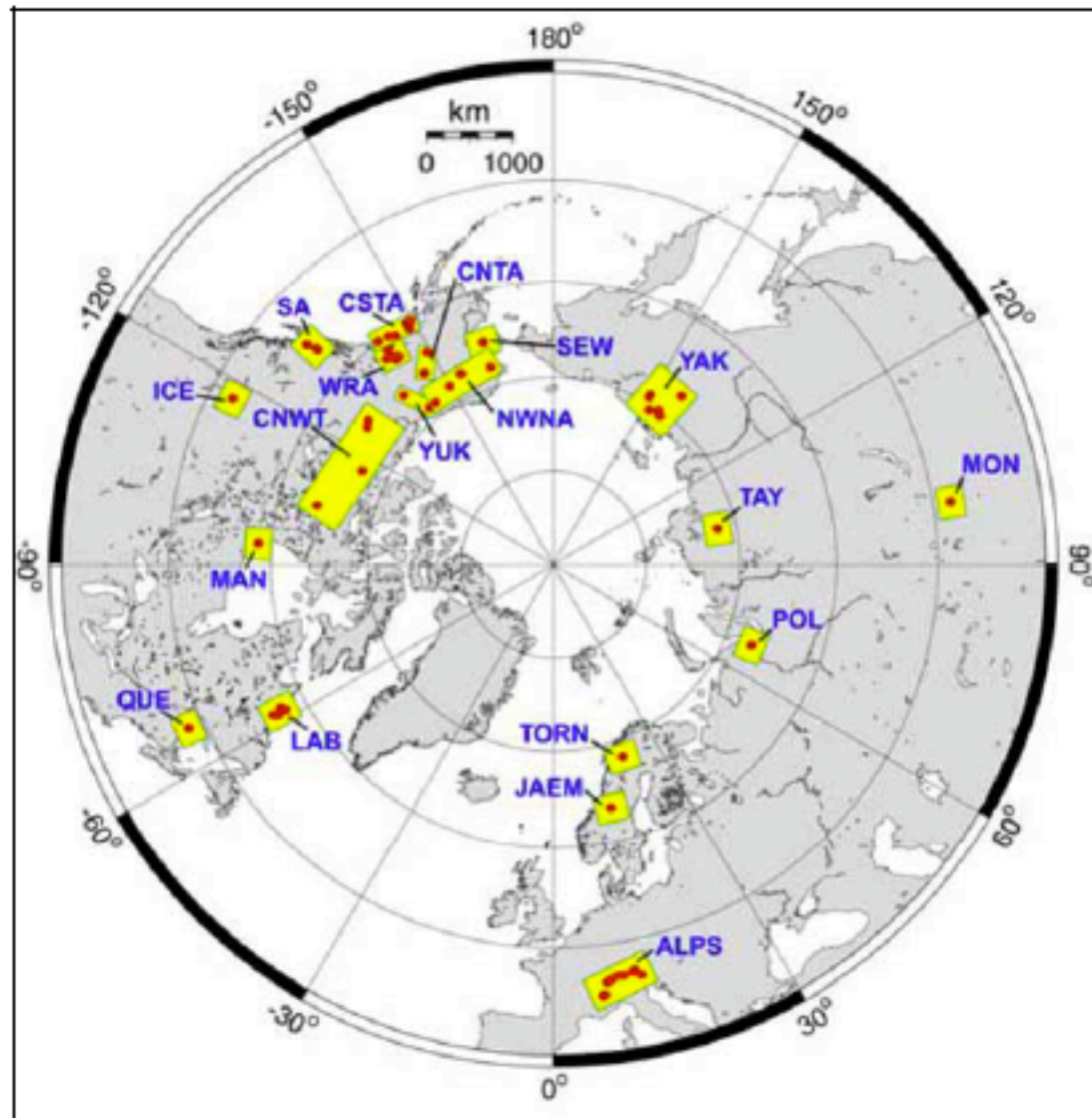


FIGURE 4-1 Location map of individual sites (red) and regional composites (yellow boxes) used to reconstruct Northern Hemisphere surface temperatures for the past millennium. SOURCE: D'Arrigo et al. (2006).

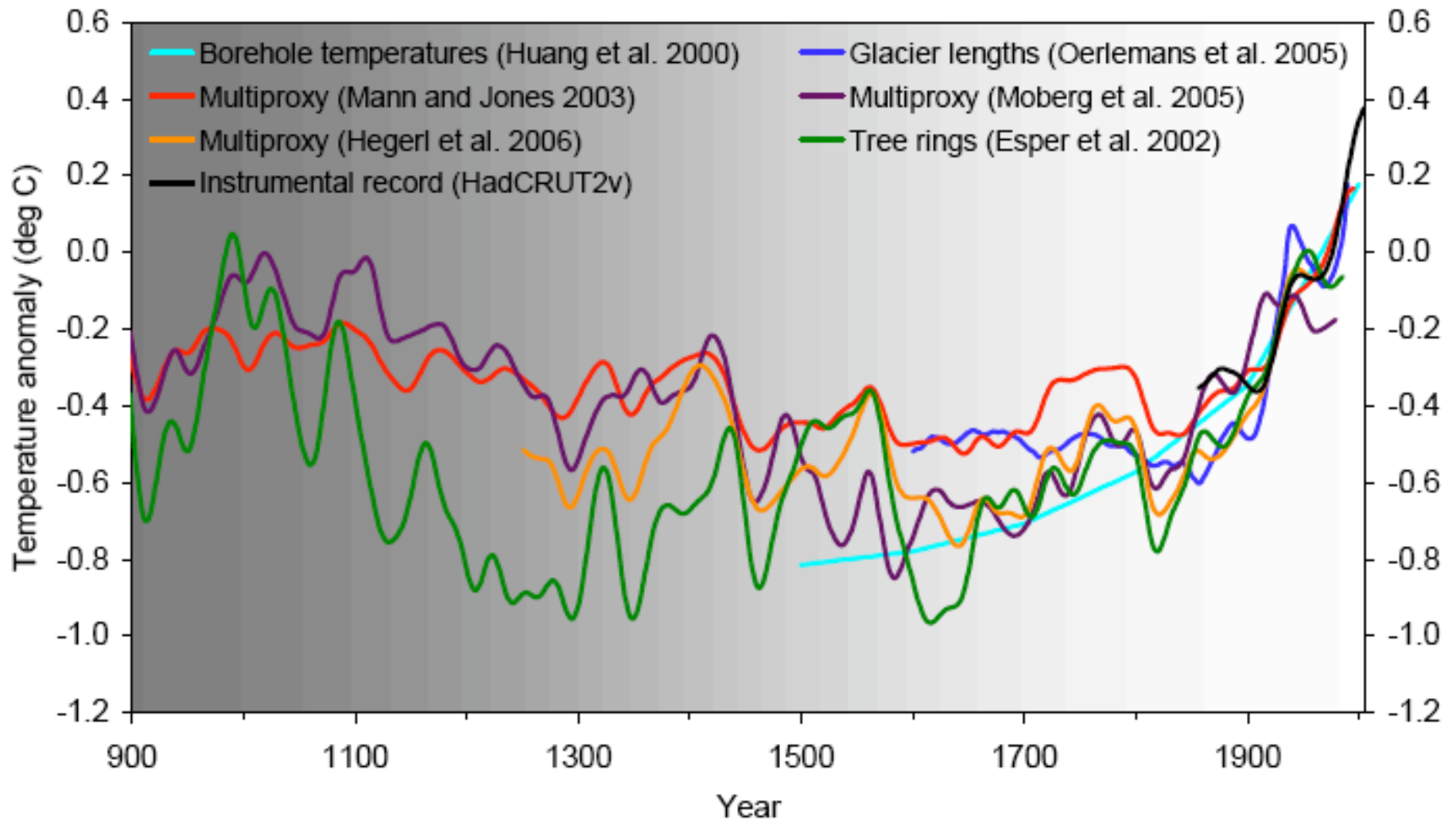
Ice Cores

- Chronology (counting back, other)
- Gas Bubbles & Volcanic Ash
- Temps from ^{18}O Isotopes
- Temperatures Down the Bore Hole

Lonnie Thompson's group collects ice cores.



NRC 2006 Graphic of STR:



NRC Committee Concludes

- Last 100 yrs: Temps up 0.6 deg C
 - (highly likely ~95% Confident)
- 30 yr averages warmest in 400 yrs
 - (likely ~2 to 1 odds)
- 30 yr averages warmest 1000 yrs
 - (Plausible ~reasonable, not possible to bring a convincing argument against - no numbers)

<http://epw.senate.gov/pressitem.cfm?id=257697&party=rep>

"Today's NAS report reaffirms what I have been saying all along, that Mann's 'hockey stick' is broken," **Senator Inhofe said**. "Today's report refutes Mann's prior assertions that there was no Medieval Warm Period or Little Ice Age."



My chance brushes with **Ben**,
Ta-Hsin, Peter, and Manny,
(Ed W., Too) have enriched
my life and I think we've
done some good.

At least we had fun tryin'.

Introducing a Fractional INAR(1) Model for Time Series of Counts

Harry Pavlopoulos

www.stat-athens.aueb.gr/~hgp/

Department of Statistics
Athens University of Economics and Business
76 Patission Str., GR-10434 Athens, Greece



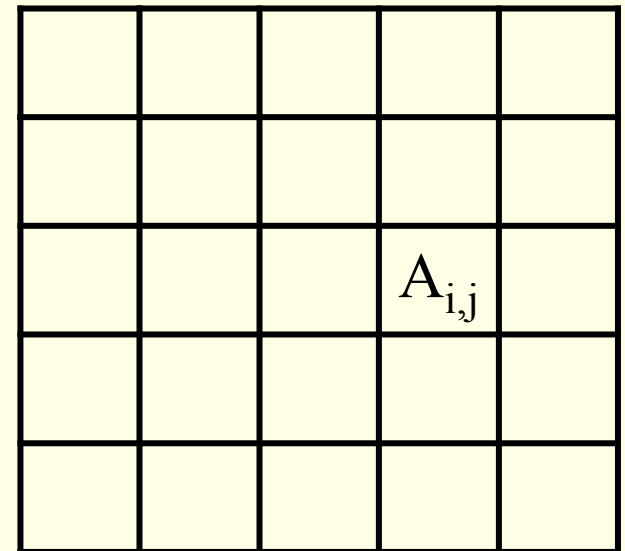
OUTLINE

1. Motivation by an interest on series of pixel counts and relevant links to the “Threshold Method” for prediction of SARR.
2. Introduce Randomized Binomial Thinning and a class of Fractional INAR(1) models.
3. Calculation of Moments up to 2nd order, under the assumption of stationarity.
4. Simulation and Inference.
5. Application on daily series of global rain-rate fields.

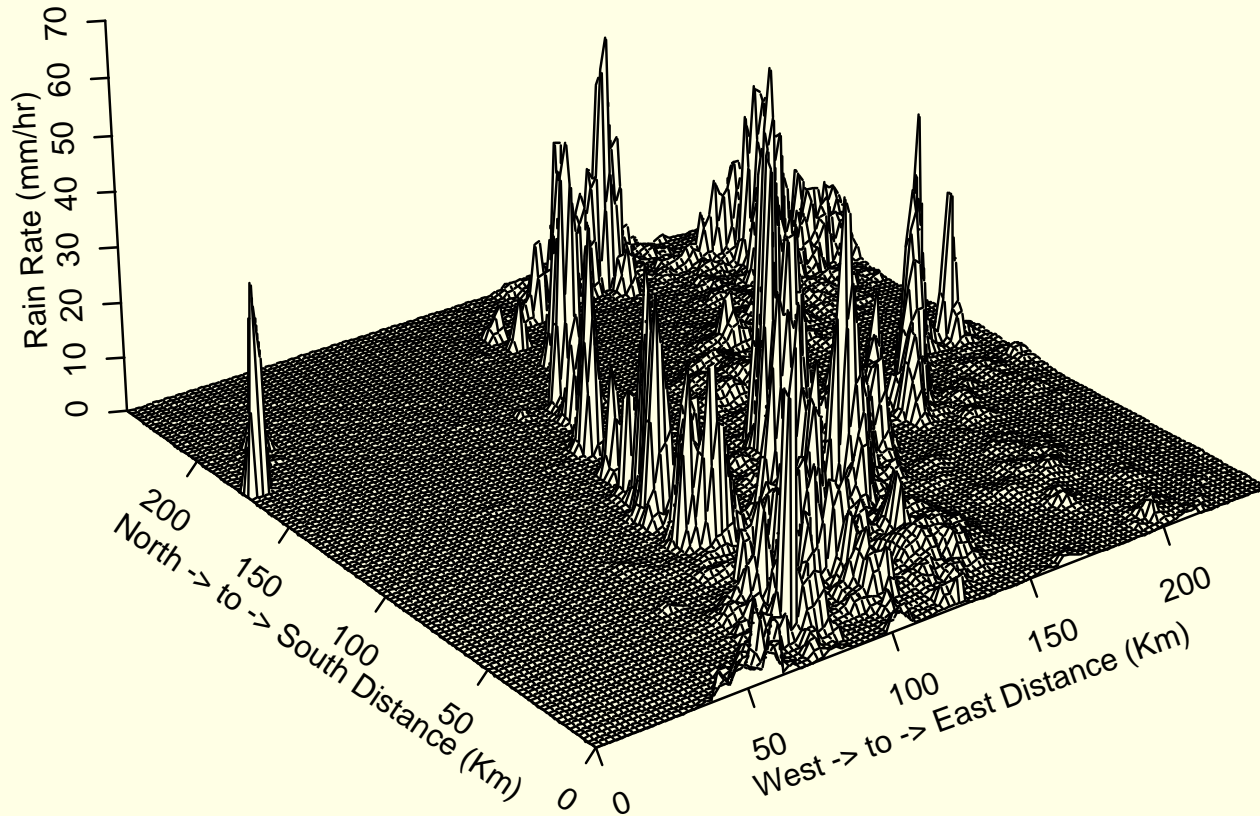
MOTIVATION

Let $\left\{ R_t(A_{i,j}) \ ; \ i, j = 1, 2, \dots, n(A) \right\}_{t=1}^T$ be a time series of instantaneous realizations (measurements) of random “*marks*” over the $n^2(A)$ pixels $\{A_{i,j}\}$ of a 2D *Marked Regular Lattice* (**MRL**) configuration of a random field probed over a fixed domain A .

WLOG: assume *pixels of square shape* (with fixed side length), and *non-negative marks* (e.g. environmental / geophysical MRL).

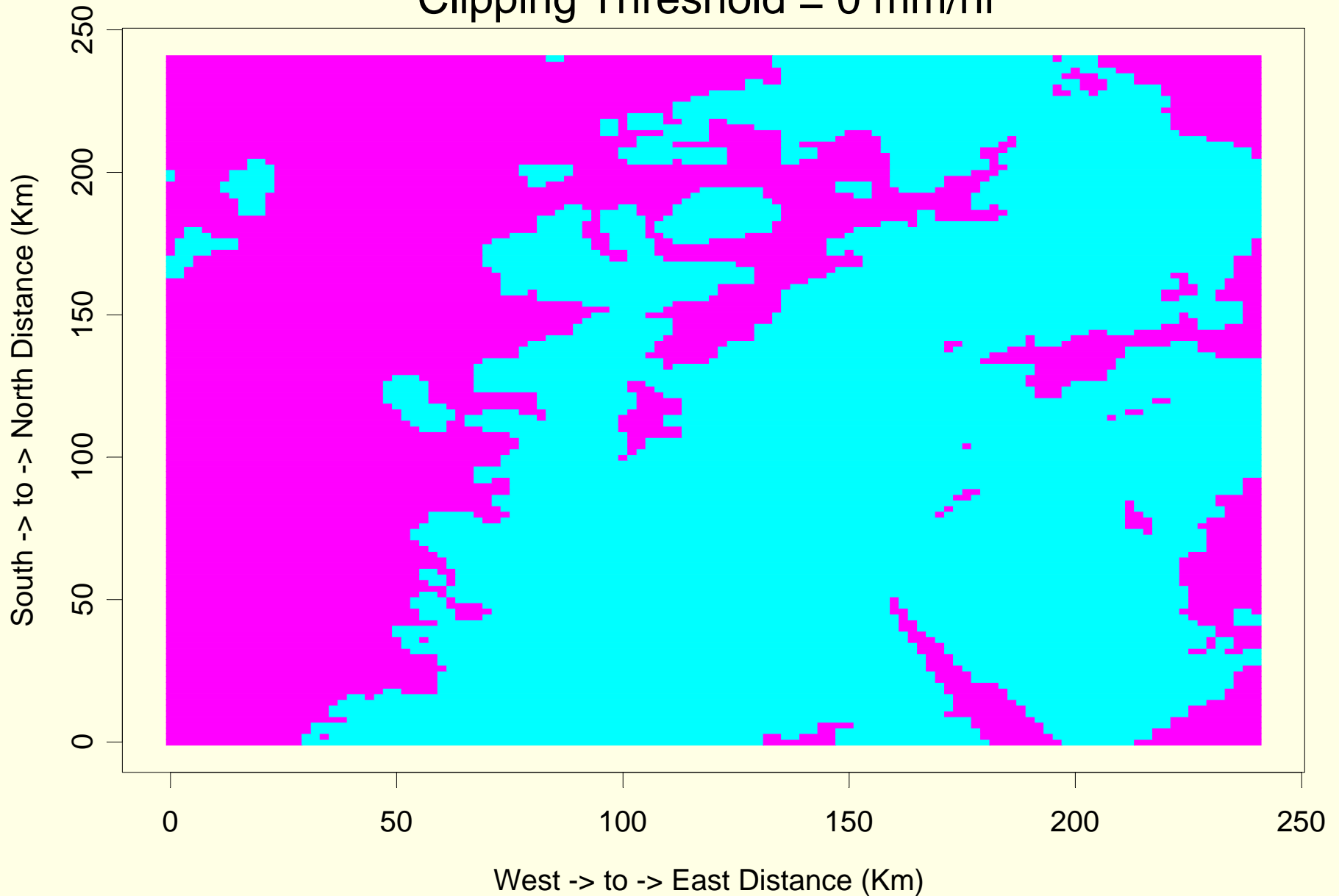


240x240 Km² TOGA-COARE Region, MIT Radar

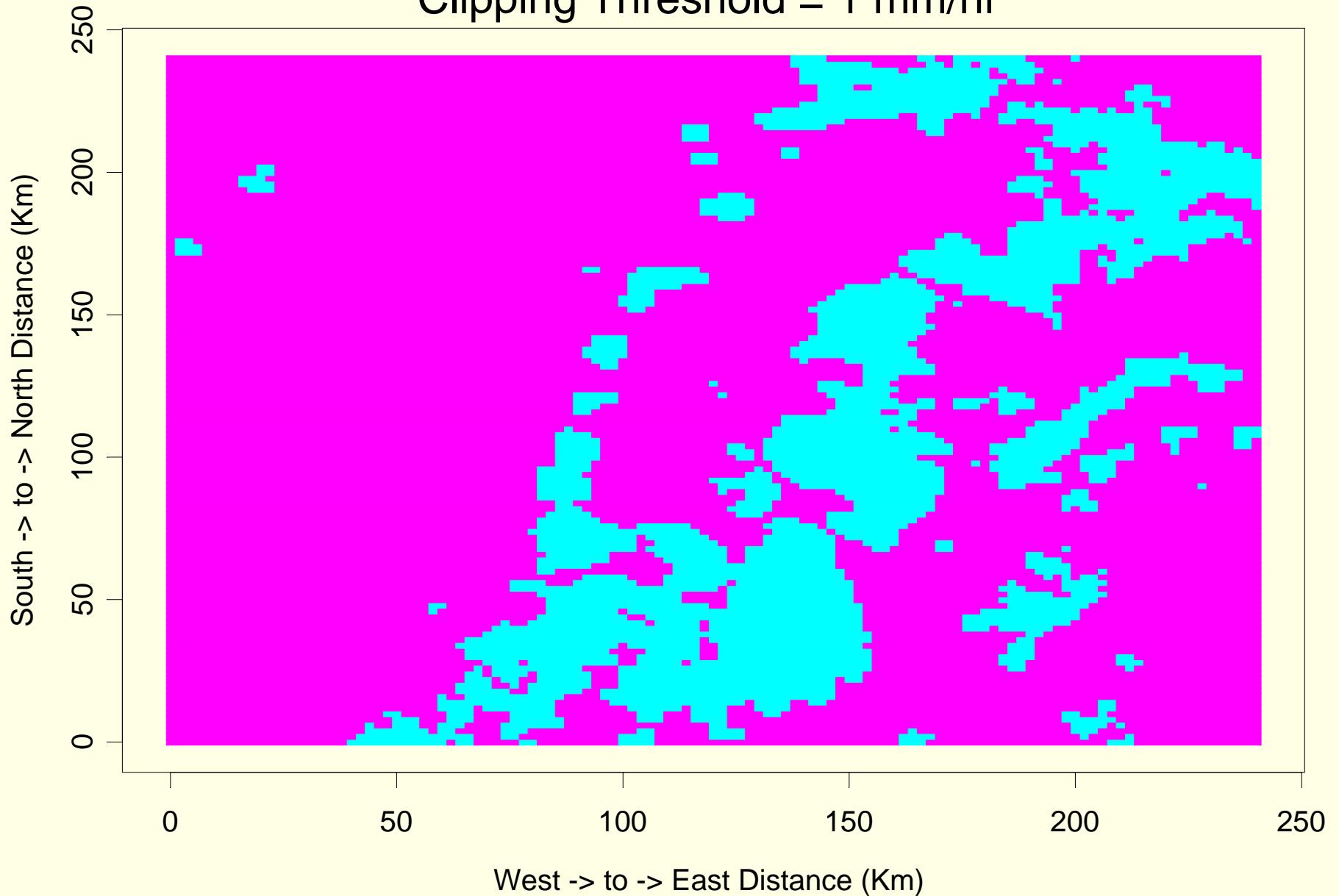


Cruise-1, Date 921110, Time 23:21 UTC

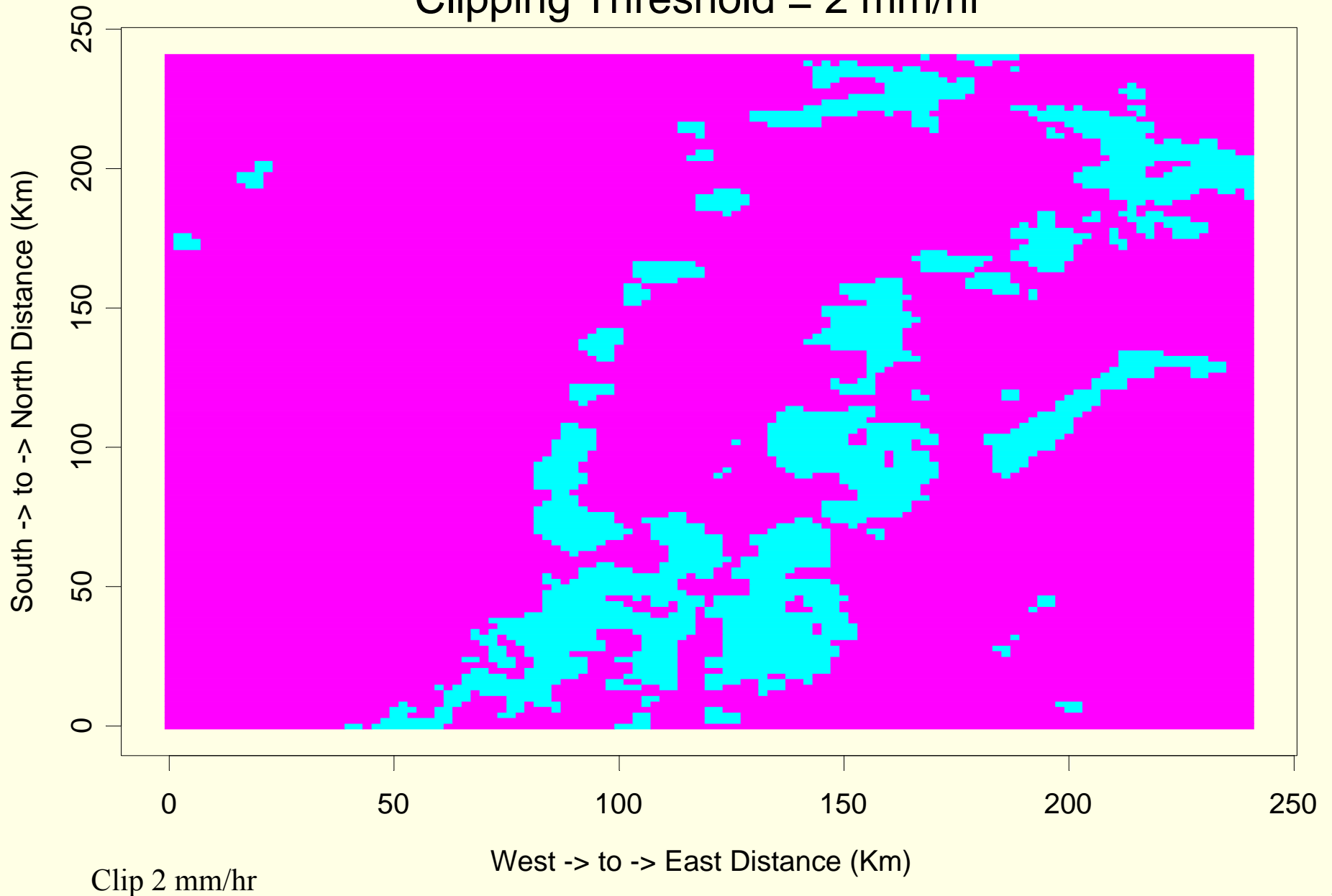
MIT Radar, Cruise-1, Date 921110, Time 23:21 UTC
Clipping Threshold = 0 mm/hr



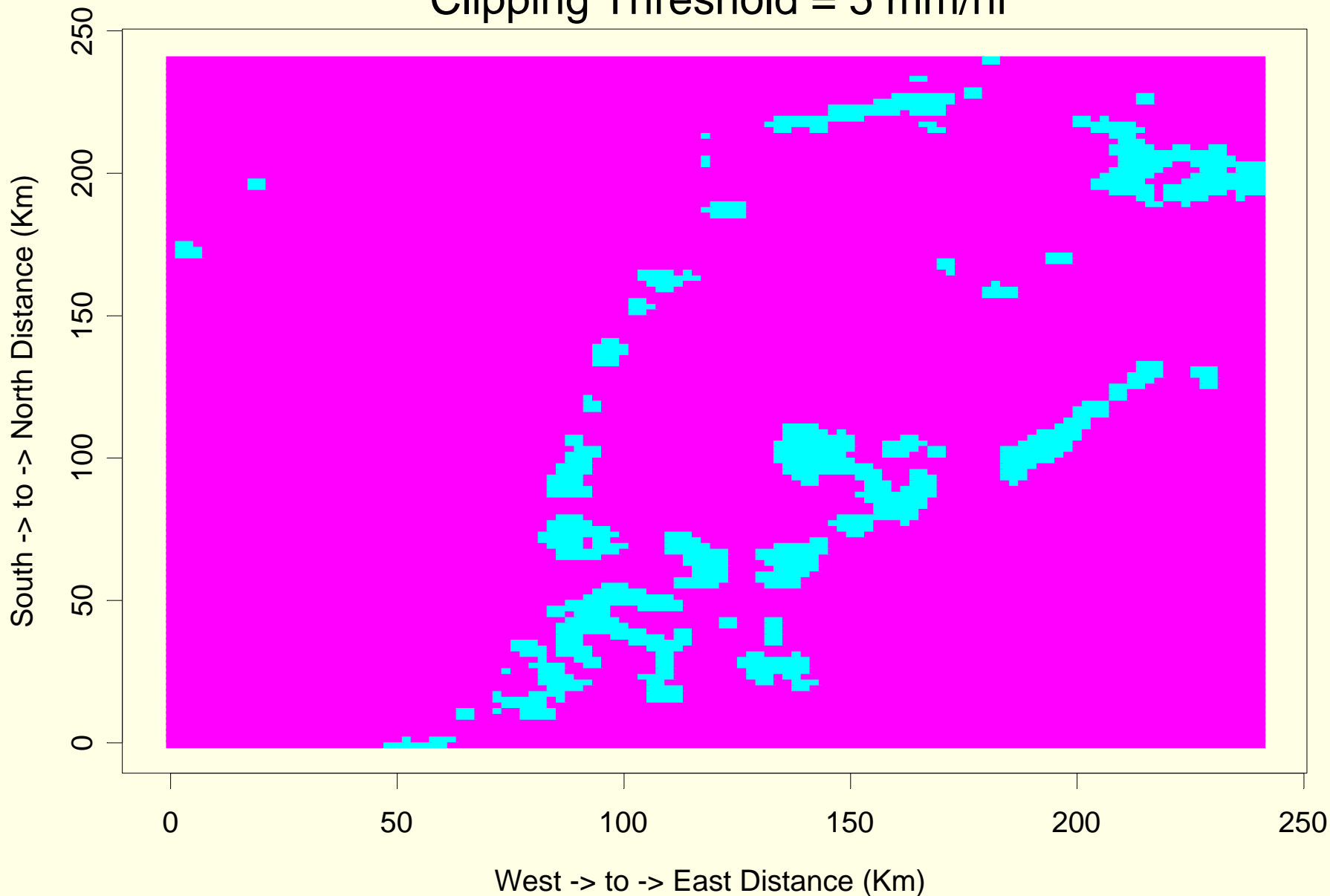
MIT Radar, Cruise-1, Date 921110, Time 23:21 UTC Clipping Threshold = 1 mm/hr



MIT Radar, Cruise-1, Date 921110, Time 23:21 UTC
Clipping Threshold = 2 mm/hr



MIT Radar, Cruise-1, Date 921110, Time 23:21 UTC
Clipping Threshold = 5 mm/hr



Clip 5 mm/hr

u-COVER: $\left\{ \bar{F}_{n,t}(u, A) = n^{-2}(A) \cdot \sum_{i,j=1}^{n(A)} I[R_t(A_{i,j}) > u] \right\}_{u \geq 0}$

is a random functional, parameterized by the *u-threshold*, providing an important statistical summary of the MRL over A at each instant t .

$\left\{ \bar{F}_{n,t}(u, A) \right\}_{u \geq 0}$ is a ***functional of random tail-probabilities***

corresponding to the ***instantaneous empirical spatial cumulative***

distribution functional (IESCDF): $\left\{ F_{n,t}(u, A) = 1 - \bar{F}_{n,t}(u, A) \right\}_{u \geq 0}$

Time series of pixel counts where a MRL exceeds a fixed u -threshold

$$\left\{ X_t(u, A) := \sum_{i,j=1}^{n(A)} I[R_t(A_{i,j}) > u] \right\}_{t=1,2,\dots,T}$$

motivate an interest to model (over-dispersed and persistent) time series of counts.

- IESCDF is a natural predictor of *instantaneous spatial-cdf* (ISCDF)

$$\left\{ F_{\infty,t}(u, A) = 1 - \bar{F}_{\infty,t}(u, A) = |A|^{-1} \cdot \int_A I[R_t(a) \leq u] da \right\}_{u \geq 0}$$

of the probed underlying spatio-temporal random field $\{R_t(a); a \in A\}$ when $n(A)$ is large enough and/or $|A_{i,j}|/|A|$ is small enough (Increasing Domain / Infill Asymptotics).

Lahiri S.N., Kaiser M.S., Cressie N., Hsu N.J.

(JASA 1999, 94: 86-97)

- Analogously, *u-COVER* is a natural predictor of *ISCDF-tails*.
- Under appropriate assumptions of spatial-temporal homogeneity, rendering stationary ISCDF, modelling the **numerator of u-COVER** by a stationary model suitable for time series of counts, could (in principle) facilitate temporal prediction of spatial tail-probabilities.

- Prediction of spatial sample moments (i.e. moments of IESCDF) may also be facilitated by implementation of the “*Threshold Method*”, provided that the underlying random field has sufficient variability in its intermittency between zero and positive marks.

$$R_t^{(k)}(A_{i,j}) := \frac{1}{|A_{i,j}|} \cdot \int_{A_{i,j}} R_t^k(a) da = \int_{-\infty}^{+\infty} u^k dF_{\infty,t}(u; A_{i,j})$$

$$R_t^{(k)}(A) := \frac{1}{n^2(A)} \cdot \sum_{i,j=1}^{n(A)} R_t^{(k)}(A_{i,j}) = \int_{-\infty}^{+\infty} u^k dF_{\infty,t}(u; A)$$

$$R_t^{[k]}(A) := \frac{1}{n^2(A)} \cdot \sum_{i,j=1}^{n(A)} R_t^k(A_{i,j}) = \int_{-\infty}^{+\infty} u^k dF_{n,t}(u; A)$$

THRESHOLD METHOD

$$R_t^{[k]}(A) = \beta_k(u) \cdot \bar{F}_{n,t}(u, A) + \varepsilon_t$$

$$\beta_k(u) = \frac{E\left\{ R_t^{(k)}(A_{i,j}) \mid R_t^{(k)}(A_{i,j}) > 0 \right\}}{P\left\{ R_t^{(k)}(A_{i,j}) > u \mid R_t^{(k)}(A_{i,j}) > 0 \right\}}$$

within an “*optimal*” range of u -thresholds, facilitating/justifying prediction of spatial moments by linear regression on ECDF-tails.

Kedem & Pavlopoulos [1991, JASA]

Short, Shimizu, Kedem [1993, JAMet.]

Shimizu, Short, Kedem [1993, J.Met.Soc.Japan]

Shimizu & Kayano [1994, J.Met.Soc.Japan]

Mase [1996, Ann.Inst.Statistical.Math.]

Sakaguchi & Mase [2003, J.Japan Statist.Soc.]

INAR(1) Model: McKenzie (1985) ; Al-Osh & Alzaid (1987)

$$X_t = p \circ X_{t-1} + \varepsilon_t$$

$\{\varepsilon_t\}$ *I.I.D. innovations* ; $\varepsilon_t \perp \{X_s\}_{s < t}$

Binomial p -Thinning: $p \circ X_{t-1} := \sum_{i=1}^{X_{t-1}} Z_{t,i}$

$\{Z_{t,i}\}$ IID array, independently of $\{X_t\}$

$$P(Z_{t,i} = 1) = 1 - P(Z_{t,i} = 0) = p \in [0,1]$$

$$(p \circ X_{t-1} \parallel X_{t-1}) \sim B(X_{t-1}, p)$$

$$\varphi_{p \circ X}(u) = E\left(e^{i \cdot u \cdot (p \circ X)}\right) = E\left\{\left(1 - p + p \cdot e^{i \cdot u}\right)^X\right\}$$

Stationary Solution

$$\{X_t\} \text{ (strictly) stationary} \Leftrightarrow p < 1 \quad \& \quad E(\varepsilon_t) < \infty$$

$$X_t \stackrel{D}{=} \sum_{j=0}^{\infty} \alpha^j \circ \varepsilon_{t-j}$$

Discrete
Self-Decomposable
Law

$$G_X(z) = E(z^{X_t}) = \prod_{j=0}^{\infty} G_{\varepsilon}(1 - \alpha^j + \alpha^j \cdot z)$$

$$G_{\varepsilon}(z) = E(z^{\varepsilon_t}), \quad |z| \leq 1$$

$$\sum_{j=1}^{\infty} \frac{P(\varepsilon_t \geq j)}{j} < \infty$$

2nd Order Moments

$$\mu_X = E(X_t) = \frac{\mu_\varepsilon}{1-p} \quad \& \quad \sigma_X^2 = \text{Var}(X_t) = \frac{p \cdot \mu_\varepsilon + \sigma_\varepsilon^2}{1-p^2}$$

$$ID(X) = \frac{\sigma_X^2}{\mu_X} = \left(1 + \frac{ID(\varepsilon)}{p}\right) \bigg/ \left(1 + \frac{1}{p}\right) \geq 1 \Leftrightarrow ID(\varepsilon) \geq 1$$

$$\gamma_X(k) = \text{Cov}(X_t, X_{t-k}) = p^{|k|} \cdot \sigma_X^2$$

$$\rho_X(k) = \text{Corr}(X_t, X_{t-k}) = p^{|k|} \geq 0$$

$$S_X(\omega) = \frac{1}{\pi} \cdot \sum_{k=-\infty}^{\infty} \gamma_X(k) \cdot e^{-ik\omega} = \frac{p \cdot \mu_\varepsilon + \sigma_\varepsilon^2}{\pi \cdot (1 - 2p \cdot \cos \omega + p^2)}$$

Model Marginal or Innovation Law ?

$$G_{\varepsilon}(z) = G_X(z) / G_X(1 - \alpha + \alpha \cdot z)$$

$$X \sim P(\lambda) \Rightarrow \varepsilon \sim P((1 - \alpha) \cdot \lambda)$$

$$X \sim \text{Geom}(\theta) \Rightarrow \varepsilon \sim B(1, 1 - \alpha) \cdot \text{Geom}(\theta)$$

$$X \sim NB(\beta, \lambda) \Rightarrow \varepsilon \sim \text{Nontrivial}$$

Model ε -law so that

$$ID(\varepsilon) \geq 1 \quad \& \quad \sum_{j=1}^{\infty} \frac{P(\varepsilon \geq j)}{j} < \infty$$

subject to feasibility of inference and simulation.

Mixed Poisson Innovations

Pavlopoulos & Karlis (*Environmetrics* 2008; **19**: 369-393)

$$P(\varepsilon_t = x) = \sum_{i=1}^m \pi_i \cdot e^{-\lambda_i} \cdot \frac{\lambda_i^x}{x!}, \quad m \geq 1, \quad x = 0, 1, 2, \dots$$

$$\lambda_1, \dots, \lambda_m > 0, \quad 0 \leq \pi_1, \dots, \pi_m \leq 1, \quad \sum_{i=1}^m \pi_i = 1$$

$$G_\varepsilon(z) = \sum_{i=1}^m \pi_i \cdot \exp\{\lambda_i \cdot (z - 1)\}$$

$$\mu_\varepsilon = \sum_{i=1}^m \pi_i \lambda_i \quad ; \quad \sigma_\varepsilon^2 = \mu_\varepsilon + \sum_{i=1}^m \pi_i (\lambda_i - \mu_\varepsilon)^2 > \mu_\varepsilon$$

$$\sum_{j=1}^{\infty} \frac{P(\varepsilon_t \geq j)}{j} \leq \sum_{j=1}^{\infty} \left(\sum_{i=1}^m \pi_i \cdot \frac{\lambda_i^j}{j!} \right) = \sum_{i=1}^m \pi_i \cdot (e^{\lambda_i} - 1) < \infty$$

$$G_X(z) = \prod_{j=0}^{\infty} \left(\sum_{i=1}^m \pi_i \cdot \exp\{\lambda_i \cdot \alpha^j \cdot (z - 1)\} \right)$$

A Convenient Representation

$$\varepsilon_t = \sum_{i=1}^m Q_t^{(i)} \cdot \Lambda_t^{(i)}$$

$$\mathbf{Q}_t = (Q_t^{(1)}, \dots, Q_t^{(m)}) \sim M(1, p_1, \dots, p_m)$$

$$\Lambda_t = (\Lambda_t^{(1)}, \dots, \Lambda_t^{(m)}) \sim_{\perp} (P(\theta_1), \dots, P(\theta_m))$$

$\{\mathbf{Q}_t\}_{t \in \mathbb{Z}} \perp \{\Lambda_t\}_{t \in \mathbb{Z}}$ independent sequences
of I.I.D. random vectors

$$\sum_{i=1}^m Q_t^{(i)} = 1 \quad \Rightarrow$$

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t = \sum_{i=1}^m Q_t^{(i)} \cdot (\alpha \circ X_{t-1} + \Lambda_t^{(i)})$$

Poisson-Binomial Convolutions

$$\left(\alpha \circ X_{t-1} + \Lambda_t^{(i)} \mid X_{t-1}\right) = \left(\alpha \circ X_{t-1} \mid X_{t-1}\right) + \Lambda_t^{(i)} \sim PB(X_{t-1}, \alpha; \lambda_i)$$

$$\pi_i(z \mid x_{t-1}) = \sum_{k=0}^{z \wedge x_{t-1}} e^{-\lambda_i} \cdot \frac{\lambda_i^k}{k!} \cdot \binom{x_{t-1}}{z-k} \cdot \alpha^{z-k} \cdot (1-\alpha)^{x_{t-1}-z+k}$$

$$\left(X_t \mid X_{t-1}\right) \sim \sum_{i=1}^m Q_t^{(i)} \cdot PB(X_{t-1}, \alpha; \lambda_i)$$

$(X_t \mid X_{t-1}) \sim$ mixture of m **Poisson-Binomial** r. v.

$$P_\theta(X_t = x_t \mid X_{t-1} = x_{t-1}) = \sum_{i=1}^m p_i \cdot \pi_i(x_t \mid x_{t-1})$$

More generally, the *k -step predictive/conditional law* of $(X_t \mid X_{t-k})$ is a mixture of m^k **Poisson-Binomial** r. v. with parameters determined by $\alpha, \lambda_1, \dots, \lambda_m, p_1, \dots, p_m$.

TOGA-COARE Data

Cruise-1: 10 November - 9 December, 1992)

MIT-Radar (Doppler Precipitation)

Location: (2°S, 156°E) China Sea, SW Pacific Ocean

Temporal resolution (sampling frequency) 20 *min*.

Regular section of 184 scans: 22/11(02:01) – 24/11(15:21)

Total duration of section: 61 *hr* ; 20 *min*

Pixel SARR Marks: $R=(Z/230)^{0.8}$

Spatial Scales of probed subregions (centrally nested):

240, 120, 60, 32, 16, 8, 4, 2 *Km*

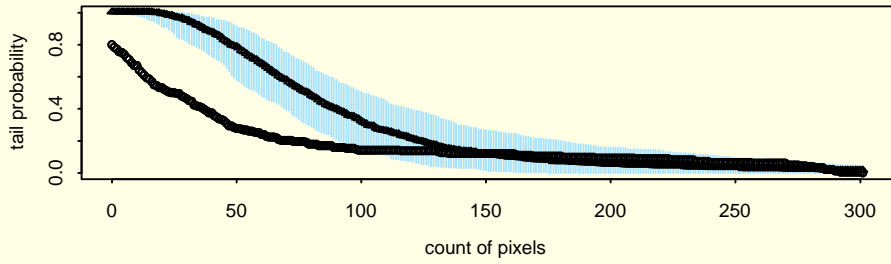
Clipping threshold levels of pixel marks:

0, 1, 2, 5, 10, 20 *mm/hr*

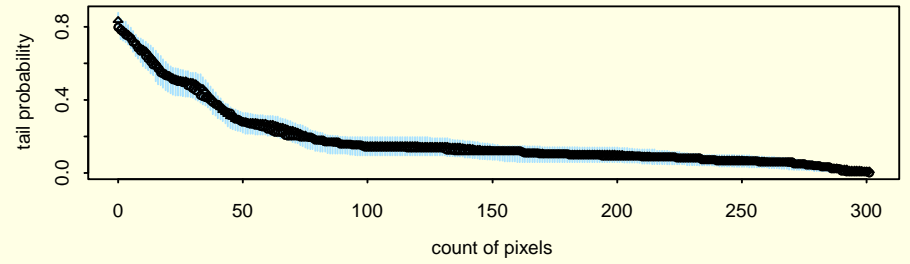
Total of 8x6 = 48 time series

| threshold | mean | ID | IS | IK | $P(X = 0)$ | acf | mean | ID | IS | IK | $P(X = 0)$ | acf | |
|-----------|---------|--------|-----------------------|---------|------------|--------|--------------|---------------|-----------------------|--------------|--------------|--------------|--|
| | | | <u>Scale = 240 Km</u> | | | | | | <u>Scale = 120 Km</u> | | | | |
| 0 mm/hr | 3903.67 | 974.84 | 0.232 | 1.905 | 0.000 | 0.823 | 1266.17 | 324.62 | 0.217 | 1.992 | 0.000 | 0.807 | |
| 1 mm/hr | 810.36 | 706.90 | 1.414 | 4.333 | 0.000 | 0.810 | 149.40 | 255.80 | 1.862 | 5.620 | 0.060 | 0.847 | |
| 2 mm/hr | 543.41 | 547.05 | 1.526 | 4.602 | 0.000 | 0.791 | 102.51 | 182.10 | 1.769 | 5.114 | 0.103 | 0.844 | |
| 5 mm/hr | 253.28 | 246.55 | 1.351 | 3.763 | 0.005 | 0.756 | 53.72 | 112.59 | 1.973 | 5.856 | 0.201 | 0.823 | |
| 10 mm/hr | 114.16 | 129.28 | 1.689 | 5.029 | 0.011 | 0.758 | 28.03 | 72.04 | 2.183 | 6.718 | 0.255 | 0.819 | |
| 20 mm/hr | 37.85 | 59.73 | 1.941 | 6.021 | 0.092 | 0.745 | 10.12 | 35.38 | 2.797 | 10.838 | 0.413 | 0.782 | |
| | | | <u>Scale = 60 Km</u> | | | | | | <u>Scale = 32 Km</u> | | | | |
| 0 mm/hr | 383.95 | 93.40 | 0.477 | 2.167 | 0.000 | 0.823 | 119.13 | 25.94 | -0.053 | 1.989 | 0.000 | 0.824 | |
| 1 mm/hr | 25.75 | 63.24 | 2.064 | 6.871 | 0.391 | 0.760 | 3.40 | 17.36 | 3.387 | 17.068 | 0.685 | 0.686 | |
| 2 mm/hr | 17.51 | 47.96 | 2.351 | 8.688 | 0.457 | 0.727 | 2.33 | 15.95 | 3.975 | 21.904 | 0.734 | 0.670 | |
| 5 mm/hr | 9.74 | 39.16 | 3.110 | 13.583 | 0.527 | 0.687 | 1.27 | 11.59 | 4.285 | 24.800 | 0.837 | 0.632 | |
| 10 mm/hr | 5.69 | 29.36 | 3.347 | 15.536 | 0.625 | 0.660 | 0.71 | 8.93 | 5.167 | 35.373 | 0.853 | 0.589 | |
| 20 mm/hr | 2.35 | 17.63 | 3.822 | 18.527 | 0.739 | 0.583 | 0.29 | 5.67 | 5.515 | 37.086 | 0.924 | 0.662 | |
| | | | <u>Scale = 16 Km</u> | | | | | | <u>Scale = 8 Km</u> | | | | |
| 0 mm/hr | 39.207 | 10.019 | -0.316 | 1.812 | 0.033 | 0.762 | 10.277 | 3.511 | -0.558 | 1.727 | 0.109 | 0.694 | |
| 1 mm/hr | 0.946 | 9.855 | 4.507 | 28.257 | 0.864 | 0.536 | 0.185 | 3.540 | 5.448 | 36.841 | 0.929 | 0.231 | |
| 2 mm/hr | 0.658 | 9.385 | 5.489 | 39.578 | 0.891 | 0.521 | 0.098 | 2.918 | 6.930 | 55.948 | 0.951 | 0.272 | |
| 5 mm/hr | 0.359 | 7.439 | 6.912 | 62.410 | 0.918 | 0.189 | 0.043 | 3.224 | 9.179 | 89.142 | 0.984 | -0.014 | |
| 10 mm/hr | 0.196 | 7.344 | 9.201 | 99.346 | 0.940 | 0.186 | 0.022 | 2.492 | 11.790 | 147.417 | 0.989 | -0.009 | |
| 20 mm/hr | 0.087 | 5.568 | 9.367 | 98.008 | 0.978 | 0.345 | 0.005 | 1.000 | 13.344 | 180.032 | 0.995 | -0.005 | |
| | | | <u>Scale = 4 Km</u> | | | | | | <u>Scale = 2 Km</u> | | | | |
| 0 mm/hr | 2.690 | 1.100 | -0.728 | 1.706 | 0.245 | 0.620 | 0.696 | 0.306 | -0.843 | 1.704 | 0.304 | 0.561 | |
| 1 mm/hr | 0.033 | 1.643 | 7.465 | 60.050 | 0.978 | -0.020 | 0 | - | - | - | 1 | - | |
| 2 mm/hr | 0.011 | 2.000 | 13.344 | 180.032 | 0.995 | -0.005 | 0 | - | - | - | 1 | - | |
| 5 mm/hr | 0.011 | 2.000 | 13.344 | 180.032 | 0.995 | -0.005 | 0 | - | - | - | 1 | - | |
| 10 mm/hr | 0.005 | 1.000 | 13.344 | 180.032 | 0.995 | -0.005 | 0 | - | - | - | 1 | - | |
| 20 mm/hr | 0 | - | - | - | 1 | - | 0 | - | - | - | 1 | - | |

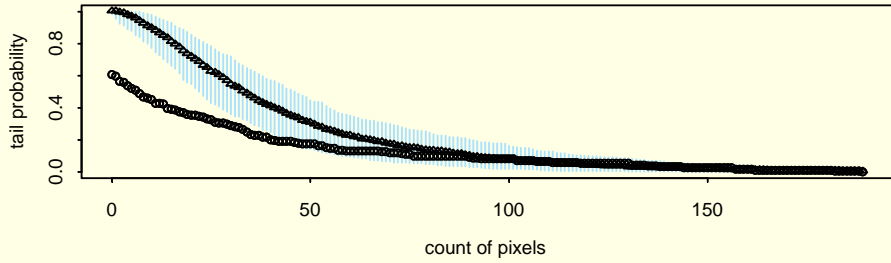
scale = 120 km, threshold=5mm/hr



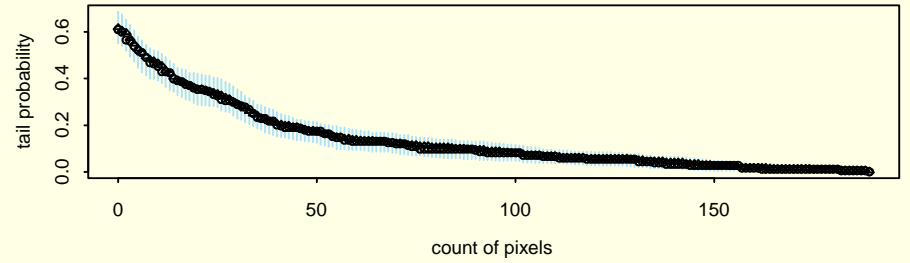
scale = 120 km, threshold=5mm/hr



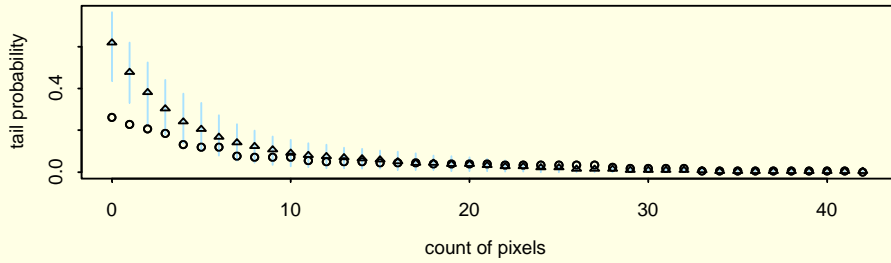
scale = 60 km, threshold=1mm/hr



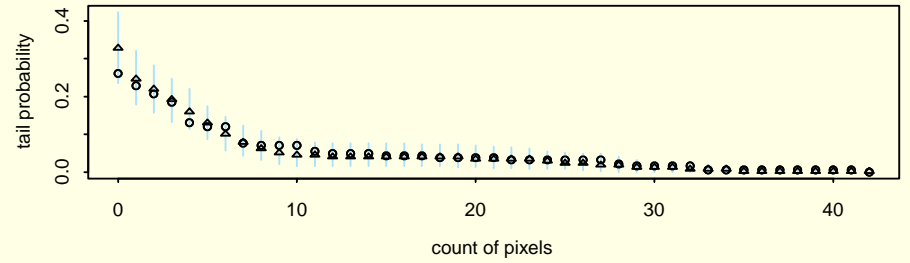
scale = 60 km, threshold=1mm/hr



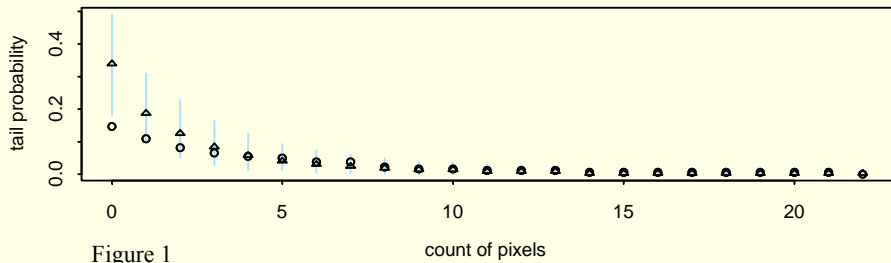
scale = 60 km, threshold=20mm/hr



scale = 60 km, threshold=20mm/hr



scale = 32 km, threshold=10mm/hr



scale = 32 km, threshold=10mm/hr

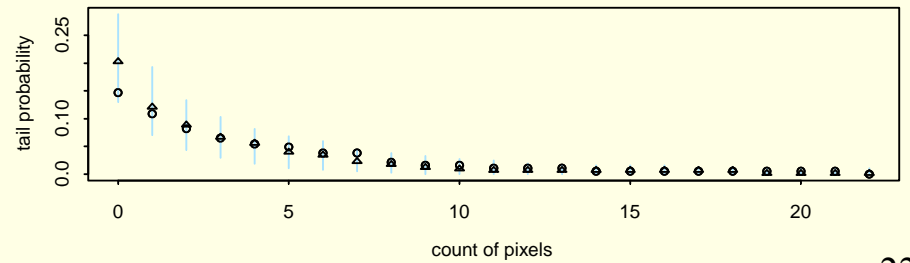
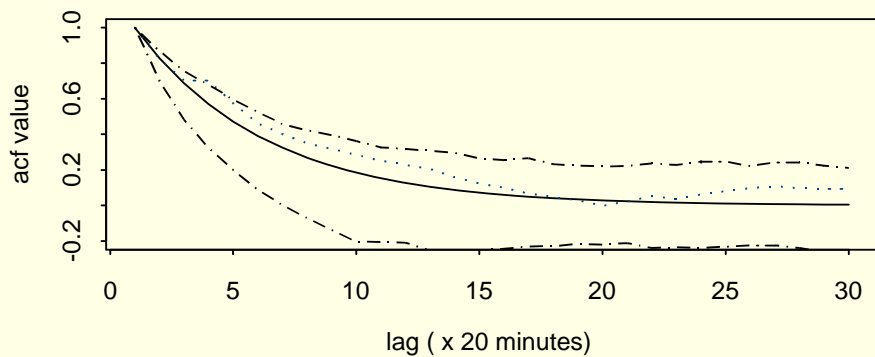
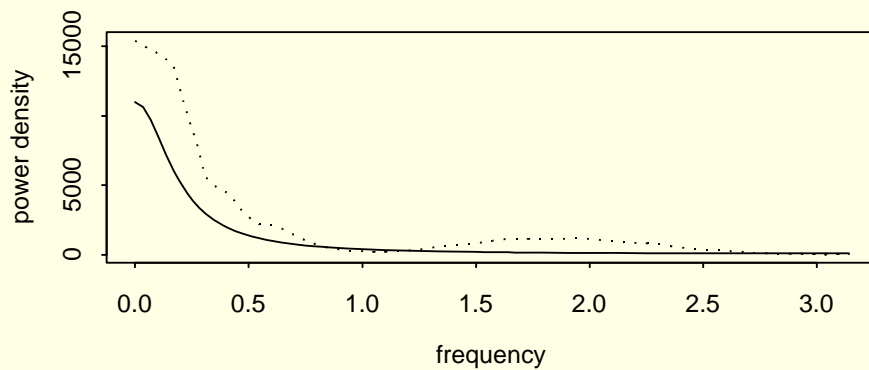


Figure 1
(Tails: MOM vs CML)

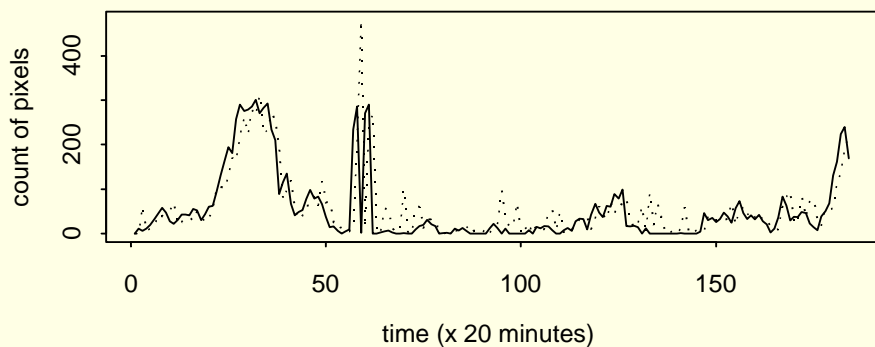
(a) ACF



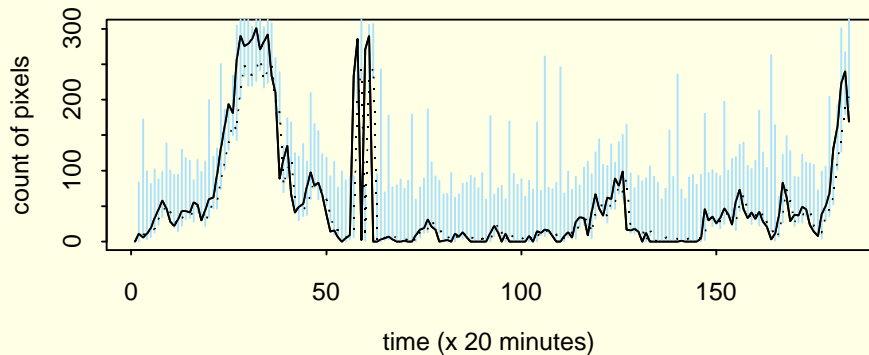
(b) Spectrum



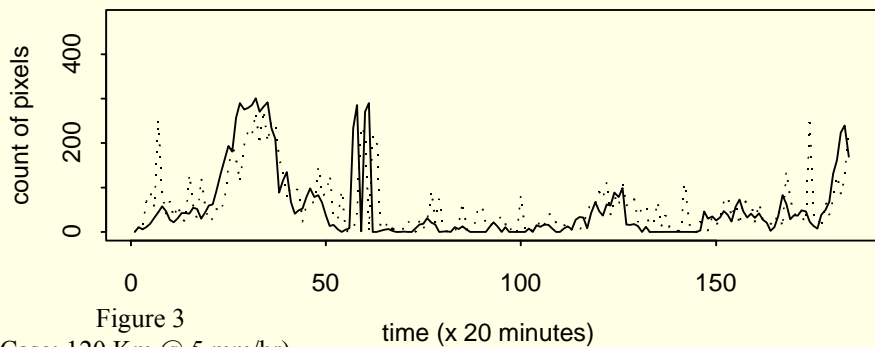
(c) 1-step predictions from one replication



(d) 1-step percentile predictions from 1000 reps



(e) 2-step predictions from one replication



(f) 2-step percentile predictions from 1000 reps

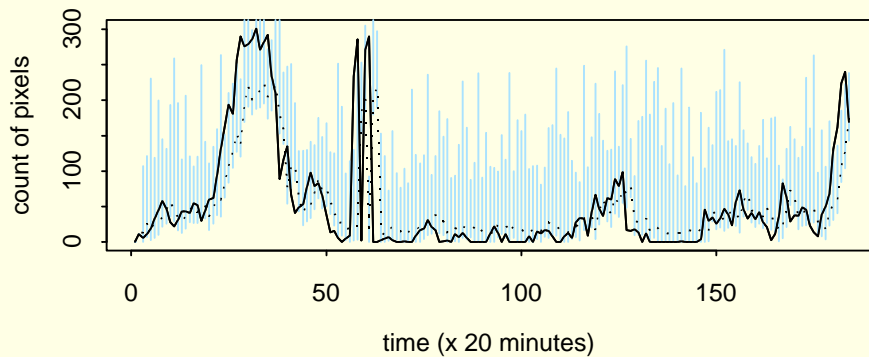
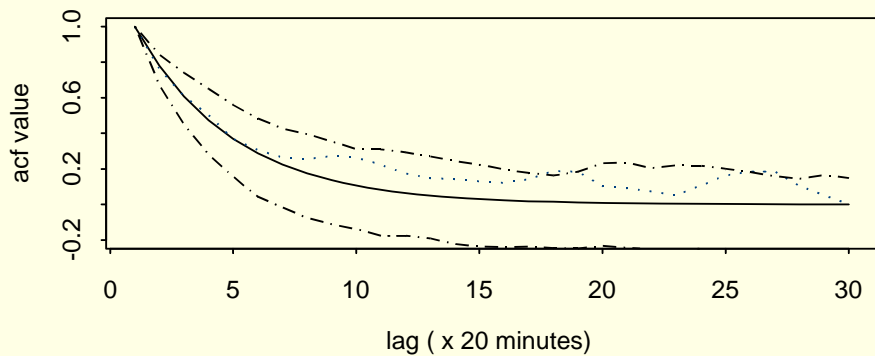
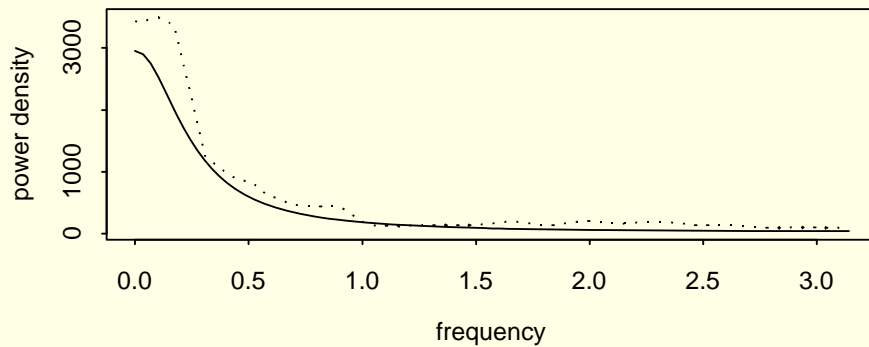


Figure 3
(Case: 120 Km @ 5 mm/hr)

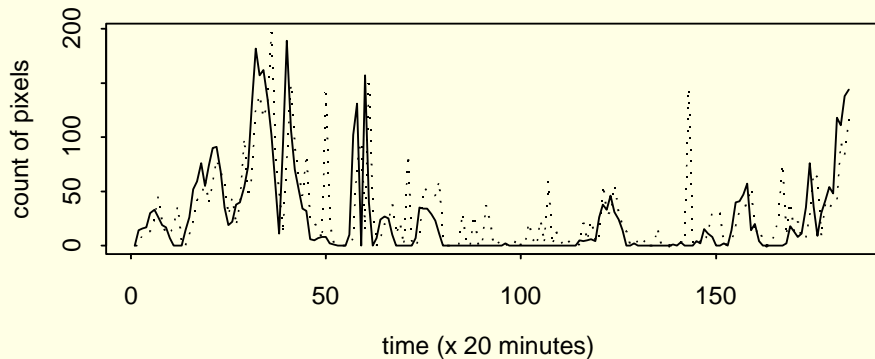
(a) ACF



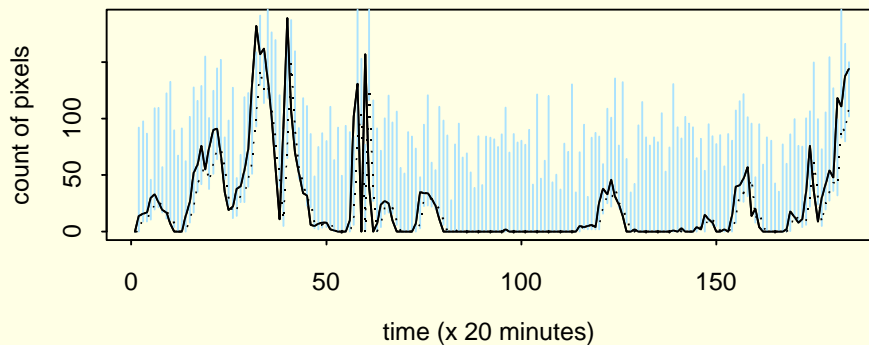
(b) Spectrum



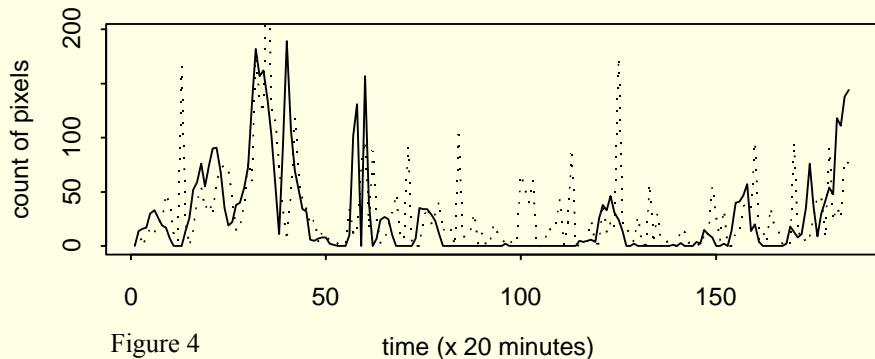
(c) 1-step predictions from one replication



(d) 1-step percentile predictions from 1000 reps



(e) 2-step predictions from one replication



(f) 2-step percentile predictions from 1000 reps

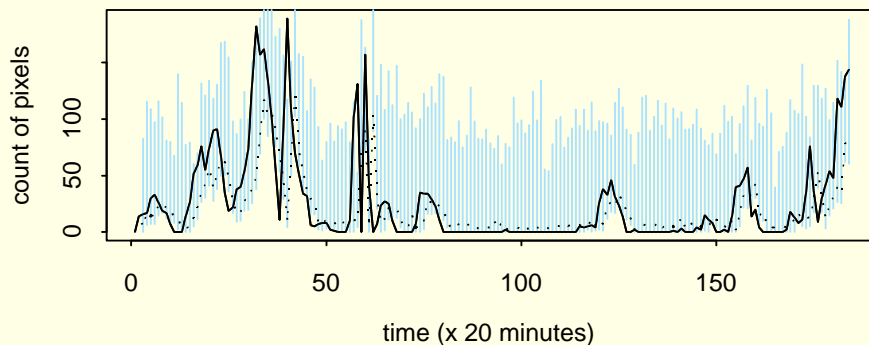


Figure 4
(Case: 60 Km @ 1 mm/hr)

Randomized Binomial Thinning

X is a non-negative integer-valued random variable,

P is a $[0,1]$ -valued random variable.

(X, P) may be jointly distributed or stochastically independent.

$$P \circ X := \sum_{i=1}^X Z_P(i)$$

$\{Z_P(i)\}_{i \geq 1}$ array of binary r.v.'s, which conditionally on $\{P = p\}$ are I.I.D., following $Bin(1,p)$, independently of X .

$$P\{Z_P(i) = 1 \parallel P = p\} = 1 - P\{Z_P(i) = 0 \parallel P = p\} = p$$

$$(P \circ X \parallel P = p) \stackrel{D}{=} p \circ X \quad ; \quad i.e. \text{ RBT} \supset \text{ BT}$$

$$(P \circ X \parallel P = p ; X = x) \stackrel{D}{=} B(x, p)$$

$$E\left(e^{i \cdot u \cdot (P \circ X)}\right) = E\left\{E\left\{e^{i \cdot u \cdot (P \circ X)} \parallel P, X\right\}\right\}$$

$$\varphi_{P \circ X}(u) = E\left\{\left(1 - P + P \cdot e^{i \cdot u}\right)^X\right\}$$

$$P_1 \circ (P_2 \circ X) \stackrel{D}{=} (P_1 P_2) \circ X \quad (\text{always})$$

$$P \circ (X + Y) \stackrel{D}{=} (P \circ X) + (P \circ Y)$$

if X,Y,P are mutually independent OR
if X,Y are independent conditionally on P.

Fractional INAR Model:

$$X_t = P_t \circ X_{t-1} + \varepsilon_t$$

$\{\varepsilon_t\}$ **I.I.D. innovations:** non-negative integer-valued noise

$\{P_t\}$ **Auto-correlated thinning process:** $[0,1]$ -valued

$\{P_t\} \perp \{\varepsilon_t\}$ and $(\forall t) (P_t, \varepsilon_t) \perp \sigma(X_s ; s < t)$

$$P_t \circ X_{t-1} := \sum_{i=1}^{X_{t-1}} Z_t(i)$$

so that, conditionally on P_t , the array $\{Z_t(i) ; i \geq 1\}$ is comprised of I.I.D. $Bin(1, P_t)$ random variables, independently of $\sigma(X_s ; s < t)$.

$$P_t = f(a_t) ; \{a_t\} \sim FGN(\mu, \sigma^2, H)$$

$$f(a) = e^{-a^2} ; e^{-|a|} ; e^{-(a \vee 0)} ; \frac{a^2}{1+a^2} ; etc. \in [0,1]$$

k-step Iteration (future to present):

$$X_{t+k} \stackrel{D}{=} \varepsilon_{t+k} + \sum_{i=1}^{k-1} \left(\prod_{j=0}^{i-1} P_{t+k-j} \right) \circ \varepsilon_{t+k-i} + \left(\prod_{i=1}^k P_{t+i} \right) \circ X_t$$

∞ -step Iteration (present to remote past):

$$X_t \stackrel{D}{=} \varepsilon_t + \sum_{i=1}^{\infty} \left(\prod_{j=0}^{i-1} P_{t-j} \right) \circ \varepsilon_{t-i}$$

Representation of Stationary Solution!

Does a Stationary Solution Exist?

Under What Conditions?

Necessary Conditions !!!

|| || ||
v v v

$$\mu_X = \mu_\varepsilon \left(1 + \sum_{i=1}^{\infty} E(P_1 \dots P_i) \right) = \frac{\mu_\varepsilon}{1 - \mu_P} \Rightarrow \begin{cases} \mu_\varepsilon = E(\varepsilon_t) < \infty \\ \mu_P = E(P_t) < 1 \\ E(P_1 \dots P_t) \rightarrow 0, \quad \text{as } t \rightarrow \infty \end{cases}$$

1-step Iteration (future to present):

$$E(X_{t+1}) = E(\varepsilon_{t+1}) + E(P_{t+1} \circ X_t)$$

$$\mu_X = \mu_\varepsilon + E\{E(P_{t+1} \circ X_t \mid P_{t+1}, X_t)\} = \mu_\varepsilon + E\{P_{t+1} X_t\} = \mu_\varepsilon + \mu_P \mu_X$$

$$\mu_X = \frac{\mu_\varepsilon}{1 - \mu_P} \quad \begin{array}{l} \text{Stationary} \\ \text{Mean} \end{array}$$

$$E(X_{t+1}^2) = E(\varepsilon_{t+1}^2) + E(P_{t+1} \circ X_t)^2 + 2E(\varepsilon_{t+1})E(P_{t+1} \circ X_t)$$

$$\sigma_X^2 + \mu_X^2 = \sigma_\varepsilon^2 + \mu_\varepsilon^2 + 2\mu_\varepsilon \mu_P \mu_X + E\{E\{(P_{t+1} \circ X_t)^2 \mid P_{t+1}, X_t\}\}$$

$$\sigma_X^2 + \mu_X^2 = \sigma_\varepsilon^2 + \mu_\varepsilon^2 + 2\mu_\varepsilon \mu_P \mu_X + E\{X_t P_{t+1} (1 - P_{t+1}) + X_t^2 P_{t+1}^2\}$$

$$\sigma_X^2 = \frac{\sigma_\varepsilon^2 + \mu_\varepsilon \mu_P}{1 - \sigma_P^2 - \mu_P^2} + \sigma_P^2 \frac{\mu_\varepsilon (\mu_\varepsilon + \mu_P - 1)}{(1 - \sigma_P^2 - \mu_P^2)(1 - \mu_P)^2} \quad \begin{array}{l} \text{Stationary} \\ \text{Variance} \end{array}$$

Index of Dispersion under Stationarity

$$\delta_X = \frac{\sigma_X^2}{\mu_X} = \frac{1 - \mu_P}{1 - \sigma_P^2 - \mu_P^2} (\delta_P + \delta_\varepsilon) + \frac{\sigma_P^2 (\mu_\varepsilon + \mu_P - 1)}{(1 - \sigma_P^2 - \mu_P^2)(1 - \mu_P)}$$

$$\delta_X > 1 \iff \delta_\varepsilon > 1 - \frac{\mu_\varepsilon \sigma_P^2}{(1 - \mu_P)^2}$$

$$\therefore \delta_\varepsilon \geq 1 \implies \delta_X > 1$$

$$E(X_{t+1}X_t) = E(\varepsilon_{t+1}X_t) + E((P_{t+1} \circ X_t)X_t)$$

$$\begin{aligned} \gamma_X(1) + \mu_X^2 &= \mu_\varepsilon \mu_X + E\{E\{(P_{t+1} \circ X_t)X_t \mid P_{t+1}, X_t\}\} \\ &= \mu_\varepsilon \mu_X + E\{P_{t+1}X_t^2\} = \mu_\varepsilon \mu_X + \mu_P(\sigma_X^2 + \mu_X^2) \end{aligned}$$

$$\mu_X = \frac{\mu_\varepsilon}{1 - \mu_P} \Rightarrow \gamma_X(1) = \mu_P \sigma_X^2 \quad \underline{\text{ACVF(1)}}$$

$$E(X_{t+2}X_t) = E(\varepsilon_{t+2}X_t) + E((P_{t+2} \circ \varepsilon_{t+1})X_t) + E(((P_{t+1}P_{t+2}) \circ X_t)X_t)$$

$$\Rightarrow \gamma_X(2) + \mu_X^2 = \mu_\varepsilon \mu_X + \mu_P \mu_\varepsilon \mu_X + (\gamma_P(1) + \mu_P^2)(\sigma_X^2 + \mu_X^2)$$

$$\Rightarrow \gamma_X(2) = \mu_P^2 \sigma_X^2 + (\sigma_X^2 + \mu_X^2) \gamma_P(1) \quad \underline{\text{ACVF(2)}}$$

$$\gamma_X(k) = -\mu_P^2 \mu_X^2 + \mu_\varepsilon \mu_X \sum_{i=1}^{k-1} E\left(\prod_{j=0}^{i-1} P_{k-j}\right) + (\sigma_X^2 + \mu_X^2) E\left(\prod_{i=1}^k P_i\right)$$

ACVF(k ≥ 3)

If $\{P_t = \exp(-a_t^2)\}$ & $\{a_t\} \sim FGN(\mu \in R, \sigma^2 > 0, 1/2 \leq H \leq 1)$,

then $E(P_1 \dots P_t) = E\left\{\exp\left(-\sum_{i=1}^t a_i^2\right)\right\} = E\left\{\exp\left(-\sum_{i=1}^t \lambda_i (W_i - \omega_i)^2\right)\right\}$

$\{W_i\}_{i=1}^t$ IID $N(0,1)$ & $\{\lambda_i \geq 0\}_{i=1}^t$ eigenvalues of

$$V_t = \left(\gamma_a(|i-j|) = \frac{\sigma^2}{2} \left((|i-j|+1)^{2H} - 2(i-j)^{2H} + (|i-j|-1)^{2H} \right) \right)_{\substack{i=1, \dots, t \\ j=1, \dots, t}}$$

$$(\omega_1, \dots, \omega_t)' = Q_t' L_t^{-1} (\mu, \dots, \mu)'$$

Q_t : orthogonal matrix of eigenvectors of V_t

L_t : non-singular lower-triangular matrix

such that $V_t = L_t L_t'$

The Case of FGN with $\mu=0$

$$\begin{aligned} E(P_1 \dots P_t) &= E \left\{ \exp \left(- \sum_{i=1}^t a_i^2 \right) \right\} = E \left\{ \exp \left(- \sum_{i=1}^t \lambda_i W_i^2 \right) \right\} \\ &= \prod_{i=1}^t (1 + 2\lambda_i)^{-1/2} = \prod_{i=1}^t (1 + \sigma^2 \lambda_i^*)^{-1/2} \end{aligned}$$

$$\left\{ \lambda_i^* \geq 0 \right\}_{i=1}^t \text{ eigenvalues of } V_t^* = \frac{2}{\sigma^2} V_t$$

For t=1: $\lambda_1^* = 2$

$$\mu_P = E(P_t) = E(P_1) = E(\exp(-a_t^2)) = (1 + 2\sigma^2)^{-1/2}$$

$$\begin{aligned} \sigma_P^2 &= E(P_t^2) - \mu_P^2 = E(\exp(-2a_t^2)) - (1 + 2\sigma^2)^{-1} \\ &= (1 + 4\sigma^2)^{-1/2} - (1 + 2\sigma^2)^{-1} \end{aligned}$$

The Case of FGN with $\mu=0$

For t=2: $\lambda_1^* = 4^H$; $\lambda_2^* = 4 - 4^H$

For t=3: $\lambda_1^* = 1 + 2 \cdot 4^H - 9^H$

$$\lambda_{2,3}^* = \frac{9^H - 2 \cdot 4^H + 5 \pm \sqrt{(9^H - 2 \cdot 4^H + 5)^2 - 8 \cdot (9^H + 2 \cdot 4^H - 16^H - 1)}}{2}$$

0-inflated Poisson innovations

$$P\{\varepsilon_t = 0\} = p \in (e^{-\theta}, 1) \quad ; \quad \theta > 0 \quad ; \quad l = \frac{1-p}{1-e^{-\theta}} \in (0,1)$$

$$P\{\varepsilon_t = k\} = e^{-\theta} \cdot \frac{\theta^k}{k!} \cdot l = (1-p) \cdot \frac{e^{-\theta}}{1-e^{-\theta}} \cdot \frac{\theta^k}{k!}$$

$$\mu_\varepsilon = l \cdot \theta$$

$$\sigma_\varepsilon^2 = l \cdot \theta + l \cdot (1-l) \cdot \theta^2$$

Overdispersion

Method of Moments Inference

$$\mu_X = \bar{X} = \frac{1}{T} \sum_{t=1}^T X_t$$

$$\sigma_X^2 = S^2 = \frac{1}{T-1} \sum_{t=1}^T (X_t - \bar{X})^2$$

$$\gamma_X(1) = \hat{\gamma}_X(1) = \frac{1}{T} \sum_{t=1}^{T-1} (X_t - \bar{X}) \cdot (X_{t+1} - \bar{X})$$

$$\gamma_X(2) = \hat{\gamma}_X(2) = \frac{1}{T} \sum_{t=1}^{T-2} (X_t - \bar{X}) \cdot (X_{t+2} - \bar{X})$$

$$\gamma_X(3) = \hat{\gamma}_X(3) = \frac{1}{T} \sum_{t=1}^{T-3} (X_t - \bar{X}) \cdot (X_{t+3} - \bar{X})$$

Concluding Remarks

- SUFFICIENT & NECESSARY CONDITIONS FOR STATIONARITY
- REMEDY INSTABILITIES OF FITTED MODEL BY INTRODUCING
 $\mu \neq 0$
- SOLVE NUMERICALLY A HIGHLY NON-LINEAR 5X5 SYSTEM
OF MOM EQUATIONS
- COME UP WITH ALTERNATIVE INFERENCE PROCEDURES
(other than Method of Moments)

**FAREWELL
DEAR
BENJAMIN**

Distributions of statistics of hidden state sequences through the sum-product algorithm

Donald E.K. Martin, NC State Univ.
John A.D. Aston, Warwick Univ.

Outline

- Introduction
- Background Information
- Computation of Distributions in Hidden State Sequences
- Conclusion

Introduction

- Let $\mathbf{O} \equiv O_1, \dots, O_n$ be an observed sequence
- $\mathbf{S} \equiv S_1, \dots, S_n$ is a corresponding hidden state sequence
- We study inference for statistics of \mathbf{S} conditional on \mathbf{O}

- States can serve as labels for $\mathbf{0}$
 - Krogh (1997), Durbin et al. (1988) DNA
 - Hamilton (1989) business cycles
(may be interested in “runs”)
- States can be the true values of a noisy observed sequence
 - McEliece (1998) transmitted codewords
 - Baxter (1982) noisy image pixels

- Past approaches to such problems:
 - determine Viterbi (most likely) sequence and obtain the statistic from that sequence
 - sample from $p(\mathbf{s}|\mathbf{o})$
 - Aston and Martin (2007) introduced a Markov chain based method to compute exact distributions for statistics of HMM states

Background Information

- The model
- Factor graphs
- Sum-product algorithm

The model

- We use the discriminative model, a conditional random field

$$p(\mathbf{s}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})} \prod_{c \in \mathbf{C}} \Psi_c(\mathbf{s}_c, \mathbf{o})$$

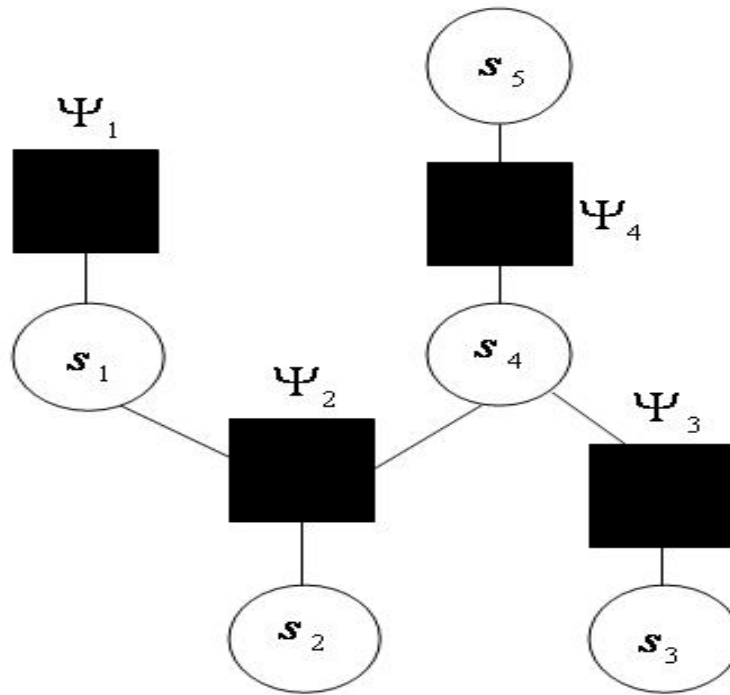
$$\Psi_c(\mathbf{s}_c, \mathbf{o}) = \exp \left[\sum_l \lambda_l f_l(\mathbf{s}_c, \mathbf{o}) \right]$$

$$Z(\mathbf{o}) = \sum_{\mathbf{s}} \prod_{c \in \mathbf{C}} \Psi_c(\mathbf{s}_c, \mathbf{o})$$

- If each s_j takes on K values then \mathbf{s} takes on K^n , and thus, in general, computation of $Z(\mathbf{o}) = \sum_{\mathbf{s}} \prod_{c \in C} \Psi_c(\mathbf{s}_c, \mathbf{o})$ is intractable for large n
- We consider models with dependence structure that allows exact inference to be performed

Factor graphs

- Corresponding to the model is a *factor graph* $F = (V, E)$, a bipartite graph with two types of nodes:
 - variable nodes
 - factor nodes (with edges connecting the functions to their arguments)
- Graph describes how a global function factors into a product of local functions



Factor graph corresponding to the joint distribution

$$p(s_1, \dots, s_5 | \mathbf{o}) = [Z(\mathbf{o})]^{-1} \Psi_4(s_4, s_5, \mathbf{o}) \Psi_3(s_3, s_4, \mathbf{o}) \Psi_2(s_1, s_2, s_4, \mathbf{o}) \Psi_1(s_1, \mathbf{o})$$

- Note that “generative” models such as Bayesian networks and HMMs can be represented by our model

e.g. for HMMs
$$p(\mathbf{s}, \mathbf{o}) = \prod_{t=1}^n \overbrace{\xi_1(s_t | s_{t-1}) \xi_2(o_t | s_t)}^{\Psi_t(s_{t-1}, s_t, o_t)}$$

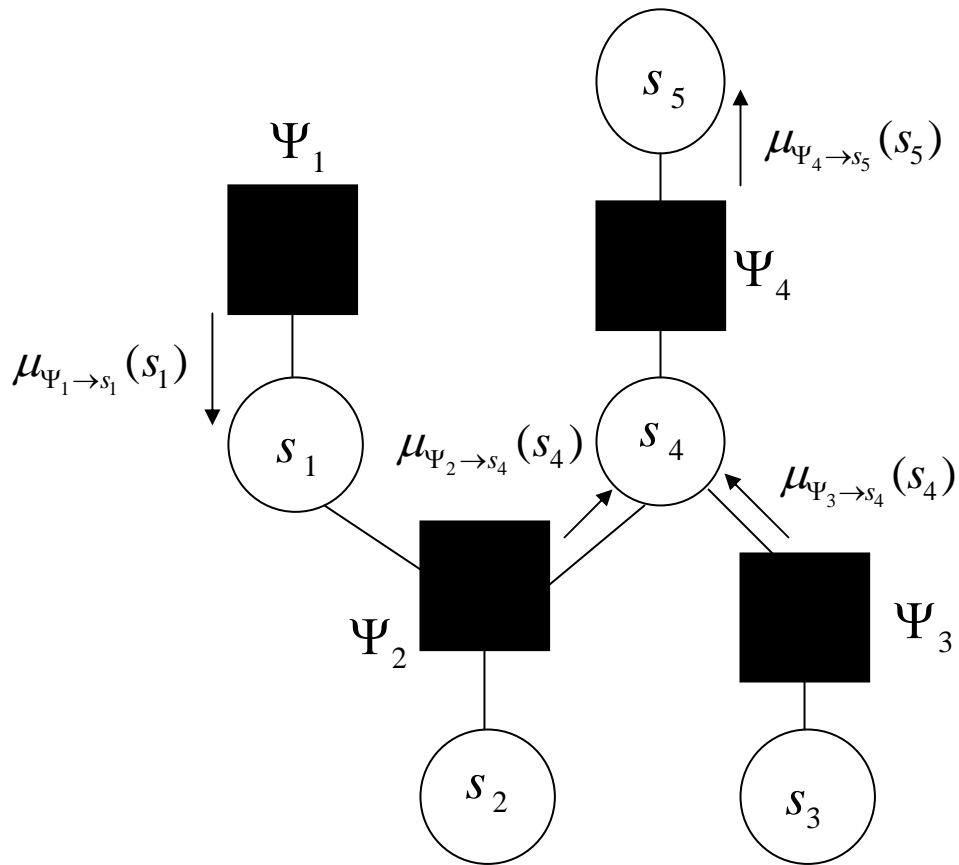
$$Z(\mathbf{o}) = \sum_{\mathbf{s}} p(\mathbf{s}, \mathbf{o})$$

$$p(\mathbf{s} | \mathbf{o}) = p(\mathbf{s}, \mathbf{o}) / Z(\mathbf{o})$$

The sum-product algorithm (SPA), Kschischang et al. (2001)

- Operates in a cycle-free factor graph to compute marginal functions by exploiting the way a global function factors
- Nodes are treated as “processors” and “messages” are sent between them
- Essentially equivalent to the “generalized distributive law” (Aji and McEliece, 2000) over junction trees

- A variety of algorithms in artificial intelligence, signal processing and digital communications can be derived as instances of the SPA
 - forward/backward algorithm for HMMs
 - Viterbi algorithm (max-product semiring)
 - iterative “turbo” decoding
 - Pearl’s belief propagation for Bayesian networks
 - Kalman filter
 - certain FFT algorithms



$$p(s_5 | \mathbf{o}) = \sum_{\sim s_5} p(\mathbf{s} | \mathbf{o}) = [Z(\mathbf{o})]^{-1} \sum_{s_4} \underbrace{\Psi_4(s_4, s_5, \mathbf{o})}_{\mu_{\Psi_4 \rightarrow s_5}(s_5)} \underbrace{\sum_{s_3} \Psi_3(s_3, s_4, \mathbf{o})}_{\mu_{\Psi_3 \rightarrow s_4}(s_4)} \underbrace{\sum_{s_1, s_2} \Psi_2(s_1, s_2, s_4, \mathbf{o})}_{\mu_{\Psi_2 \rightarrow s_4}(s_4)} \underbrace{\Psi_1(s_1, \mathbf{o})}_{\mu_{\Psi_1 \rightarrow s_1}(s_1)}$$

$$Z(\mathbf{o}) = \sum_{s_5} \underbrace{\mu_{\Psi_4 \rightarrow s_5}(s_5)}_{\mu_{\Psi_4 \rightarrow s_5}(s_5)}$$

- A perfect elimination sequence of the SPA is one where marginalization can be carried out without enlarging local domains
- More than one perfect elimination sequence can exist

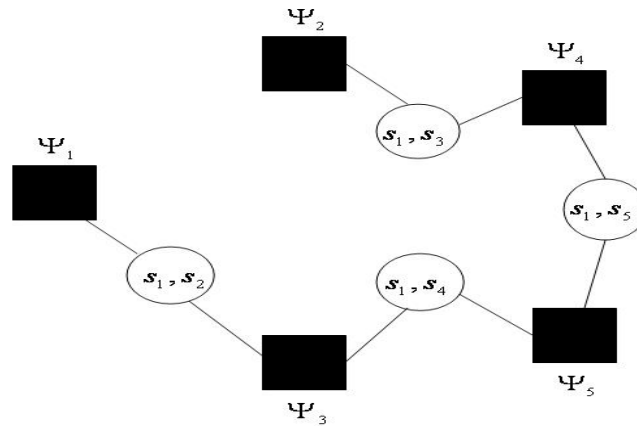
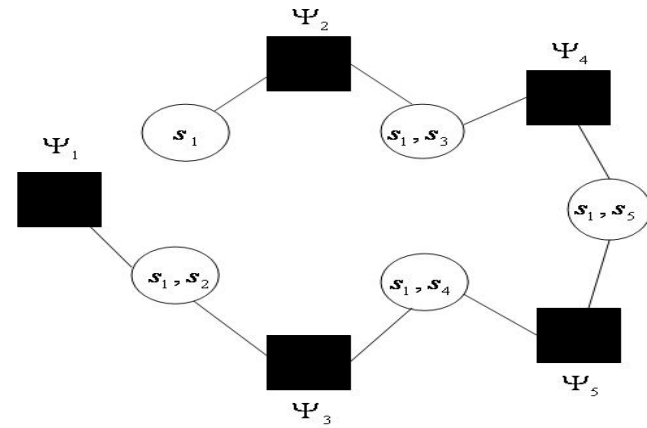
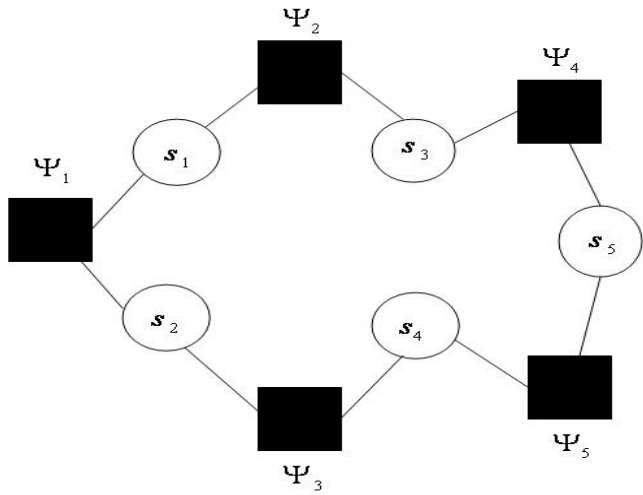
- To obtain $p(s_j | \mathbf{o})$:
- s_j is treated as the root of the graph
- Initial messages are sent from leaves
- Message sent from node to its parent after it has received messages from its children

$$\mu_{s_t \rightarrow \Psi_c}(s_t) = \prod_{\Psi_d \in ne(s_t) \setminus \Psi_c} \mu_{\Psi_d \rightarrow s_t}(s_t)$$

$$\mu_{\Psi_c \rightarrow s_v}(s_v) = \sum_{\mathbf{s}_c \setminus s_v} \left(\Psi_c(\mathbf{s}_c, \mathbf{o}) \prod_{s_u \in \mathbf{s}_c \setminus s_v} \mu_{s_u \rightarrow \Psi_c}(s_u) \right)$$

For a factor graph F with cycles:

- Obtain a spanning tree T
- A variable is “stretched” by including it in variable nodes along the unique path of T from the variable to nodes of F that can be reached from it on a path of length two
- Delete redundant edges



$$p(\mathbf{s}|\mathbf{o}) = [Z(\mathbf{o})]^{-1} \Psi_5(s_4, s_5, \mathbf{o}) \Psi_4(s_3, s_5, \mathbf{o}) \Psi_3(s_2, s_4, \mathbf{o}) \Psi_2(s_1, s_3, \mathbf{o}) \Psi_1(s_1, s_2, \mathbf{o})$$

Computation of Distributions of Statistics of a Hidden State Sequence

- We compute distributions of a count statistic $\theta_n = h_n(\sigma_1, \dots, \sigma_n) \in \Theta = (\theta^{(1)}, \dots, \theta^{(|\Theta|)})$ associated with a pattern
- $\sigma_1, \dots, \sigma_n$ is the order of evaluating θ_n

- A sequential ordering $\sigma_1, \dots, \sigma_n = s_1, \dots, s_n$ is most convenient when the pattern has words of length greater than one
- For that case we introduce $q_t = g(s_1, \dots, s_t)$ to keep track of pattern progress
- If the pattern consists of singletons, we can evaluate the statistic in any order (e.g. a multinomial distribution)

- Let $Q \times \Theta = \{(q, \theta)^{(1)}, \dots, (q, \theta)^{(\omega)}\}$
- We first obtain an $\omega \times 1$ vector $\tilde{\gamma}_n$ that has entries $\Pr[(q_n, \theta_n) = (q, \theta)^j]$
- Then we obtain a $|\Theta| \times 1$ vector γ_n holding probabilities for θ_n through

$$\gamma_n = \underbrace{\Delta}_{|\Theta| \times \omega} \tilde{\gamma}_n$$

- For each component $\tilde{\gamma}_n((q, \theta)^{(j)})$ of $\tilde{\gamma}_n$,

$$\begin{aligned}\tilde{\gamma}_n((q, \theta)^{(j)}) &= \sum_{\mathbf{s}} \Pr(\mathbf{s}, (q_n, \theta_n)^{(j)} = (q, \theta)^{(j)} \mid \mathbf{o}) \\ &= \sum_{\mathbf{s}} p(\mathbf{s} \mid \mathbf{o}) I \left[\left((q, \theta)^{(j)} \right) = \left(g(\mathbf{s}), h(\mathbf{s}) \right) \right]\end{aligned}$$

- The computation is similar to computing $Z(\mathbf{o})$ (we also must keep track of the value of the statistic)

- Vectors φ_n serve as indicators of the value of the statistic

$$\varphi_0 = (1 \ 0 \ \dots \ 0)^T \qquad \varphi_n = \Delta_{\sigma_n} \Delta_{\sigma_{n-1}} \dots \Delta_{\sigma_1} \varphi_0$$

where Δ_{σ_j} are $\omega \times \omega$ matrices with a one in the (i, j) position if and only if $(q, \theta)^{(j)} \rightarrow (q, \theta)^{(i)}$ with the occurrence of σ_j .

- Using our representation of $p(\mathbf{s}|\mathbf{o})$ we obtain:

$$\tilde{\gamma}_n = Z(\mathbf{o})^{-1} \sum_{\mathbf{s}} \left(\prod_c \Psi_c(\mathbf{s}_c, \mathbf{o}) \right) \left(\prod_{k=1}^n \Delta_{\sigma_k} \right) \varphi_0 = \frac{\sum_{\sigma_n} \Delta_{\sigma_n} \cdots \sum_{\sigma_1} \Delta_{\sigma_1} \left(\prod_c \Psi_c(\mathbf{s}_c, \mathbf{o}) \right) \varphi_0}{\sum_{\mathbf{s}} \prod_c \Psi_c(\mathbf{s}_c, \mathbf{o})}$$

- When $\sigma_1, \dots, \sigma_n$ is a perfect elimination sequence of the SPA, factors can be efficiently distributed throughout sums

$$\tilde{\mu}_{\sigma_t \rightarrow \Psi_c}(\sigma_t, q_t, \theta_t) = \prod_{\Psi_d \in ne(\sigma_t) \setminus \Psi_c} \tilde{\mu}_{\Psi_d \rightarrow \sigma_t}(\sigma_t, q_t, \theta_t)$$

$$\tilde{\mu}_{\Psi_c \rightarrow \sigma_u}(\sigma_u, q_u, \theta_u) = \sum_{\mathbf{s}_c \setminus \sigma_u} \left(\Psi_c(\mathbf{s}_c, \mathbf{o}) \prod_{s_r \in \mathbf{s}_c \setminus \sigma_u} \Delta_{s_r} \tilde{\mu}_{s_r \rightarrow \Psi_c}(s_r, q_r, \theta_r) \right)$$

$$\tilde{\gamma}_n = Z(\mathbf{o})^{-1} \sum_{s_u \in \mathbf{s}_c^{(n)}} \prod_{s_u \in \mathbf{s}_c^{(n)}} \Delta_{s_u} \left(\prod_{\Psi_d \in ne(s_u)} \tilde{\mu}_{\Psi_d \rightarrow s_u}(s_u, q_n, \theta_n) \right)$$

$$\gamma_n = \Lambda \tilde{\gamma}_n$$

Examples

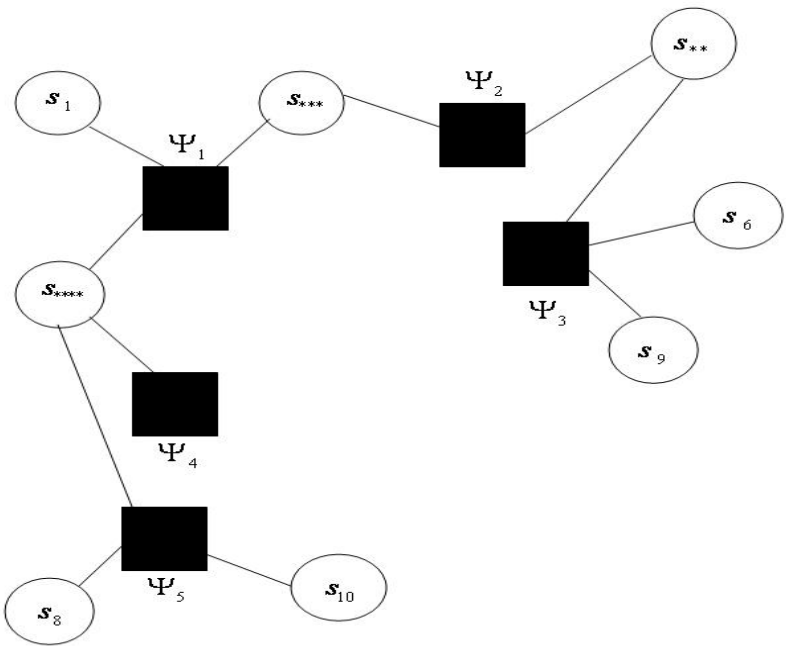
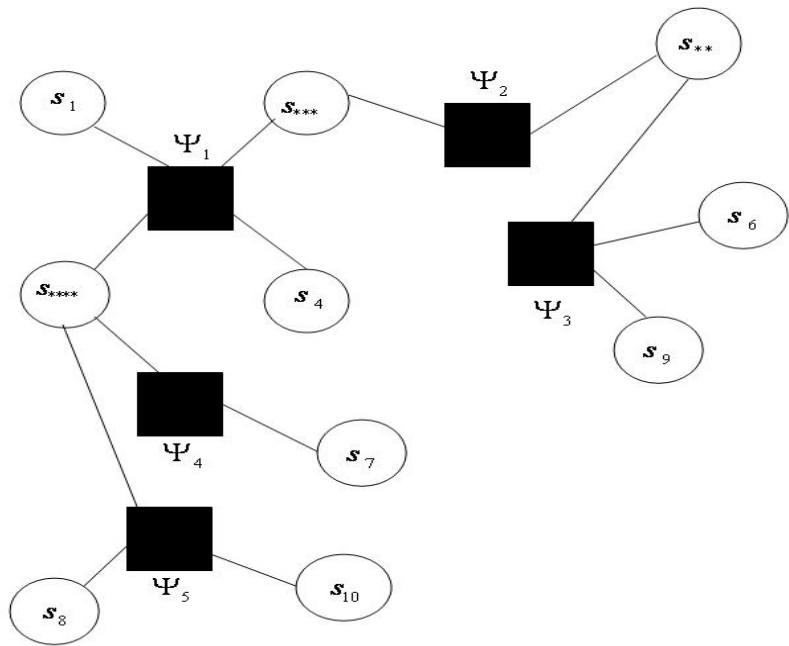
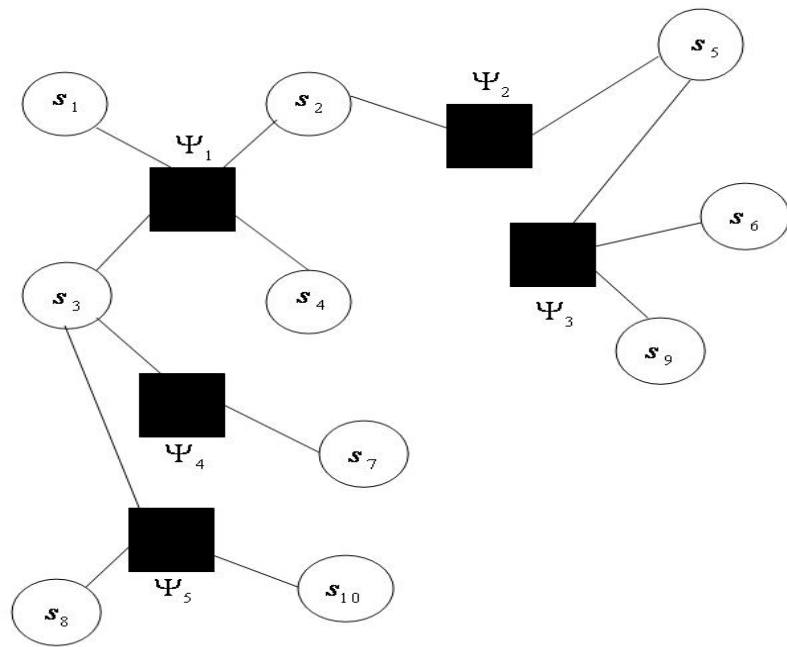
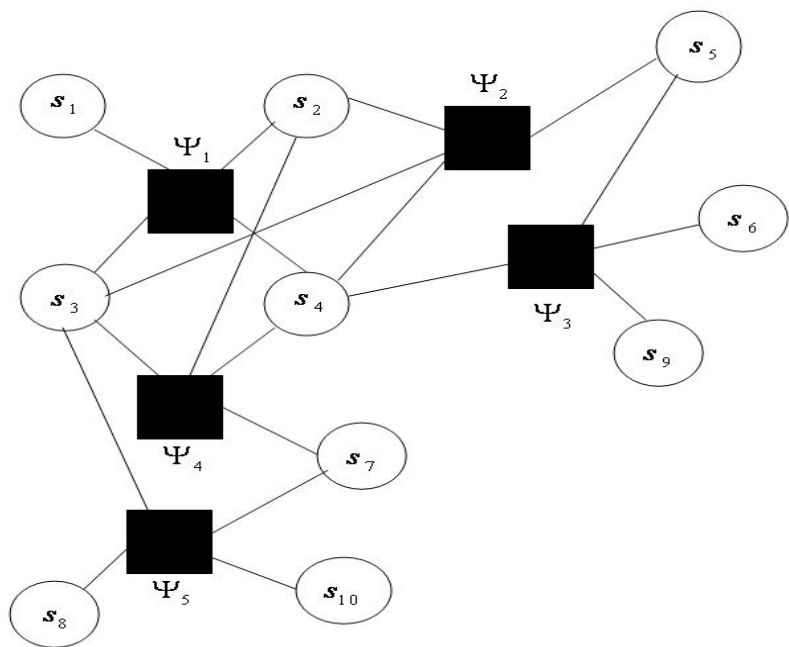
- Let $\theta_t = (\eta_{1t}, \dots, \eta_{Kt})$

$$p(\mathbf{s}|\mathbf{o}) = [Z(\mathbf{o})]^{-1} \Psi_1(s_1, s_2, s_3, s_4, \mathbf{o}) \Psi_2(s_2, s_3, s_4, s_5, \mathbf{o}) \Psi_3(s_4, s_5, s_6, s_9, \mathbf{o}) \\ \times \Psi_4(s_2, s_3, s_4, s_7, \mathbf{o}) \Psi_5(s_3, s_7, s_8, s_{10}, \mathbf{o})$$

- No pattern progress is needed (patterns are the symbols $s^{(1)}, \dots, s^{(K)}$)

$$\omega = |\Theta| \equiv \binom{n + K - 1}{K - 1}$$

- The matrices Δ_{σ_t} for the transitions based on $\sigma_t = s^{(l)}$ have a one in location (i, j) if column j corresponds to $(\eta_{t-1,1}, \dots, \eta_{t-1,l}, \dots, \eta_{t-1,K})$ and row i to $(\eta_{t-1,1}, \dots, \eta_{t-1,l} + 1, \dots, \eta_{t-1,K})$



- Passed messages:

$$\tilde{\mu}_{s_j \rightarrow \Psi_d}(s_j, \theta_1) = 1 \quad (j, d) = (1, 1), (6, 3), (9, 3), (8, 5), (10, 5)$$

$$\tilde{\mu}_{\Psi_5 \rightarrow s_{****}}(s_{****}, \theta_3) = \sum_{s_8, s_{10}} \Delta_{s_{10}} \Delta_{s_8} \Psi_5(s_3, s_7, s_8, s_{10}, \mathbf{o}) \varphi_0$$

$$\tilde{\mu}_{\Psi_4 \rightarrow s_{****}}(s_{****}, \theta_1) = \Psi_4(s_2, s_3, s_4, s_7, \mathbf{o})$$

$$\tilde{\mu}_{s_{****} \rightarrow \Psi_1}(s_{****}, \theta_3) = \tilde{\mu}_{\Psi_4 \rightarrow s_{****}}(s_{****}, \theta_1) \tilde{\mu}_{\Psi_5 \rightarrow s_{****}}(s_{****}, \theta_3)$$

$$\tilde{\mu}_{\Psi_1 \rightarrow s_{***}}(s_{***}, \theta_5) = \sum_{s_1, s_7} \Delta_{s_1} \Delta_{s_7} \Psi_1(s_1, s_2, s_3, s_4, \mathbf{o}) \tilde{\mu}_{s_1 \rightarrow \Psi_1}(s_1, \theta_1) \tilde{\mu}_{s_{****} \rightarrow \Psi_1}(s_{****}, \theta_3) = \tilde{\mu}_{s_{***} \rightarrow \Psi_2}(s_{***}, \theta_5)$$

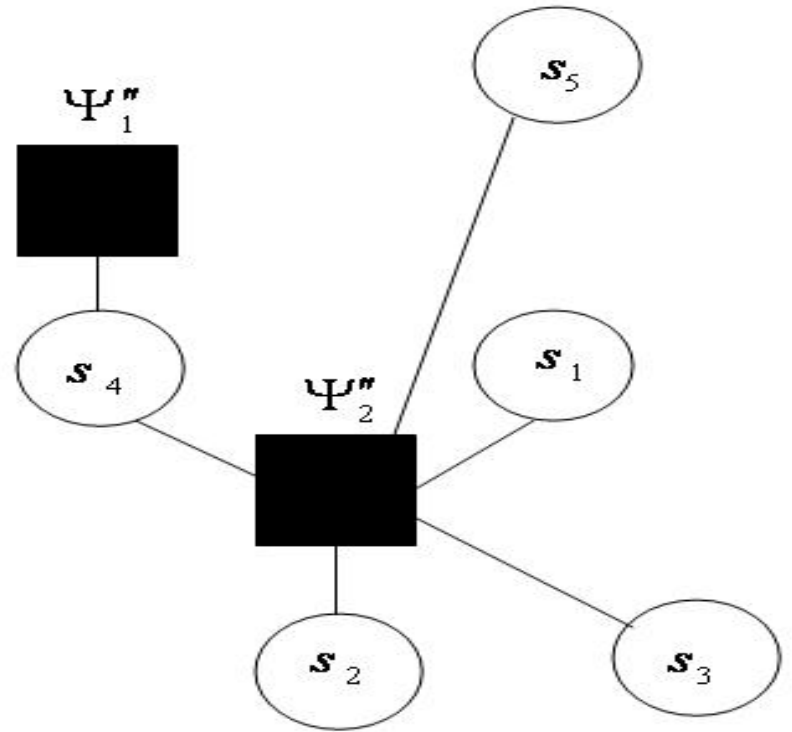
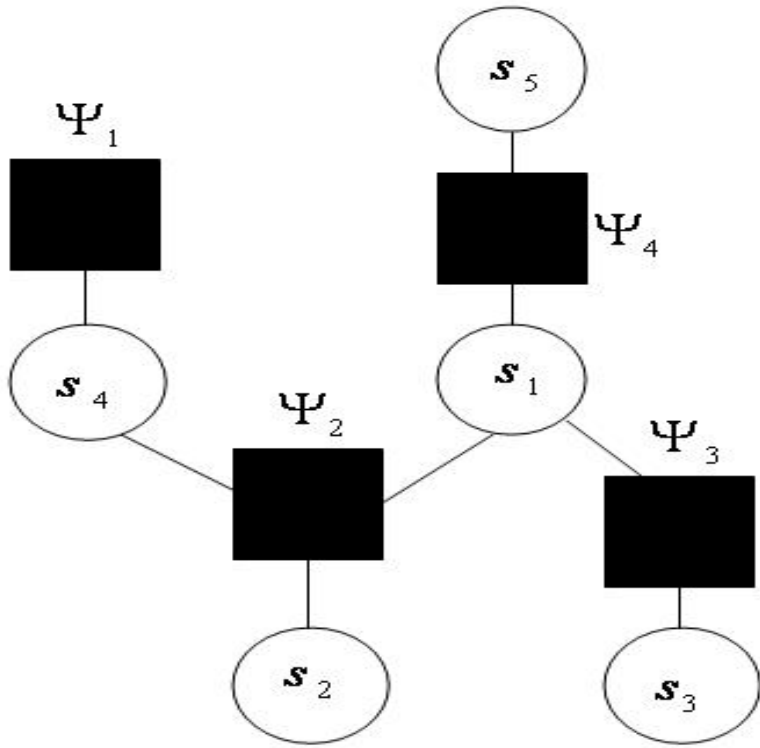
$$\tilde{\mu}_{\Psi_2 \rightarrow s_{**}}(s_{**}, \theta_7) = \sum_{s_2, s_3} \Delta_{s_2} \Delta_{s_3} \Psi_2(s_2, s_3, s_4, s_5, \mathbf{o}) \tilde{\mu}_{s_{***} \rightarrow \Psi_2}(s_{***}, \theta_5)$$

$$\tilde{\mu}_{\Psi_3 \rightarrow s_{**}}(s_{**}, \theta_9) = \sum_{s_6, s_9} \Delta_{s_6} \Delta_{s_9} \Psi_3(s_4, s_5, s_6, s_9, \mathbf{o})$$

$$\gamma_{10} = Z(\mathbf{o})^{-1} \sum_{s_4} \sum_{s_5} \Delta_{s_4} \Delta_{s_5} \tilde{\mu}_{\Psi_2 \rightarrow s_{**}}(s_{**}, \theta_7) \tilde{\mu}_{\Psi_3 \rightarrow s_{**}}(s_{**}, \theta_9)$$

- Consider the computation of the distribution of the number of overlapping occurrences of $L_1 = 111$ in $\mathbf{s} = (s_1, \dots, s_5)$

Let $p(\mathbf{s}|\mathbf{o}) = [Z(\mathbf{o})]^{-1} \Psi_4(s_1, s_5, \mathbf{o}) \Psi_3(s_1, s_3, \mathbf{o}) \Psi_2(s_1, s_2, s_4, \mathbf{o}) \Psi_1(s_4, \mathbf{o})$



$$\Psi_1'' = \Psi_1$$

$$\Psi_2''(s_1, s_2, s_3, s_4, s_5, \mathbf{o}) = \Psi_4(s_1, s_5, \mathbf{o}) \Psi_3(s_1, s_3, \mathbf{o}) \Psi_2(s_1, s_2, s_4, \mathbf{o})$$

- Transitions of (q, θ) values (for Δ matrices)

| (q, θ) | after one | after zero |
|---------------|-----------|------------|
|---------------|-----------|------------|

| | | |
|--------------------|------------|--------------------|
| $(\varepsilon, 0)$ | $(1, 0)$ | $(\varepsilon, 0)$ |
| $(1, 0)$ | $(11, 0)$ | $(\varepsilon, 0)$ |
| $(11, 0)$ | $(111, 1)$ | $(\varepsilon, 0)$ |
| $(\varepsilon, 1)$ | $(1, 1)$ | $(\varepsilon, 1)$ |
| $(111, 1)$ | $(111, 2)$ | $(\varepsilon, 1)$ |
| $(111, 2)$ | $(111, 3)$ | $(\varepsilon, 2)$ |

- If states are listed in the order

$(\varepsilon, 0)$ $(1, 0)$ $(11, 0)$ $(\varepsilon, 1)$ $(1, 1)$ $(111, 1)$ $(\varepsilon, 2)$ $(111, 2)$ $(111, 3)$

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

- To reduce the number of q values, we introduce an Aho-Corasick automaton

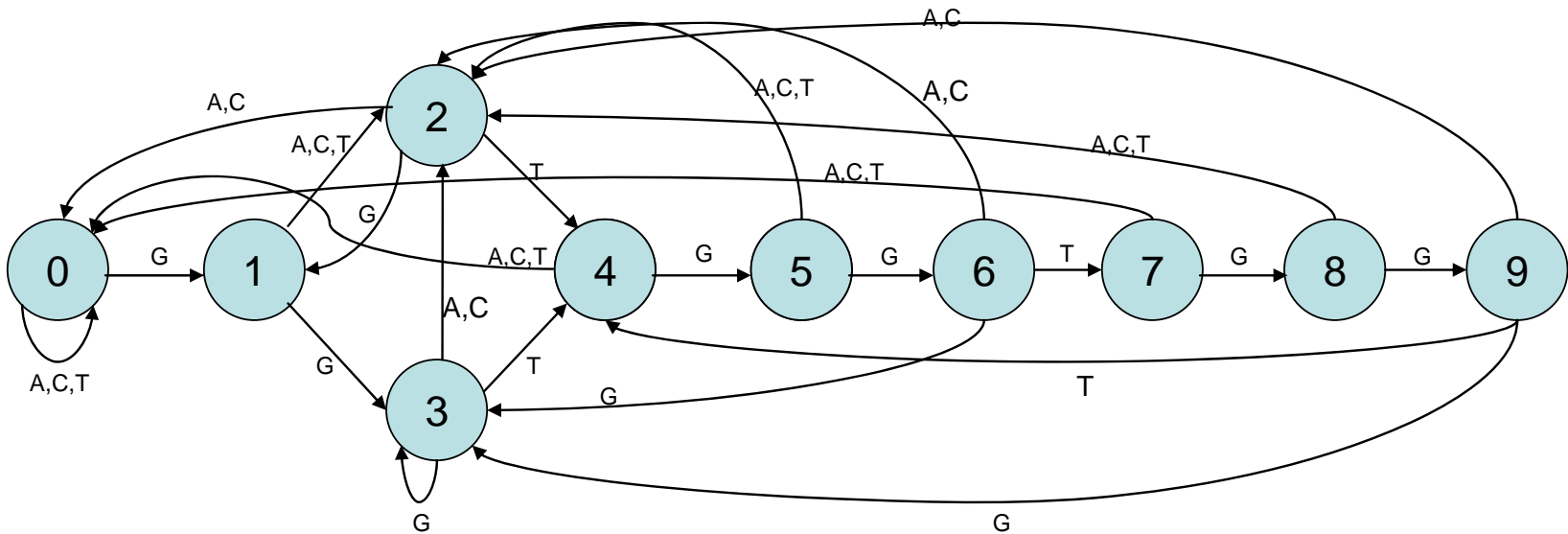
$$D = (Q, \chi, \delta_Q, \varepsilon, T)$$

Q set of prefixes of pattern

$\delta_Q : Q \times \chi \rightarrow Q$ such that $\delta_Q(q, a)$ is the longest suffix of qa that is in Q
terminal states T are words of the pattern

- The automaton is then minimized using the Hopcroft (1971) algorithm

- Supposed that we are interested in the number of overlapping occurrences of the Chi motif $L = GNTGGTGG$, $N \in \chi = \{A, C, G, T\}$
- There are 30 pattern prefixes, but only 10 states in the minimal automaton



Conclusion

- We have given a method to compute exact distributions of statistics of hidden state sequences
- The method can be applied to a variety of models, both discriminative and generative
- For certain elimination orders, computation is as efficient as computing marginals

Future work

- Analyze data sets
- Applications, e.g. in information/coding theory: can we facilitate decoding based on new decoding criteria?
- Approximate inference (bounds on deviation from true distributions?)

References

- A. M. Aji and R. J. McEliece, *IEEE Transactions on Information Theory*, **46**, 325-343 (2000).
- J. A. D. Aston and D. E. K. Martin, *Annals of Applied Statistics*, **1**(2), 585-611 (2007).
- F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, *IEEE Transactions on Information Theory*, **47**(2), 498-519, (2001).

Thank you!!

Classification of Eye Movement Data

Ritaja Sur and Benjamin Kedem

Department of Mathematics,
University of Maryland, College Park

Advances in Statistics and Applied Probability: Unified Approaches

July 31, 2009



Science News

 Share  Blog  Cite

Eye-Movement Study Shows Glimpse Of How Brain Plans Movement

ScienceDaily (Jan. 11, 2001) — January 11, 2001 — Researchers are getting closer to understanding how the brain is able to hold a moving image on the retina, yielding smooth, relatively stable images of the moving object. Howard Hughes Medical Institute (HHMI) researchers have discovered that a region of the brain that was formerly believed to control eye movement is actually involved in the high-level planning of movement.

See also:

Health & Medicine

- [Eye Care](#)
- [Psychology Research](#)
- [Nervous System](#)

Mind & Brain

- [Perception](#)
- [Neuroscience](#)

Their findings offer new insight into how the area in the brain's motor cortex adjusts eye movement to track objects, say the researchers. And their experimental approach—which involves measuring and altering via electrical stimulation the tiny eye movements by which rhesus monkeys track a spot of light—provides a precise, quantitative approach to studying the basic

Ads by Google

[Advertise here](#)

MindWell Psychology

Testing: Educ, Devel, Emotional Therapy: Goal oriented, family focus
www.mindwellpsychology.com

Mental Illness Symptoms

Find Info On Mental Health & Diagnosing Neurological Symptoms
www.wyethneuroscience.com

Forensic Psychology

Take Your Psychology Career to the Next Level. Learn Online Anytime.
online.argosy.edu/Psychology

NCI Cancer Trials

Clinical Trial Info from NCI at NIH Learn More About Resources/Options
BethesdaTrials.Cancer.gov

Brain Rules - Brain Rules

Read the new book that will change your brain

Science News

 Share  Blog  Cit

Eye Movement Studies To Help Diagnose Mental Illness

ScienceDaily (June 18, 2003) — Researchers at the University of Illinois at Chicago are studying subtle abnormalities in eye movements that may one day be used to diagnose psychiatric disease.

Ads by Google

Advertise her

See also:

Health & Medicine

- [Eye Care](#)
- [Diseases and Conditions](#)
- [Mental Health Research](#)

Mind & Brain

- [Intelligence](#)
- [Disorders and Syndromes](#)

Irregularities in how the eyes track a moving object reflect defects in the neural circuitry of the brain and appear to correspond with particular types of mental disorders.

Schizophrenic patients, for example, have difficulty keeping their eyes focused on slow-moving objects. With new technology, these abnormalities can be measured precisely and compared with normal patterns.

Wet Macular Degeneration

Info on wet AMD and possible treatment option. Learn more today.

www.LivingWithAMD.com

Treat Bulging Veins

Treating Varicose Veins Without Surgery Since 1981. Free Consults.

www.VeinClinics.com

Mental Illness Symptoms

Find Info On Mental Health & Diagnosing Neurological Symptoms

www.wyethneuroscience.com

Get Bulimia Treatment

Providing Quality Care For Over 20 Yrs. Call Us Fo

Background

- Land and Mcleod (2000) - For a successful hit cricketers fixate on future points of contact of the ball and the ground.
- Reina and Schwartz (2003) studied monkey's fixation points in a repeated drawing task.
- Mataric and Pomplun (1998) studied movement imitation and showed that people tend to fixate on end effectors.

Questions

- What visual inputs are captured when we intend to imitate a movement?
- What are the computations performed by the brain in movement imitation?
- Whether there is any difference when a person intends to imitate?

Data Description

- The eye movement data was obtained from the work of Lior Noy.
- Eye gaze as a response to the motion of the hand was recorded using an eye tracker.
- Coordinates of the wrist, elbow and shoulder were recorded.
- Number of human subjects = 7.
- Number of distinct hand movements = 10.
- Number of conditions (Watch and Imitate) = 2.
- Data recorded in x- and y- coordinates.
- Length of time series ranges from 801 (6.68s) to 1609 (13.41s) data points.
- r-coordinate $r_t = (x_t^2 + y_t^2)^{1/2}$ considered for analysis.

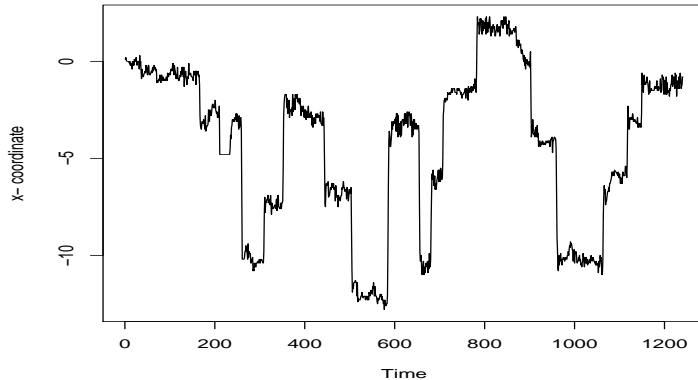


Figure: x- coordinate for Subject 5 Movement 2 (Watch).

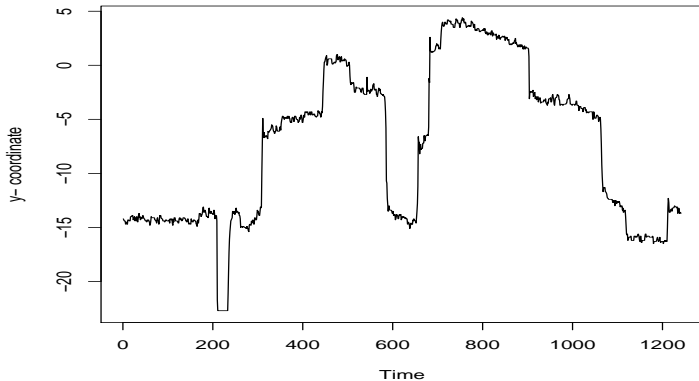


Figure: y- coordinate for Subject 5 Movement 2 (Watch).

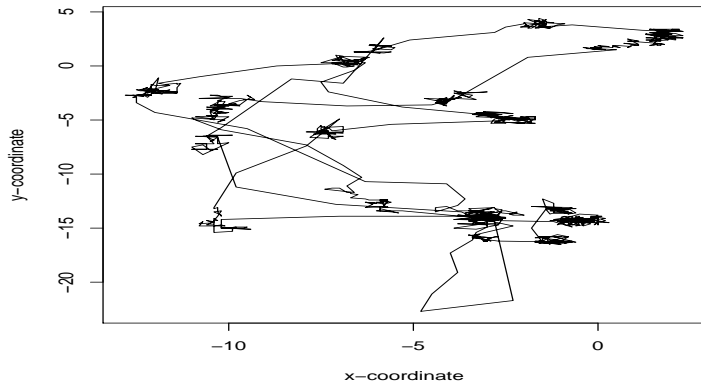


Figure: x- coordinate vs y- coordinate for Subject 5 Movement 2 (Watch).

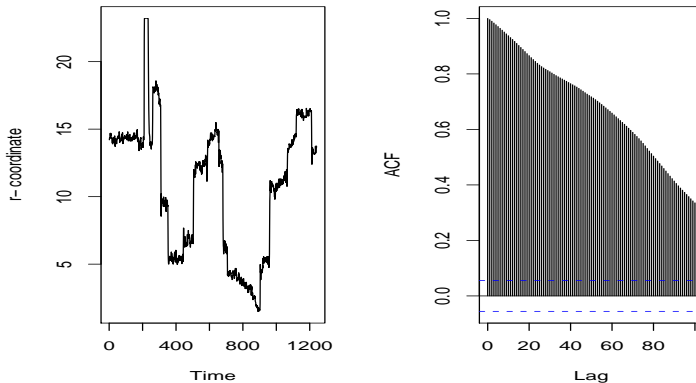


Figure: Time series plot and Autocorrelation plot of the r- coordinate for Subject 5 Movement 2 (Watch).

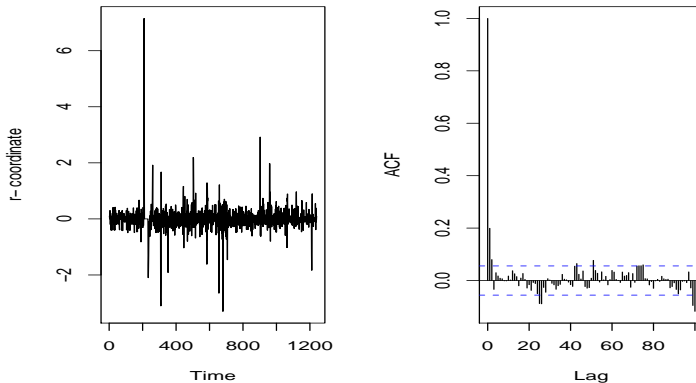


Figure: Time series plot and Autocorrelation plot of the differenced r-coordinate for Subject 5 Movement 2 (Watch).

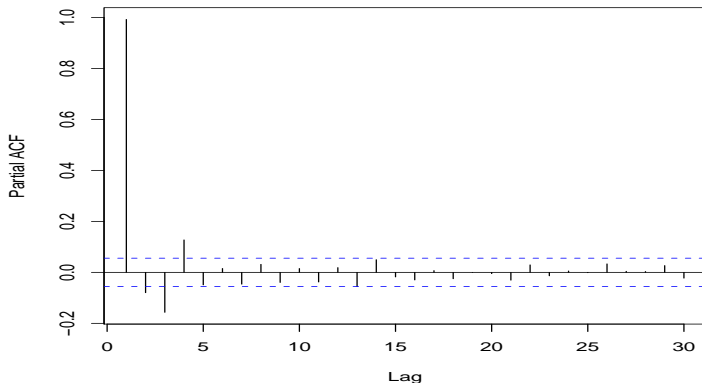


Figure: Partial Autocorrelation plot of the differenced x- coordinate for Subject 5 Movement 2 (Watch).

Non-parametric approach using HOC

Let ∇ be the difference operator defined as.

$$\nabla Z_t \equiv Z_t - Z_{t-1}$$

In general, for $k = 0, 1, 2, \dots$, $\nabla^k Z_t$ is given by

$$\nabla^k Z_t = \sum_{j=0}^k \binom{k}{j} (-1)^j Z_{t-j}$$

where $\nabla^0 Z_t \equiv Z_t$. For each $k = 1, 2, \dots$, we further obtain the binary clipped process $X_t(k)$ and the HOC counts D_k .

$$X_t(k) = \begin{cases} 1, & \text{if } \nabla^{k-1} Z_t \geq 0 \\ 0, & \text{if } \nabla^{k-1} Z_t < 0 \end{cases}$$

HOC continued

$$D_k = \sum_{t=1}^N [X_t(k) - X_{t-1}(k)]^2$$

$$\Delta_k \equiv \begin{cases} D_1 & \text{if } k = 1 \\ D_k - D_{k-1} & \text{if } k = 2, 3, \dots, K - 1 \\ (N - 1) - D_{K-1} & \text{if } k = K \end{cases}$$

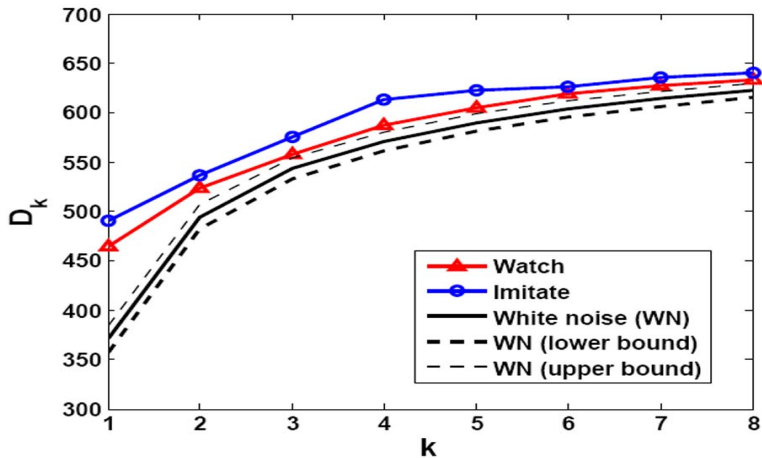
Distance measure from white Gaussian noise:

$$\psi^2 = \sum_{k=1}^K \frac{(\Delta_k - m_k)^2}{m_k}$$

where $m_k = E(\Delta_k)$ can be obtained using

$$E[D_k] = (N - 1) \left[\frac{1}{2} + \frac{1}{\pi} \sin^{-1} \left(\frac{k-1}{k} \right) \right]$$

Subject 1, Movement 9



Application to eye data

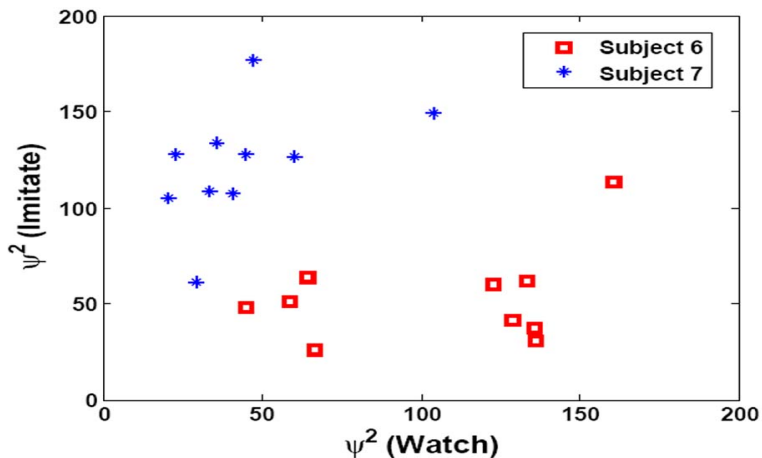
- For each eye signal, the calculations of the D_k 's are based on a 751-point segments, corresponding to $t = 50, 51, \dots, 800$.
- The respective ψ^2 distances from white noise were computed using D_k 's for $k = 1, 2, \dots, 8$.
- For each subject, the average ψ^2 values were computed from the 10 movements.
- To test the hypothesis that ψ^2 values differ significantly under the watch and imitate conditions for each subject, wilcoxon rank-sum test was performed.
- To get further insight from the 70 cases for each of the two conditions, each of the 70 time series was divided into two equal parts. Essentially, this implies as if the subjects were viewing 20 movements instead of 10.

Results using ψ^2 Statistic

Table: r- Average ψ^2 distance from white noise.

| Subject | Watch1 | Imitate1 | $H_0 : W1 = I1$ | Watch2 | Imitate2 | $H_0 : W2 = I2$ |
|---------|---------|----------|-----------------|--------|----------|-----------------|
| 1 | 53.317 | 85.527 | 0.0116 | 33.237 | 50.629 | 0.0037 |
| 2 | 37.064 | 48.627 | 0.1088 | 27.884 | 32.373 | 0.1917 |
| 3 | 92.091 | 66.617 | 0.0526 | 52.357 | 44.101 | 0.0637 |
| 4 | 75.271 | 79.421 | 0.2894 | 40.421 | 46.265 | 0.1917 |
| 5 | 51.149 | 50.656 | 0.4559 | 32.863 | 33.162 | 0.3490 |
| 6 | 105.241 | 53.506 | 0.0014 | 60.175 | 34.186 | 0.0026 |
| 7 | 43.894 | 122.453 | 1.08E-05 | 26.643 | 74.597 | < 0.005 |
| Average | 65.43 | 72.40 | | 39.08 | 45.04 | |
| Median | 53.2 | 66.62 | | 33.24 | 44.10 | |

Subject 6, Subject 7



Distance measures - 1

Let x_t and y_t be two zero mean stationary time series of length n .
The various distance measures considered are:

- Euclidean Distance:

$$d_{EUCL}(x, y) = \sqrt{\sum_{t=1}^n (x_t - y_t)^2}$$

- Distance based on estimated autoregressive weights:

$$d_{AR}(x, y) = \sqrt{\sum_{j=1}^{\infty} (\hat{\pi}_{j,x} - \hat{\pi}_{j,y})^2}$$

Distance measures - 2

- Distance based on estimated autocorrelations:

$$d_{ACFU}(x, y) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)'(\hat{\rho}_x - \hat{\rho}_y)}$$

$$d_{ACFG}(x, y) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)' \Omega (\hat{\rho}_x - \hat{\rho}_y)}$$

- Distance based on periodogram coordinates:

$$d_{NP}(x, y) = \sqrt{\sum_{j=1}^{\lfloor n/2 \rfloor} [NP_x(w_j) - NP_y(w_j)]^2}$$

$$d_{LNP}(x, y) = \sqrt{\sum_{j=1}^{\lfloor n/2 \rfloor} [\log NP_x(w_j) - \log NP_y(w_j)]^2}$$

$$d_{KLF D}(x, y) = \sum_{j=1}^{\lfloor n/2 \rfloor} \left[\frac{NP_x(w_j)}{NP_y(w_j)} - \log \frac{NP_x(w_j)}{NP_y(w_j)} - 1 \right]$$

Distance measures - 3

- Distance based on HOC:

$$d_{Delta}(x, y) = \sqrt{\sum_{k=1}^K (\Delta_{k,x} - \Delta_{k,y})^2}$$

$$d_D(x, y) = \sqrt{\sum_{k=1}^K (D_{k,x} - D_{k,y})^2}$$

Simulation study

- Simulated 1000 replications of AR(0.3) and AR(0.4) processes.
- Length of each time series = n .
- ACFU, ACFG calculated for lag L .
- Ω is a diagonal matrix in ACFG with geometrically decaying weights with $p = 0.05$.
- Clustering algorithm - Complete linkage (hierarchical) and K-means (non-hierarchical).
- Calculate missclassification percentages.

Table: Misclassification Percentages in Hierarchical clustering for classification of AR(0.3) and AR(0.4) processes.

| n | EUCL | AR | L | ACFU | ACFG | FREQ | NP | LNP | KL | Delta | D |
|------|------|----|-----|------|------|------|----|-----|----|-------|----|
| 100 | 50 | 50 | 5 | 45 | 47 | LOW | 49 | 50 | 50 | 42 | 44 |
| | | | 8 | 46 | 47 | HIGH | 50 | 50 | 45 | | |
| | | | 10 | 48 | 47 | ALL | 50 | 50 | 46 | | |
| | | | 25 | 45 | 49 | | | | | | |
| 200 | 49 | 46 | 5 | 46 | 47 | LOW | 49 | 49 | 50 | 42 | 47 |
| | | | 8 | 45 | 50 | HIGH | 50 | 50 | 50 | | |
| | | | 10 | 47 | 49 | ALL | 50 | 50 | 44 | | |
| | | | 50 | 46 | 47 | | | | | | |
| | | | 100 | 50 | 44 | | | | | | |
| 500 | 45 | 50 | 5 | 34 | 36 | LOW | 50 | 49 | 43 | 35 | 34 |
| | | | 8 | 46 | 43 | HIGH | 50 | 50 | 50 | | |
| | | | 10 | 47 | 40 | ALL | 50 | 49 | 43 | | |
| | | | 50 | 37 | 39 | | | | | | |
| | | | 100 | 50 | 45 | | | | | | |
| | | | 250 | 50 | 38 | | | | | | |
| 1000 | 47 | 37 | 5 | 45 | 38 | LOW | 49 | 50 | 42 | 18 | 34 |
| | | | 8 | 44 | 39 | HIGH | 50 | 50 | 49 | | |
| | | | 50 | 38 | 42 | ALL | 50 | 50 | 49 | | |
| | | | 100 | 39 | 39 | | | | | | |
| | | | 250 | 49 | 41 | | | | | | |

Non-Hierarchical Clustering - K-Means

Table: Misclassification Percentages in Non-Hierarchical clustering (K-Means) for classification of AR(0.3) and AR(0.4) processes.

| n | EUCL | AR | L | ACFU | FREQ | NP | LNP | Delta | D |
|------|------|----|-----|------|------|----|-----|-------|----|
| 100 | 50 | 33 | 5 | 39 | LOW | 48 | 49 | 40 | 44 |
| | | | 8 | 41 | HIGH | 48 | 39 | | |
| | | | 10 | 41 | ALL | 48 | 49 | | |
| | | | 25 | 42 | | | | | |
| 200 | 50 | 24 | 5 | 34 | LOW | 48 | 50 | 36 | 42 |
| | | | 8 | 37 | HIGH | 49 | 35 | | |
| | | | 10 | 38 | ALL | 48 | 50 | | |
| | | | 50 | 41 | | | | | |
| 500 | 50 | 14 | 5 | 21 | LOW | 47 | 47 | 26 | 34 |
| | | | 8 | 23 | HIGH | 46 | 20 | | |
| | | | 10 | 25 | ALL | 47 | 46 | | |
| | | | 50 | 29 | | | | | |
| | | | 100 | 31 | | | | | |
| 1000 | 50 | 5 | 5 | 9 | LOW | 48 | 49 | 16 | 16 |
| | | | 8 | 9 | HIGH | 24 | 10 | | |
| | | | 50 | 9 | ALL | 20 | 10 | | |
| | | | 100 | 9 | | | | | |
| | | | 250 | 1 | | | | | |

Clustering

- Consider a particular movement.
- Corresponding to each movement, there are eye gaze time series of seven subjects under Watch condition and seven under imitate condition.
- Use Delta metric for classification.
- Use Complete linkage algorithm.
- 1, 2, 3, ..., 7 correspond to Watch condition.
- 8, 9, 10, ..., 14 correspond to Imitate condition.

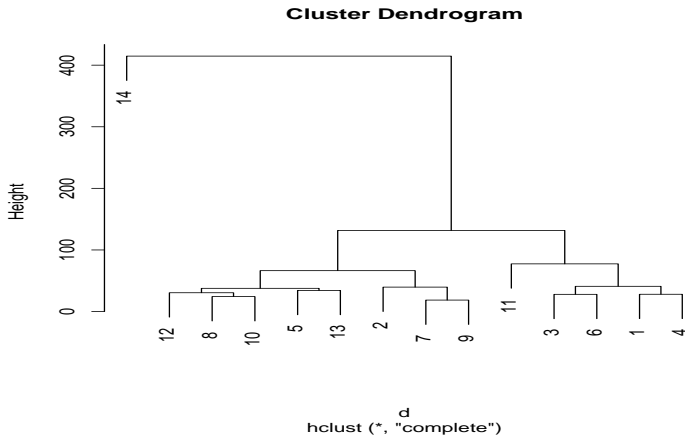


Figure: Classification of eye data for movement 1 using the Delta metric.

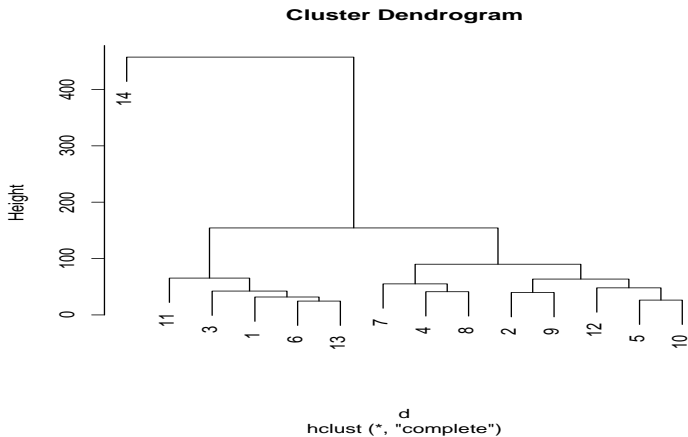


Figure: Classification of eye data for movement 4 using the Delta metric.

Summary

- Based on the ψ^2 distances, we conclude that in some cases the individuals (Subjects 1, 6 and 7) viewed the movements differently under the two conditions.
- Using the HOC measure in clustering, it can be said that Subject 7 viewed the movements differently under the two conditions.
- The results obtained from clustering do not in general give a clear picture.

References

- Kedem, B.(1994).Time Series Analysis by Higher Order Crossings. IEEE Press,Piscataway, N.J.
- Noy, L.(2009).Studying Eye Movements in Movement Imitation. PhD Thesis, Weizmann Institute of Science, Rehovot, Israel.
- Caiado, J., and Crato, N., and Pena, D.(2006).A periodogram-based metric for time series classification.Computational Statistics & Data Analysis,50,2668-2684.

Acknowledgements

We are grateful to Lior Noy (Harvard Medical School) for his advice and for providing the datasets.

THANK YOU

Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models

Guanhua Lu

Dr. Ben Kedem Symposium

July 31, 2009

Outline

- 1 Introduction
 - A Historic Review of Biased Sampling Models
 - Semiparametric Density Ratio Model
 - Estimation
- 2 Asymptotic Theory for $\hat{\theta}$ and \hat{G}
 - Asymptotic Theory for $\hat{\theta}$
 - Asymptotic Theory for $\hat{G}(t)$
- 3 Data Analysis
 - A Simulation Study for Estimation of Parameters
 - Goodness of Fit and Confidence Bands
 - An Application to Coronary Heart Disease Data

- Length-biased Sampling

- **Sampling Scheme:** Lengths of items are distributed according to the cdf G (**to be estimated**). The probability of selecting any particular item is proportional to its length, then the lengths of sampled items are distributed according to the *length-biased* cdf

$$F(y) = \frac{1}{\mu} \int_0^y x dG(x), \quad y \geq 0, \quad (1)$$

where $\mu = \int_0^\infty x dG(x) < \infty$.

- The Nonparametric maximum likelihood estimator (NPMLE) for G was obtained through empirical likelihood.

- Biased Sampling/Selection Bias:
 - Vardi[1985] and Gill, Vardi and Wellner[1988] considered the following biased sampling model by assuming general weight functions

$$F_j(y) = W_j(G)^{-1} \int_{-\infty}^y w_j(x) dG(x), \quad j = 1, \dots, m, \quad (2)$$

where the normalization constant $W_j(G) = \int_{-\infty}^{\infty} w_j(x) dG(x)$.

- NPMLE for G was obtained by combining information from several independent samples, and was shown asymptotically efficient.

- Case-control Study: (Prentice and Pyke[1979])
 - Suppose that m disease groups are defined. Let $D = j$ denote the development of the j th disease, and let $D = 0$ indicate the disease-free group. The probability that an individual with characteristics x develops disease $D = j$ can be specified in terms of a logistic regression model as

$$P(D = j|x) = \exp(\alpha_j + \beta_j'x) / \sum_{i=0}^m \exp(\alpha_i + \beta_i'x). \quad (3)$$

- Case-control Study: (Prentice and Pyke[1979])
 - Suppose that m disease groups are defined. Let $D = j$ denote the development of the j th disease, and let $D = 0$ indicate the disease-free group. The probability that an individual with characteristics x develops disease $D = j$ can be specified in terms of a logistic regression model as

$$P(D = j|x) = \exp(\alpha_j + \beta_j'x) / \sum_{i=0}^m \exp(\alpha_i + \beta_i'x). \quad (3)$$

- Formula (3) leads to the density ratio model

$$P(x | D = j) / P(x | D = 0) = \exp(\alpha_j^* + \beta_j'x).$$

- Case-control Study: (Prentice and Pyke[1979])
 - Suppose that m disease groups are defined. Let $D = j$ denote the development of the j th disease, and let $D = 0$ indicate the disease-free group. The probability that an individual with characteristics x develops disease $D = j$ can be specified in terms of a logistic regression model as

$$P(D = j|x) = \exp(\alpha_j + \beta_j'x) / \sum_{i=0}^m \exp(\alpha_i + \beta_i'x). \quad (3)$$

- Formula (3) leads to the density ratio model

$$P(x | D = j) / P(x | D = 0) = \exp(\alpha_j^* + \beta_j'x).$$

- Let $g_j(x)$ and $g(x)$ be the densities of the j th group the disease-free group respectively,

$$g_j(x) = \exp(\alpha_j^* + \beta_j'x)g(x), \quad j = 1, \dots, m. \quad (4)$$

- Recent Extensions
 - Gilbert, Lele & Vardi[1999] applied the selection biased model for assessing from vaccine trial data how efficacy of an HIV vaccine varies with characteristics (genotype and phenotype) of exposing virus.

- Recent Extensions

- Gilbert, Lele & Vardi[1999] applied the selection biased model for assessing from vaccine trial data how efficacy of an HIV vaccine varies with characteristics (genotype and phenotype) of exposing virus.
- Qin & Zhang[1997] and Zhang[2000] considered the two-sample case, and studied the asymptotic theory. A Kolmogorov-Smirnov type statistic was constructed for testing the goodness of fit of model (4).

- Recent Extensions

- Gilbert, Lele & Vardi[1999] applied the selection biased model for assessing from vaccine trial data how efficacy of an HIV vaccine varies with characteristics (genotype and phenotype) of exposing virus.
- Qin & Zhang[1997] and Zhang[2000] considered the two-sample case, and studied the asymptotic theory. A Kolmogorov-Smirnov type statistic was constructed for testing the goodness of fit of model (4).
- Fokianos et al.[2001] studied model (4) based on multiple samples for one-way layout with the distortion function x replaced by a more general form $h(x)$ and designed test for homogeneity among different samples.

- **Data Structure:** Suppose we have $m + 1$ independent samples,

$$\begin{aligned} X_0 &= (x_{01}, \dots, x_{0n_0})' \sim g(x) \\ X_1 &= (x_{11}, \dots, x_{1n_1})' \sim g_1(x) \\ &\vdots \\ X_m &= (x_{m1}, \dots, x_{mn_m})' \sim g_m(x). \end{aligned} \quad (5)$$

X_0 is referred as the reference sample with unknown distribution.

- **Data Structure:** Suppose we have $m + 1$ independent samples,

$$\begin{aligned} X_0 &= (x_{01}, \dots, x_{0n_0})' \sim g(x) \\ X_1 &= (x_{11}, \dots, x_{1n_1})' \sim g_1(x) \\ &\vdots \\ X_m &= (x_{m1}, \dots, x_{mn_m})' \sim g_m(x). \end{aligned} \quad (5)$$

X_0 is referred as the reference sample with unknown distribution.

- **Density Ratio Model:**

$$g_j(x) = \exp(\alpha_j + \beta_j' h(x)) g(x), \quad j = 1, \dots, m, \quad (6)$$

where α_j is a scalar, β_j is a $p \times 1$ vector, $h(x)$ is a $p \times 1$ predetermined distortion function.

- Example 1. (Normal Distribution)

$X_j \sim g_j(x) = N(\mu_j, \sigma_j^2)$, $j = 0, 1, \dots, m$. The weight function

$$\begin{aligned}w_j(x|\alpha_j, \beta_j) &= g_j(x)/g_0(x) = \exp(\alpha_j + \beta_j(x, x^2)'), \\ \alpha_j &= \log(\sigma_0/\sigma_j) + \mu_j^2/(2\sigma_j^2) - \mu_0^2/(2\sigma_0^2), \\ \beta_j &= (\mu_0/\sigma_0^2 - \mu_j/\sigma_j^2, 1/(2\sigma_0^2) - 1/(2\sigma_j^2))' .\end{aligned}$$

The distortion function

$$h(x) = (x, x^2)' .$$

$h(x)$ degenerates to x^2 if $\mu_j = 0$, and weight functions reduce to

$$w_j(x) = \exp(\alpha_j + \beta_j x^2) .$$

- Example 2. (Gamma Distribution)

$X_j \sim g_j(x) = \text{Gamma}(\alpha_{\gamma j}, \beta_\gamma)$, $j = 0, 1, \dots, m$ with a common scale parameter β_γ .

The weight function

$$\begin{aligned}w_j(x|\alpha_j, \beta_j) &= g_j(x)/g_0(x) = \exp(\alpha_j + \beta_j \log(x)), \\ \alpha_j &= \log \frac{\Gamma(\alpha_{\gamma 0})}{\Gamma(\alpha_{\gamma j})} + (\Gamma(\alpha_{\gamma j}) - \Gamma(\alpha_{\gamma 0})) \log \beta_\gamma, \\ \beta_j &= \alpha_{\gamma j} - \alpha_{\gamma 0}.\end{aligned}$$

The distortion function is

$$h(x) = \log(x).$$

- Example 3. (Log Normal Distribution)

$X_j \sim g_j(x) = \text{LN}(\alpha_j, \sigma^2)$, $j = 0, 1, \dots, m$ with a common σ^2 parameter.

The weight function

$$\begin{aligned}w_j(x|\alpha_j, \beta_j) &= g_j(x)/g_0(x) = \exp(\alpha_j + \beta_j \log(x)), \\(\alpha_j, \beta_j) &= \left(\frac{\mu_0^2 - \mu_j^2}{2\sigma^2}, \frac{\mu_0 - \mu_j}{\sigma^2}\right).\end{aligned}$$

The distortion function is

$$h(x) = \log(x).$$

Preparation for Estimation

- **Parameters:** $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)'$, $\boldsymbol{\beta} = (\beta'_1, \dots, \beta'_m)'$,
 $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$.

Preparation for Estimation

- **Parameters:** $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)'$, $\boldsymbol{\beta} = (\beta'_1, \dots, \beta'_m)'$,
 $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$.
- **Pooled Sample:** $\mathbf{t} = (t_1, \dots, t_n)' = (X'_0, X'_1, \dots, X'_m)'$, where
 $n = n_0 + n_1 + \dots + n_m$, the total sample size.

Preparation for Estimation

- **Parameters:** $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)'$, $\boldsymbol{\beta} = (\beta'_1, \dots, \beta'_m)'$,
 $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$.
- **Pooled Sample:** $\mathbf{t} = (t_1, \dots, t_n)' = (X'_0, X'_1, \dots, X'_m)'$, where
 $n = n_0 + n_1 + \dots + n_m$, the total sample size.
- **Empirical Likelihood:**

$$\mathcal{L}(\boldsymbol{\theta}, G) = \prod_{i=1}^n p_i \prod_{j=1}^m \prod_{i=1}^{n_j} \exp(\alpha_j + \beta'_j h(x_{ji})), \quad (7)$$

where $p_i = dG(t_i)$ is the mass at t_i .

A Profiling Procedure for Estimation

1. For fixed θ , maximize only the product term $\prod_{i=1}^n p_i$ from the empirical likelihood subject to the $m + 1$ constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i [w_j(t_i) - 1] = 0, \quad j = 1, \dots, m, \quad (8)$$

where the weights $w_j(t) = \exp(\alpha_j + \beta_j' h(t))$, $j = 1, \dots, m$.

A Profiling Procedure for Estimation

1. For fixed θ , maximize only the product term $\prod_{i=1}^n p_i$ from the empirical likelihood subject to the $m + 1$ constraints

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i [w_j(t_i) - 1] = 0, \quad j = 1, \dots, m, \quad (8)$$

where the weights $w_j(t) = \exp(\alpha_j + \beta_j' h(t))$, $j = 1, \dots, m$.

2. Solve p_i by applying the method of Lagrange multipliers,

$$p_i = (n + \sum_{j=1}^m \lambda_j [w_j(t_i) - 1])^{-1}, \quad i = 1, \dots, n. \quad (9)$$

3. Substituting p_i 's back into (7), the **profile log likelihood** as a function of θ only,

$$\begin{aligned} \ell(\theta) = & -n \log n_0 - \sum_{i=1}^n \log[1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)] \\ & + \sum_{j=1}^m \sum_{i=1}^{n_j} (\alpha_j + \beta_j' h(x_{ji})). \end{aligned} \quad (10)$$

3. Substituting p_i 's back into (7), the **profile log likelihood** as a function of θ only,

$$\begin{aligned} \ell(\theta) &= -n \log n_0 - \sum_{i=1}^n \log[1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)] \\ &\quad + \sum_{j=1}^m \sum_{i=1}^{n_j} (\alpha_j + \beta_j' h(x_{ji})). \end{aligned} \quad (10)$$

4. $\hat{\theta}$ obtained from the **score equations**,

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha_j} &= - \sum_{i=1}^n \frac{\rho_j w_j(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} + n_j = 0 \\ \frac{\partial \ell}{\partial \beta_j} &= - \sum_{i=1}^n \frac{\rho_j w_j(t_i) h(t_i)}{\sum_{k=0}^m \rho_k w_k(t_i)} + \sum_{i=1}^{n_j} h(x_{ji}) = 0. \end{aligned} \quad (11)$$

5. The solution of the score equations gives the MLE $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')$, and consequently by substitution also

$$\hat{p}_i = \frac{1}{n_0} \cdot \frac{1}{\sum_{j=0}^m \rho_j \exp(\hat{\alpha}_j + \hat{\beta}_j h(t_i))} \quad (12)$$

5. The solution of the score equations gives the MLE $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')$, and consequently by substitution also

$$\hat{p}_i = \frac{1}{n_0} \cdot \frac{1}{\sum_{j=0}^m \rho_j \exp(\hat{\alpha}_j + \hat{\beta}_j h(t_i))} \quad (12)$$

6. The estimator for G

$$\begin{aligned} \hat{G}(t) &= \sum_{i=1}^n \hat{p}_i I(t_i \leq t) \\ &= \frac{1}{n_0} \cdot \sum_{i=1}^n \frac{I(t_i \leq t)}{\sum_{j=0}^m \rho_j \exp(\hat{\alpha}_j + \hat{\beta}_j h(t_i))}. \end{aligned} \quad (13)$$

Outline

- 1 Introduction
 - A Historic Review of Biased Sampling Models
 - Semiparametric Density Ratio Model
 - Estimation
- 2 Asymptotic Theory for $\hat{\theta}$ and \hat{G}
 - Asymptotic Theory for $\hat{\theta}$
 - Asymptotic Theory for $\hat{G}(t)$
- 3 Data Analysis
 - A Simulation Study for Estimation of Parameters
 - Goodness of Fit and Confidence Bands
 - An Application to Coronary Heart Disease Data

Assumptions

- The first and second moments of $h(t)$ with respect to each sample are finite,

$$\int h(t)w_j(t)dG(t) < \infty, \quad \int h(t)h'(t)w_j(t)dG(t) < \infty, \quad (14)$$

$$j = 0, 1, \dots, m.$$

Assumptions

- The first and second moments of $h(t)$ with respect to each sample are finite,

$$\int h(t)w_j(t)dG(t) < \infty, \quad \int h(t)h'(t)w_j(t)dG(t) < \infty, \quad (14)$$

$$j = 0, 1, \dots, m.$$

- Sample fractions $\rho_j = n_j/n_0$, $j = 0, 1, \dots, m$ are finite and remain fixed as the total sample size

$$n = \sum_{j=0}^m n_j \rightarrow \infty.$$

- Taylor expansion of $\partial\ell(\hat{\theta})/\partial\theta$ at "true" θ_0 ,

$$0 = \frac{\partial\ell(\hat{\theta})}{\partial\theta} = \frac{\partial\ell(\theta_0)}{\partial\theta} + \frac{\partial^2\ell(\theta^*)}{\partial\theta^2}(\hat{\theta} - \theta_0), \quad (15)$$

where θ^* is between $\hat{\theta}$ and θ_0 .

- Taylor expansion of $\partial\ell(\hat{\theta})/\partial\theta$ at "true" θ_0 ,

$$0 = \frac{\partial\ell(\hat{\theta})}{\partial\theta} = \frac{\partial\ell(\theta_0)}{\partial\theta} + \frac{\partial^2\ell(\theta^*)}{\partial\theta^2}(\hat{\theta} - \theta_0), \quad (15)$$

where θ^* is between $\hat{\theta}$ and θ_0 .

- Provided that $S_n(\theta) = \partial^2\ell(\theta)/\partial\theta^2$ is positive-definite,

$$\sqrt{n}(\hat{\theta} - \theta_0) = - \left[\frac{1}{n} S_n(\theta^*) \right]^{-1} \frac{1}{\sqrt{n}} \frac{\partial\ell(\theta_0)}{\partial\theta}.$$

- Taylor expansion of $\partial\ell(\hat{\theta})/\partial\theta$ at "true" θ_0 ,

$$0 = \frac{\partial\ell(\hat{\theta})}{\partial\theta} = \frac{\partial\ell(\theta_0)}{\partial\theta} + \frac{\partial^2\ell(\theta^*)}{\partial\theta^2}(\hat{\theta} - \theta_0), \quad (15)$$

where θ^* is between $\hat{\theta}$ and θ_0 .

- Provided that $S_n(\theta) = \partial^2\ell(\theta)/\partial\theta^2$ is positive-definite,

$$\sqrt{n}(\hat{\theta} - \theta_0) = - \left[\frac{1}{n} S_n(\theta^*) \right]^{-1} \frac{1}{\sqrt{n}} \frac{\partial\ell(\theta_0)}{\partial\theta}.$$

- Under the density ratio model (6)

$$E_{\theta_0}(\partial\ell(\theta)/\partial\theta)^2 \neq -E_{\theta_0}S_n(\theta)$$

since contributions to the score statistic $\partial\ell(\theta_0)/\partial\theta$ from individual samples do not in general have mean zero.

Procedure to Develop the Asymptotic Theory for $\hat{\theta}$

1. Derive the structure of limit matrix $-\frac{1}{n}S_n \xrightarrow{a.s.} S$

Procedure to Develop the Asymptotic Theory for $\hat{\theta}$

1. Derive the structure of limit matrix $-\frac{1}{n}S_n \xrightarrow{a.s.} S$
2. Derive the covariance matrix of the score statistic $\frac{\partial \ell(\theta_0)}{\partial \theta}$.

Procedure to Develop the Asymptotic Theory for $\hat{\theta}$

1. Derive the structure of limit matrix $-\frac{1}{n}S_n \xrightarrow{a.s.} S$
2. Derive the covariance matrix of the score statistic $\frac{\partial \ell(\theta_0)}{\partial \theta}$.
3. Prove the strong consistency of $\hat{\theta}$ as an estimator of θ .

Procedure to Develop the Asymptotic Theory for $\hat{\theta}$

1. Derive the structure of limit matrix $-\frac{1}{n}S_n \xrightarrow{a.s.} S$
2. Derive the covariance matrix of the score statistic $\frac{\partial \ell(\theta_0)}{\partial \theta}$.
3. Prove the strong consistency of $\hat{\theta}$ as an estimator of θ .
4. Formulate the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta_0)$.

Notation

$$A_{jj'} = \int \frac{w_j(t)w_{j'}(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t), \quad B_{jj'} = \int \frac{w_j(t)w_{j'}(t)h(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)$$

$$C_{jj'} = \int \frac{w_j(t)w_{j'}(t)h(t)h'(t)}{\sum_{k=0}^m \rho_k w_k(t)} dG(t)$$

$$E_j = E(h(x_{ji})) = \int w_j(t)h(t)dG(t) \quad \bar{E}_j = \int w_j(t)h(t)h'(t)dG(t)$$

$$\begin{aligned} V_j &= \text{Var}(h(x_{ji})) \\ &= \int w_j h(t) h'(t) dG(t) - \int h(t) w_j dG(t) \int h'(t) w_j dG(t) \\ &= \bar{E}_j - E_j E_j', \end{aligned}$$

where $B_{jj'}$ and E_j are $p \times 1$ vectors, and $C_{jj'}$, \bar{E}_j and V_j are all $p \times p$ matrices.

Notation Cont'd

$$A = (A_{ij})_{m \times m}, \quad B = (B_{ij})_{mp \times m}, \quad C = (C_{ij})_{mp \times mp}$$

$$\rho = \text{diag}(\rho_1, \dots, \rho_m)_{m \times m}$$

$$E = \begin{pmatrix} E_1 & \cdots & \tilde{0} \\ \vdots & \ddots & \vdots \\ \tilde{0} & \cdots & E_m \end{pmatrix}_{mp \times m} \quad \bar{E} = \begin{pmatrix} \bar{E}_1 & \cdots & \hat{0} \\ \vdots & \ddots & \vdots \\ \hat{0} & \cdots & \bar{E}_m \end{pmatrix}_{mp \times mp}$$

$$\mathbf{1}_m = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{m \times m} \quad V = \begin{pmatrix} V_1 & \cdots & \hat{0} \\ \vdots & \ddots & \vdots \\ \hat{0} & \cdots & V_m \end{pmatrix}_{mp \times mp}$$

where $\tilde{0}$ is a $p \times 1$ vector of 0's, and $\hat{0}$ is a $p \times p$ matrix of 0's.

- Structure of the Limit Matrix

$$S = \frac{1}{\sum_{k=0}^m \rho_k} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}, \quad (16)$$

where

$$\begin{aligned} S_{11} &= \rho - \rho A \rho \\ S_{12} &= \rho E' - \rho B' (\rho \otimes I_p) \\ S_{21} &= S'_{12} = E \rho - (\rho \otimes I_p) B \rho \\ S_{22} &= (\rho \otimes I_p) \bar{E} - (\rho \otimes I_p) C (\rho \otimes I_p). \end{aligned} \quad (17)$$

I_p is the $p \times p$ identity matrix, \otimes denotes the *kroncker* product.

Structure of the Covariance Matrix

$$\Lambda = \text{Var}\left(\frac{1}{\sqrt{n}} \frac{\partial \ell}{\partial \theta}\right) = \frac{1}{\sum_{k=0}^m \rho_k} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix},$$

$$\Lambda_{11} = \rho A \rho - \rho A \rho A \rho - \rho \mathbf{1}_m \rho + \rho A \rho \mathbf{1}_m \rho + \rho \mathbf{1}_m \rho A \rho - \rho A \rho \mathbf{1}_m \rho A \rho$$

$$\Lambda_{12} = \rho A E'(\rho \otimes I_p) - \rho A \rho B'(\rho \otimes I_p) - \rho \mathbf{1}_m E'(\rho \otimes I_p) + \rho A \rho \mathbf{1}_m E'(\rho \otimes I_p) + \rho \mathbf{1}_m \rho B'(\rho \otimes I_p) - \rho A \rho \mathbf{1}_m B'(\rho \otimes I_p)$$

$$\Lambda_{21} = \Lambda'_{12} = (\rho \otimes I_p) E A \rho - (\rho \otimes I_p) B \rho A \rho - (\rho \otimes I_p) E \mathbf{1}_m \rho + (\rho \otimes I_p) E \mathbf{1}_m \rho A \rho + (\rho \otimes I_p) B \rho \mathbf{1}_m \rho - (\rho \otimes I_p) B \mathbf{1}_m \rho A \rho$$

$$\Lambda_{22} = -(\rho \otimes I_p) C(\rho \otimes I_p) - (\rho \otimes I_p) B \rho B'(\rho \otimes I_p) + (\rho \otimes I_p) B E'(\rho \otimes I_p) + (\rho \otimes I_p) E B'(\rho \otimes I_p) + (\rho \otimes I_p) V - (\rho \otimes I_p) E \mathbf{1}_m E'(\rho \otimes I_p) + (\rho \otimes I_p) B \rho \mathbf{1}_m E'(\rho \otimes I_p) + (\rho \otimes I_p) E \mathbf{1}_m \rho B'(\rho \otimes I_p) - (\rho \otimes I_p) B \rho \mathbf{1}_m \rho B'(\rho \otimes I_p).$$

Lemma 2.1 (Connection between S and Λ)

The limit matrix S and the covariance matrix Λ is connected by

$$\begin{aligned}\Lambda_{11} &= S_{11} - S_{11}(\mathbf{1}_m + \rho^{-1})S_{11} \\ \Lambda_{12} &= S_{12} - S_{11}(\mathbf{1}_m + \rho^{-1})S_{12} \\ \Lambda_{21} &= S_{21} - S_{21}(\mathbf{1}_m + \rho^{-1})S_{11} \\ \Lambda_{22} &= S_{22} - S_{21}(\mathbf{1}_m + \rho^{-1})S_{12}.\end{aligned}$$

Therefore, we have

$$\Sigma \stackrel{\text{def}}{=} S^{-1}\Lambda S^{-1} = S^{-1} - \sum_{k=0}^m \rho_k \begin{pmatrix} \mathbf{1}_m + \rho^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \quad (18)$$

Theorem 2.1 (Asymptotic Theory for $\hat{\theta}$)

Suppose that the density ratio model (6) and the Assumption (14) hold, and S is positive-definite, then

- (a) the solution $\hat{\theta}$ to the score equation system (11) is a strongly consistent estimator for θ_0 .
- (b) as $n \rightarrow \infty$,

$$\sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} \xrightarrow{d} N_{(p+1)m}(\mathbf{0}, \Sigma), \quad (19)$$

where $\Sigma = S^{-1} \Lambda S^{-1}$.

A Brief Review of Empirical Process

Let T_1, \dots, T_n be a real-valued random sample from a distribution function F .

Empirical measure

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{T_i}.$$

Empirical distribution

$$F_n(t) = P_n I_{[x < t]} = \frac{1}{n} \sum_{i=1}^n I_{[T_i < t]}.$$

Consider $F_n(t)$ as a random function $F_n(t, \omega)$.

Glivenko-Cantelli Theorem (SLLN)

$$|F_n - F|_\infty \xrightarrow{a.s.} 0$$

under the supremum norm $|F_n - F|_\infty = \sup_t |F_n(t) - F(t)|$.

Consider $F_n(t)$ as a random function $F_n(t, \omega)$.

Glivenko-Cantelli Theorem (SLLN)

$$|F_n - F|_\infty \xrightarrow{a.s.} 0$$

under the supremum norm $|F_n - F|_\infty = \sup_t |F_n(t) - F(t)|$.

Donsker Theorem (CLT)

$$\sqrt{n}(F_n - F) \xrightarrow{d} \mathbb{G}_F$$

in the Skorohod space $D[-\infty, \infty]$. The Brownian bridge \mathbb{G}_F has mean 0 and covariance

$$\mathbb{E}\mathbb{G}_F(t)\mathbb{G}_F(s) = F(t \wedge s) - F(t)F(s).$$

Define the empirical process of the reference sample:

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{[x_{0i} < t]}.$$

Define the empirical process of the reference sample:

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{[x_{0i} < t]}.$$

Procedure to Prove Weak Convergence:

- **Approximation:** $\hat{G}(t) \approx H_1(t) - H_2(t)$ uniformly in t .

Define the empirical process of the reference sample:

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{[x_{0i} < t]}.$$

Procedure to Prove Weak Convergence:

- **Approximation:** $\hat{G}(t) \approx H_1(t) - H_2(t)$ uniformly in t .
- **Decomposition:**

$$\begin{aligned} \sqrt{n}(\hat{G}(t) - G(t)) &= \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)) \\ &= \sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)) + \sqrt{n}(\tilde{G}(t) - G(t)). \end{aligned}$$

Define the empirical process of the reference sample:

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{[x_{0i} < t]}.$$

Procedure to Prove Weak Convergence:

- **Approximation:** $\hat{G}(t) \approx H_1(t) - H_2(t)$ uniformly in t .
- **Decomposition:**

$$\begin{aligned} \sqrt{n}(\hat{G}(t) - G(t)) &= \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)) \\ &= \sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)) + \sqrt{n}(\tilde{G}(t) - G(t)). \end{aligned}$$

- **Covariance Structure and Finite-dimensional Convergence.**

Define the empirical process of the reference sample:

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{[x_{0i} < t]}.$$

Procedure to Prove Weak Convergence:

- **Approximation:** $\hat{G}(t) \approx H_1(t) - H_2(t)$ uniformly in t .
- **Decomposition:**

$$\begin{aligned} \sqrt{n}(\hat{G}(t) - G(t)) &= \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)) \\ &= \sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)) + \sqrt{n}(\tilde{G}(t) - G(t)). \end{aligned}$$

- **Covariance Structure and Finite-dimensional Convergence.**
- **Tightness.**

An Approximation of $\hat{G}(t)$

- Define

$$H_1(t; \theta) = \frac{1}{n_0} \cdot \sum_{i=1}^n \frac{I(t_i \leq t)}{\sum_{k=0}^m \rho_k w_k(t_i; \alpha_k, \beta_k)}.$$

Assume $H_1(t) = H_1(t; \theta_0)$. $\hat{G}(t)$ is a realization of $H_1(t; \theta)$ at $\hat{\theta}$.

An Approximation of $\hat{G}(t)$

- Define

$$H_1(t; \theta) = \frac{1}{n_0} \cdot \sum_{i=1}^n \frac{I(t_i \leq t)}{\sum_{k=0}^m \rho_k w_k(t_i; \alpha_k, \beta_k)}.$$

Assume $H_1(t) = H_1(t; \theta_0)$. $\hat{G}(t)$ is a realization of $H_1(t; \theta)$ at $\hat{\theta}$.

- As $n \rightarrow \infty$, we have, uniformly in t

$$\begin{aligned} \frac{\partial H_1(t; \theta_0)}{\partial \alpha_j} &\xrightarrow{a.s.} -\rho_j A_j(t) = \rho_j \int \frac{w_j(y) I(y \leq t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \\ \frac{\partial H_1(t; \theta_0)}{\partial \beta_j} &\xrightarrow{a.s.} -\rho_j B_j(t) = \rho_j \int \frac{w_j(y) h(y) I(y \leq t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y). \end{aligned}$$

Denote $\bar{A}(t) = (A_1(t), \dots, A_m(t))'$, $\bar{B}(t) = (B'_1(t), \dots, B'_m(t))'$.

Lemma 3.1 (Approximation of $\hat{G}(t)$)

$\hat{G}(t)$ has an approximation uniformly in t ,

$$\hat{G}(t) = H_1(t) - H_2(t) + R_n(t),$$

where $H_1(t)$ is defined as before, and

$$H_2(t) = \frac{1}{n} \left(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p) \right) S^{-1} \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \end{pmatrix}, \quad (20)$$

and the remainder term $R_n(t)$ satisfies
 $\sup_{-\infty < t < \infty} |R_n(t)| = o_p(n^{-1/2})$.

Covariance Structure

Notice that $E_{\theta_0}[H_1(t) - \tilde{G}] = 0$, $E_{\theta_0}[H_2(t)] = 0$,

$$\begin{aligned} & \text{Cov} \left[\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)), \sqrt{n}(H_1(s) - \tilde{G}(s) - H_2(s)) \right] \\ &= n \left[E \left((H_1(t) - \tilde{G}(t))(H_1(s) - \tilde{G}(s)) \right) \right. \\ & \quad - E \left((H_1(t) - \tilde{G}(t))H_2(s) \right) \\ & \quad \left. - E \left(H_2(t)(H_1(s) - \tilde{G}(s)) \right) + E \left(H_2(t)H_2'(s) \right) \right]. \end{aligned}$$

Covariance Structure

$$\begin{aligned} & \text{Cov} \left[\sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)), \sqrt{n}(H_1(s) - \tilde{G}(s) - H_2(s)) \right] \\ &= \left(\sum_{k=0}^m \rho_k \right) \sum_{j=1}^m \rho_j A_j(t \wedge s) \\ & \quad - \left[\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p) \right] S^{-1} \begin{bmatrix} \boldsymbol{\rho} \bar{A}(s) \\ (\boldsymbol{\rho} \otimes I_p) \bar{B}(s) \end{bmatrix}. \quad (21) \end{aligned}$$

Lemma 3.2 (Finite-dimensional Convergence)

For any finite set (t_1, \dots, t_k) of points on the real line, let G_n denote $\sqrt{n}(H_1(t) - \tilde{G} - H_2(t))$, then we have

$$(G_n(t_1), \dots, G_n(t_k)) \xrightarrow{d} N_k(0, \Delta),$$

where N_k is a mean-zero k -dimensional multivariate normal distribution with covariance matrix Δ , of which the (i, j) th element is determined by (21).

Techniques for Proving Tightness

Donsker Class

Assume $\sup_{f \in \mathcal{F}} \|f(x) - Pf\| < \infty$, for every x . Under this condition, the empirical process $\{P_n f : f \in \mathcal{F}\}$ can be viewed as a map into $\ell^\infty(\mathcal{F})$, where $\ell^\infty(\mathcal{F})$ is a set of uniformly bounded real functions in \mathcal{F} .

\mathcal{F} is called **P-Donsker class** if the empirical processes based on P and indexed by \mathcal{F} satisfy

$$G_n = \sqrt{n}(P_n - P) \xrightarrow{d} G, \quad \text{in } \ell^\infty(\mathcal{F}),$$

where G is a tight Borel measurable element in $\ell^\infty(\mathcal{F})$.

Examples of Donsker Classes

- The collection of all indicator functions of the form $I_{(-\infty, t]}$ is a Donsker class.
- If \mathcal{F} is a Donsker class with $\|P\|_{\mathcal{F}} < \infty$ and g is a uniformly bounded, measurable function, then $\mathcal{F} \cdot g$ is a Donsker class.

Refer to Example 2.10.10 of Van der Vaart and Wellner (1996), p.192

- Any finite dimensional vector space \mathcal{F} of measurable functions is a Donsker class.

Refer to Lemma 2.6.15, Van der Vaart and Wellner (1996) Page 146

Tightness of $\sqrt{n}(H_1(t) - \tilde{G}(t))$

$$\begin{aligned}H_1(t) - \tilde{G}(t) &= \frac{1}{n_0} \sum_{j=1}^m \sum_{i=1}^{n_j} \frac{I(x_{ji} \leq t)}{\sum_{k=0}^m \rho_k w_k(x_{ji})} \\ &\quad - \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i} \leq t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})} \\ H_{1j}(t) &= \frac{1}{n_0} \sum_{i=1}^{n_j} \frac{I(x_{ji} \leq t)}{\sum_{k=0}^m \rho_k w_k(x_{ji})}, \\ H_{10}(t) &= \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\sum_{k=1}^m \rho_k w_k(x_{0i}) I(x_{0i} \leq t)}{\sum_{k=0}^m \rho_k w_k(x_{0i})}\end{aligned}$$

$$\mathbb{E}[H_{1j}(t)] = \rho_j A_j(t), \quad \mathbb{E}[H_{10}(t)] = \sum_{j=1}^m \rho_j A_j(t)$$

$$\begin{aligned} H_1(t) - \tilde{G}(t) &= \sum_{j=1}^m H_{1j}(t) - H_{10}(t) \\ &= \sum_{j=1}^m [H_{1j}(t) - \rho_j A_j(t)] - \left[H_{10}(t) - \sum_{j=1}^m \rho_j A_j(t) \right]. \end{aligned}$$

- Let \mathcal{F} be the collection of all indicator functions of the form $I_{(-\infty, t]}$.

- Let \mathcal{F} be the collection of all indicator functions of the form $I_{(-\infty, t]}$.
- P_{X_j} is the law of X_j , the j th sample, $j = 0, 1, \dots, m$.

- Let \mathcal{F} be the collection of all indicator functions of the form $I_{(-\infty, t]}$.
- P_{X_j} is the law of X_j , the j th sample, $j = 0, 1, \dots, m$.
- Define uniformly bounded functions

$$f_0(y) = \frac{\sum_{k=1}^m \rho_k w_k(y)}{\sum_{k=0}^m \rho_k w_k(y)}, \quad f_j(y) = \frac{\rho_j}{\sum_{k=0}^m \rho_k w_k(y)},$$

$j=1, \dots, m$.

- Let \mathcal{F} be the collection of all indicator functions of the form $I_{(-\infty, t]}$.
- P_{X_j} is the law of X_j , the j th sample, $j = 0, 1, \dots, m$.
- Define uniformly bounded functions

$$f_0(y) = \frac{\sum_{k=1}^m \rho_k w_k(y)}{\sum_{k=0}^m \rho_k w_k(y)}, \quad f_j(y) = \frac{\rho_j}{\sum_{k=0}^m \rho_k w_k(y)},$$

$j=1, \dots, m$.

- From the previous example, $\mathcal{F} \cdot f_i$, $i = 0, 1, \dots, m$, are P_{X_j} -Donsker classes, $j = 0, 1, \dots, m$

Let $P_{nj} = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{x_{ji}}$ be the empirical measure of the j th sample.
Then we have

$$\begin{aligned} & \sqrt{n_j}(P_{nj} - P_{X_j})(I_{(-\infty, t]}f_j) \\ &= \sqrt{n_j} \left[\frac{1}{n_j} \sum_{i=1}^{n_j} \frac{\rho_j I(x_{ji} \leq t)}{\sum_{k=0}^m \rho_k w_k(x_{ji})} - \rho_j \int \frac{w_j(y) I(y \leq t)}{\sum_{k=0}^m \rho_k w_k(y)} dG(y) \right] \\ &= (\rho_j / \sum_{k=0}^m \rho_k)^{1/2} \sqrt{n} (H_{1j} - \rho_j A_j(t)), \quad j = 1, \dots, m, \end{aligned}$$

and, similarly,

$$\sqrt{n_0}(P_{n0} - P_{X_0})(I_{(-\infty, t]}f_0) = (1 / \sum_{k=0}^m \rho_k)^{1/2} \sqrt{n} \left[H_{10}(t) - \sum_{j=1}^m \rho_j A_j(t) \right]$$

By Donsker's Theorem,

$$\sqrt{n_j}(P_{nj} - P_{X_j})(I_{(-\infty, t]}f_0) \xrightarrow{d} W_j \quad \text{in } D[-\infty, \infty],$$

where $j = 0, 1, \dots, m$, and W_j 's are mean-zero Gaussian processes. Therefore,

$$\sqrt{n}(H_{1j} - \rho_j A_j(t)), j = 0, 1, \dots, m, \text{ are tight in } D[-\infty, \infty].$$

From the decomposition of $\sqrt{n}(H_1(t) - \tilde{G}(t))$,

Lemma 3.3 Tightness of $\sqrt{n}(H_1(t) - \tilde{G}(t))$

$\sqrt{n}(H_1(t) - \tilde{G}(t))$ is tight in $D[-\infty, \infty]$.

Tightness of $\sqrt{n}H_2(t)$

- Define functions

$$U_l(y) = \frac{\rho_l w_l(y)}{\sum_{k=0}^m \rho_k w_k(y)}, \quad V_l(y) = \frac{\rho_l w_l(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)},$$

where $l = 0, 1, \dots, m$.

Tightness of $\sqrt{n}H_2(t)$

- Define functions

$$U_l(y) = \frac{\rho_l w_l(y)}{\sum_{k=0}^m \rho_k w_k(y)}, \quad V_l(y) = \frac{\rho_l w_l(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)},$$

where $l = 0, 1, \dots, m$.

- Define spaces

$$\mathcal{U} = \text{Span}\{U_k : k = 0, 1, \dots, m\}$$

$$\mathcal{V} = \text{Span}\{V_k : k = 0, 1, \dots, m\}$$

Tightness of $\sqrt{n}H_2(t)$

- Define functions

$$U_l(y) = \frac{\rho_l w_l(y)}{\sum_{k=0}^m \rho_k w_k(y)}, \quad V_l(y) = \frac{\rho_l w_l(y) h(y)}{\sum_{k=0}^m \rho_k w_k(y)},$$

where $l = 0, 1, \dots, m$.

- Define spaces

$$\mathcal{U} = \text{Span}\{U_k : k = 0, 1, \dots, m\}$$

$$\mathcal{V} = \text{Span}\{V_k : k = 0, 1, \dots, m\}$$

- \mathcal{U} and \mathcal{V} are both Donsker classes

Let $(a_1(t), \dots, a_m(t), b'_1(t), \dots, b'_m(t)) = (\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes I_p))S^{-1}$,
Then we have a decomposition for $\sqrt{n}H_2(t)$,

$$\begin{aligned} \sqrt{n}H_2(t) &= \frac{\sqrt{n}}{n} \left(a_1(t), \dots, a_m(t), b'_1(t), \dots, b'_m(t) \right) \frac{\partial \ell}{\partial \boldsymbol{\theta}} \\ &= \frac{n_0}{n} \left[\sum_{j=1}^m \frac{\sum_{k=0}^m \rho_k}{\rho_j} \sqrt{n_j} (P_{n_j} - P_{X_j}) \left(\sum_{\substack{l=0 \\ l \neq j}}^m \rho_l a_j(t) U_l \right) \right. \\ &\quad - \sum_{j=1}^m \sum_{\substack{l=0 \\ l \neq j}}^m \frac{\sum_{k=0}^m \rho_k}{\rho_l} \sqrt{n_l} (P_{n_l} - P_{X_l}) (\rho_l a_j(t) U_j) \\ &\quad + \sum_{j=1}^m \frac{\sum_{k=0}^m \rho_k}{\rho_j} \sqrt{n_j} (P_{n_j} - P_{X_j}) \left(\sum_{\substack{l=0 \\ l \neq j}}^m \rho_l b'_j(t) V_l \right) \\ &\quad \left. - \sum_{j=1}^m \sum_{\substack{l=0 \\ l \neq j}}^m \frac{\sum_{k=0}^m \rho_k}{\rho_l} \sqrt{n_l} (P_{n_l} - P_{X_l}) \left(\rho_l b'_j(t) V_j \right) \right]. \quad (22) \end{aligned}$$

Lemma 3.4 Tightness of $\sqrt{n}H_2(t)$

$\sqrt{n}H_2(t)$ is tight in $D[-\infty, \infty]$.

Lemma 3.4 Tightness of $\sqrt{n}H_2(t)$

$\sqrt{n}H_2(t)$ is tight in $D[-\infty, \infty]$.

Theorem 3.1 Weak Convergence of $\sqrt{n}(\hat{G} - \tilde{G})$

The process $\sqrt{n}(\hat{G} - \tilde{G})$ converges weakly to a zero-mean Gaussian process W with continuous sample paths in $D[-\infty, \infty]$, and the covariance matrix is determined by

$$\begin{aligned} EW(t)W(s) &= \left(\sum_{k=0}^m \rho_k \right) \sum_{j=1}^m \rho_j A_j(t \wedge s) \\ &\quad - \left(\bar{A}'(t)\rho, \bar{B}'(t)(\rho \otimes I_p) \right) S^{-1} \begin{pmatrix} \rho \bar{A}(s) \\ (\rho \otimes I_p) \bar{B}(s) \end{pmatrix}. \end{aligned}$$

Weak Convergence of $\sqrt{n}(\hat{G}(t) - G(t))$:

- Decomposition:

$$\begin{aligned}\sqrt{n}(\hat{G}(t) - G(t)) &= \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)) \\ &\approx \sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)) + \sqrt{n}(\tilde{G}(t) - G(t)).\end{aligned}$$

Weak Convergence of $\sqrt{n}(\hat{G}(t) - G(t))$:

- Decomposition:

$$\begin{aligned}\sqrt{n}(\hat{G}(t) - G(t)) &= \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)) \\ &\approx \sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)) + \sqrt{n}(\tilde{G}(t) - G(t)).\end{aligned}$$

- Variance-covariance structure and finite-dimensional convergence can be obtained similarly as for $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$.

Weak Convergence of $\sqrt{n}(\hat{G}(t) - G(t))$:

- Decomposition:

$$\begin{aligned}\sqrt{n}(\hat{G}(t) - G(t)) &= \sqrt{n}(\hat{G}(t) - \tilde{G}(t)) + \sqrt{n}(\tilde{G}(t) - G(t)) \\ &\approx \sqrt{n}(H_1(t) - \tilde{G}(t) - H_2(t)) + \sqrt{n}(\tilde{G}(t) - G(t)).\end{aligned}$$

- Variance-covariance structure and finite-dimensional convergence can be obtained similarly as for $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$.
- Tightness is followed from the fact that both $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ and $\sqrt{n}(\tilde{G}(t) - G(t))$ are tight.

Theorem 3.2 Weak Convergence of $\sqrt{n}(\hat{G} - G)$

The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges weakly to a zero-mean Gaussian process in $D[-\infty, \infty]$, with covariance matrix given by

$$\begin{aligned} \text{Cov}\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = & \\ & \left(\sum_{k=0}^m \rho_k \right) \left(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^m \rho_j A_j(t \wedge s) \right) \\ & + \left(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p) \right) S^{-1} \begin{pmatrix} \boldsymbol{\rho} \bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p) \bar{B}(t) \end{pmatrix}. \quad (23) \end{aligned}$$

Outline

- 1 Introduction
 - A Historic Review of Biased Sampling Models
 - Semiparametric Density Ratio Model
 - Estimation
- 2 Asymptotic Theory for $\hat{\theta}$ and \hat{G}
 - Asymptotic Theory for $\hat{\theta}$
 - Asymptotic Theory for $\hat{G}(t)$
- 3 Data Analysis
 - A Simulation Study for Estimation of Parameters
 - Goodness of Fit and Confidence Bands
 - An Application to Coronary Heart Disease Data

Simulation Study for Parameters

- Generate random samples $X_0 \sim N(0, 1)$, $X_1 \sim N(0, 2)$ and $X_2 \sim N(0, 4)$ with density functions $g(x)$, $g_1(x)$ and $g_2(x)$ respectively.

Simulation Study for Parameters

- Generate random samples $X_0 \sim N(0, 1)$, $X_1 \sim N(0, 2)$ and $X_2 \sim N(0, 4)$ with density functions $g(x)$, $g_1(x)$ and $g_2(x)$ respectively.
- Fit the following density ratio model:

$$\begin{aligned}g_1(x) &= g(x) \exp(\alpha_1 + \beta_1 x^2), \\g_2(x) &= g(x) \exp(\alpha_2 + \beta_2 x^2).\end{aligned}\tag{24}$$

Simulation Study for Parameters

- Generate random samples $X_0 \sim N(0, 1)$, $X_1 \sim N(0, 2)$ and $X_2 \sim N(0, 4)$ with density functions $g(x)$, $g_1(x)$ and $g_2(x)$ respectively.
- Fit the following density ratio model:

$$\begin{aligned}g_1(x) &= g(x) \exp(\alpha_1 + \beta_1 x^2), \\g_2(x) &= g(x) \exp(\alpha_2 + \beta_2 x^2).\end{aligned}\tag{24}$$

- True parameters
 $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (-0.34657, -0.69315, 0.25000, 0.37500)$.

Simulation Study for Parameters

- Generate random samples $X_0 \sim N(0, 1)$, $X_1 \sim N(0, 2)$ and $X_2 \sim N(0, 4)$ with density functions $g(x)$, $g_1(x)$ and $g_2(x)$ respectively.
- Fit the following density ratio model:

$$\begin{aligned}g_1(x) &= g(x) \exp(\alpha_1 + \beta_1 x^2), \\g_2(x) &= g(x) \exp(\alpha_2 + \beta_2 x^2).\end{aligned}\tag{24}$$

- True parameters
 $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (-0.34657, -0.69315, 0.25000, 0.37500)$.
- Calculate average bias and sample variance based on 1000 combined random samples.

Table: Bias and variance of parameter estimates.

| Sample Size | Bias($\hat{\alpha}_1$) | Bias($\hat{\beta}_1$) | Bias($\hat{\alpha}_2$) | Bias($\hat{\beta}_2$) |
|-----------------|--------------------------|-------------------------|--------------------------|-------------------------|
| (50, 50, 50) | -0.01663 | 0.02337 | -0.03752 | 0.03508 |
| (50, 50, 100) | -0.00022 | 0.00856 | -0.02041 | 0.02142 |
| (50, 100, 50) | -0.01865 | 0.02550 | -0.03797 | 0.03338 |
| (100, 50, 50) | -0.00326 | 0.00511 | -0.02925 | 0.01811 |
| (200, 200, 200) | -0.00017 | 0.00217 | -0.00303 | 0.00439 |
| Sample Size | Var($\hat{\alpha}_1$) | Var($\hat{\beta}_1$) | Var($\hat{\alpha}_2$) | Var($\hat{\beta}_2$) |
| (50, 50, 50) | 0.02425 | 0.01731 | 0.03362 | 0.01672 |
| (50, 50, 100) | 0.02085 | 0.01497 | 0.02276 | 0.01421 |
| (50, 100, 50) | 0.01961 | 0.01623 | 0.03168 | 0.01658 |
| (100, 50, 50) | 0.01773 | 0.00929 | 0.02674 | 0.00828 |
| (200, 200, 200) | 0.00611 | 0.00391 | 0.00837 | 0.00374 |

Table: 95% Confidence intervals for parameters.

| Sample Size | α_1 | α_2 |
|-----------------|----------------------|----------------------|
| (50, 50, 50) | (-0.64010, -0.05305) | (-1.04119, -0.34511) |
| (50, 50, 100) | (-0.63505, -0.05810) | (-0.98391, -0.40238) |
| (50, 100, 50) | (-0.60031, -0.09284) | (-1.02833, -0.35796) |
| (100, 50, 50) | (-0.60319, -0.08996) | (-1.00900, -0.37730) |
| (200, 200, 200) | (-0.49333, -0.19981) | (-0.86717, -0.51913) |
| Sample Size | β_1 | β_2 |
| (50, 50, 50) | (0.02029, 0.47971) | (0.14742, 0.60258) |
| (50, 50, 100) | (0.02349, 0.47651) | (0.15889, 0.59111) |
| (50, 100, 50) | (0.03402, 0.46598) | (0.15327, 0.59673) |
| (100, 50, 50) | (0.06846, 0.43154) | (0.19693, 0.55307) |
| (200, 200, 200) | (0.13515, 0.36485) | (0.26121, 0.48879) |

Statistic for Goodness-of-fit Test

- Define test statistic

$$\Delta_n(t) = \sqrt{n} |\hat{G} - \tilde{G}|, \quad \Delta_n = \sup_{-\infty \leq t \leq \infty} \Delta_n(t)$$

Statistic for Goodness-of-fit Test

- Define test statistic

$$\Delta_n(t) = \sqrt{n} |\hat{G} - \tilde{G}|, \quad \Delta_n = \sup_{-\infty \leq t \leq \infty} \Delta_n(t)$$

- Let w_α be the α -quantile of the distribution of $\sup_{-\infty \leq t \leq \infty} |W(t)|$. By weak convergence of $\sqrt{n}(\hat{G} - \tilde{G})$,

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\Delta_n \geq w_{1-\alpha}) &= \lim_{n \rightarrow \infty} P\left(\sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G} - \tilde{G}| \geq w_{1-\alpha}\right) \\ &= P\left(\sup_{-\infty \leq t \leq \infty} |W(t)| \geq w_{1-\alpha}\right) = \alpha. \end{aligned}$$

Statistic for Goodness-of-fit Test

- Define test statistic

$$\Delta_n(t) = \sqrt{n} |\hat{G} - \tilde{G}|, \quad \Delta_n = \sup_{-\infty \leq t \leq \infty} \Delta_n(t)$$

- Let w_α be the α -quantile of the distribution of $\sup_{-\infty \leq t \leq \infty} |W(t)|$. By weak convergence of $\sqrt{n}(\hat{G} - \tilde{G})$,

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\Delta_n \geq w_{1-\alpha}) &= \lim_{n \rightarrow \infty} P\left(\sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G} - \tilde{G}| \geq w_{1-\alpha}\right) \\ &= P\left(\sup_{-\infty \leq t \leq \infty} |W(t)| \geq w_{1-\alpha}\right) = \alpha. \end{aligned}$$

- We reject the density ratio model (6) at level α if

$$\Delta_n > w_{1-\alpha}.$$

Test Procedure Based on Bootstrap

1. Estimate $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ from (X_0, X_1, \dots, X_m) and calculate Δ_n .

Test Procedure Based on Bootstrap

1. Estimate $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ from (X_0, X_1, \dots, X_m) and calculate Δ_n .
2. Generate $X_0^*, X_1^*, \dots, X_m^*$ from $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ respectively.

Test Procedure Based on Bootstrap

1. Estimate $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ from (X_0, X_1, \dots, X_m) and calculate Δ_n .
2. Generate $X_0^*, X_1^*, \dots, X_m^*$ from $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ respectively.
3. Obtain the estimate \hat{G}^* for \hat{G} based on $(X_0^*, X_1^*, \dots, X_m^*)$ and empirical cdf \tilde{G}^* for X_0^* , then calculate

$$\Delta_n^* = \sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G}^* - \tilde{G}^*|.$$

Test Procedure Based on Bootstrap

1. Estimate $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ from (X_0, X_1, \dots, X_m) and calculate Δ_n .
2. Generate $X_0^*, X_1^*, \dots, X_m^*$ from $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ respectively.
3. Obtain the estimate \hat{G}^* for \hat{G} based on $(X_0^*, X_1^*, \dots, X_m^*)$ and empirical cdf \tilde{G}^* for X_0^* , then calculate

$$\Delta_n^* = \sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G}^* - \tilde{G}^*|.$$

!! $\sqrt{n}(\hat{G}^* - \tilde{G}^*) \xrightarrow{d} W$, W is also the limit process of $\sqrt{n}(\hat{G} - \tilde{G})$.

Test Procedure Based on Bootstrap

1. Estimate $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ from (X_0, X_1, \dots, X_m) and calculate Δ_n .
2. Generate $X_0^*, X_1^*, \dots, X_m^*$ from $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ respectively.
3. Obtain the estimate \hat{G}^* for \hat{G} based on $(X_0^*, X_1^*, \dots, X_m^*)$ and empirical cdf \tilde{G}^* for X_0^* , then calculate

$$\Delta_n^* = \sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G}^* - \tilde{G}^*|.$$

- !! $\sqrt{n}(\hat{G}^* - \tilde{G}^*) \xrightarrow{d} W$, W is also the limit process of $\sqrt{n}(\hat{G} - \tilde{G})$.
4. Repeat step 3 by bootstrapping from $(X_0^*, X_1^*, \dots, X_m^*)$.

Test Procedure Based on Bootstrap

1. Estimate $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ from (X_0, X_1, \dots, X_m) and calculate Δ_n .
2. Generate $X_0^*, X_1^*, \dots, X_m^*$ from $\hat{G}, \hat{G}_1, \dots, \hat{G}_m$ respectively.
3. Obtain the estimate \hat{G}^* for \hat{G} based on $(X_0^*, X_1^*, \dots, X_m^*)$ and empirical cdf \tilde{G}^* for X_0^* , then calculate

$$\Delta_n^* = \sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G}^* - \tilde{G}^*|.$$

- !! $\sqrt{n}(\hat{G}^* - \tilde{G}^*) \xrightarrow{d} W$, W is also the limit process of $\sqrt{n}(\hat{G} - \tilde{G})$.
4. Repeat step 3 by bootstrapping from $(X_0^*, X_1^*, \dots, X_m^*)$.
 5. Thus we can approximate the quantiles of Δ_n by those of Δ_n^* .

Example 1. (Correctly Specified Model)

- Simulated samples $X_0 \sim N(0, 1)$, $X_1 \sim N(0, 2)$ and $X_2 \sim N(0, 4)$ with sample sizes $(n_0, n_1, n_2) = (50, 60, 80)$.
- Fit the following model with distortion function $h(x) = x^2$

$$\begin{aligned}g_1(x) &= g(x) \exp(\alpha_1 + \beta_1 x^2), \\g_2(x) &= g(x) \exp(\alpha_2 + \beta_2 x^2).\end{aligned}$$

- $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (-0.576, -0.84, 0.436, 0.535)$. The value of the proposed test statistic $\Delta_n = 1.05$, and the observed p -value is $P(\Delta_n^* > 1.05) = 0.904$ based on 1000 bootstrap replications of Δ_n^* .

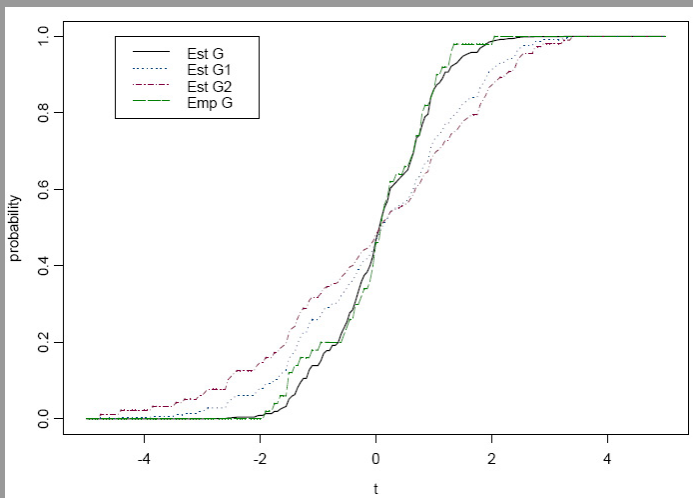


Figure: \hat{G} (solid curve), \hat{G}_1 (blue dotted curve), \hat{G}_2 (red dash-dot curve), empirical cdf \tilde{G} (green dashed curve).

Example 2. (Misspecified Model)

- Use the same data as in Example 1.
- Intentionally fit the following misspecified model with distortion function $h(x) = x$

$$\begin{aligned}g_1(x) &= g(x) \exp(\alpha_1 + \beta_1 x), \\g_2(x) &= g(x) \exp(\alpha_2 + \beta_2 x).\end{aligned}$$

- $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (-0.00072, -0.03, -0.0015, 0.032)$. The value of the proposed test statistic $\Delta_n = 2.31$, and the observed p -value is $P(\Delta_n^* > 2.31) = 0.007$ based on 1000 bootstrap replications of Δ_n^* .

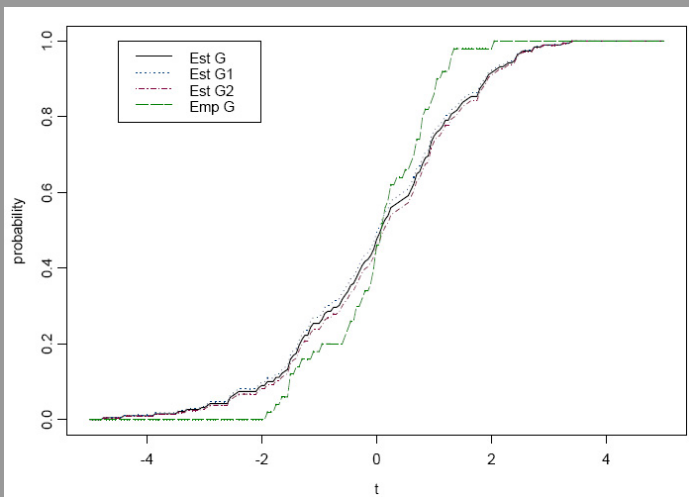


Figure: \hat{G} (solid curve), \hat{G}_1 (blue dotted curve), \hat{G}_2 (red dash-dot curve), empirical cdf \tilde{G} (green dashed curve).

Example 3. (With Relaxed Weight Functions)

- Use the same data as in Example 1.
- Fit the following model with distortion function $h(x) = (x, x^2)'$

$$g_1(x) = g(x) \exp(\alpha_1 + \gamma_1 x + \beta_1 x^2),$$

$$g_2(x) = g(x) \exp(\alpha_2 + \gamma_2 x + \beta_2 x^2).$$

- $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\beta}_1, \hat{\beta}_2) = (-0.562, -0.860, 0.023, 0.139, 0.427, 0.539)$. The value of the proposed test statistic $\Delta_n = 0.92$, and the observed p -value is $P(\Delta_n^* > 0.92) = 0.89$ based on 1000 bootstrap replications of Δ_n^* .

Confidence Bands (with Equal Widths)

- The limit of $\Delta_n = \sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G}(t) - G(t)|$ agrees with the limit of its bootstrap counterpart $\Delta_n^* = \sup_{-\infty \leq t \leq \infty} \sqrt{n} |\hat{G}^*(t) - \hat{G}(t)|$ almost surely.
- Approximate the quantiles of Δ_n by those of the distribution of Δ_n^* .
- For $\alpha \in (0, 1)$, let

$$w_{1-\alpha}^n = \inf\{y | P^*(\Delta_n^* \leq y) \geq 1 - \alpha\},$$

then a $1 - \alpha$ level bootstrap confidence band for G is given by

$$\left(\hat{G}(\cdot) - w_{1-\alpha}^n / \sqrt{n}, \hat{G}(\cdot) + w_{1-\alpha}^n / \sqrt{n} \right). \quad (25)$$

Pointwise Confidence Intervals

Let $V(t_i)$ be the estimated variance for $\hat{G}(t_i)$ at each point t_i from the covariance matrix (23). Then a $1 - \alpha$ level pointwise confidence interval for $\hat{G}(t_i)$ is given by

$$\left(\hat{G}(t_i) - z_{1-\alpha/2} \sqrt{V(t_i)}, \hat{G}(t_i) + z_{1-\alpha/2} \sqrt{V(t_i)} \right), \quad (26)$$

where $z_{1-\alpha/2}$ satisfies $P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$ with $Z \sim N(0, 1)$.

Bonferroni Simultaneous Confidence Intervals

The $1 - \alpha$ Bonferroni simultaneous confidence intervals given by

$$\left(\hat{G}(t_i) - t_{1-\alpha/2n}^{n-1} \sqrt{V(t_i)}, \hat{G}(t_i) + t_{1-\alpha/2n}^{n-1} \sqrt{V(t_i)} \right), \quad (27)$$

where $t_{1-\alpha/2n}^{n-1}$ is the $(1 - \frac{\alpha}{2n})$ percent cutoff point of the t_{n-1} distribution with degree of freedom $n - 1$.

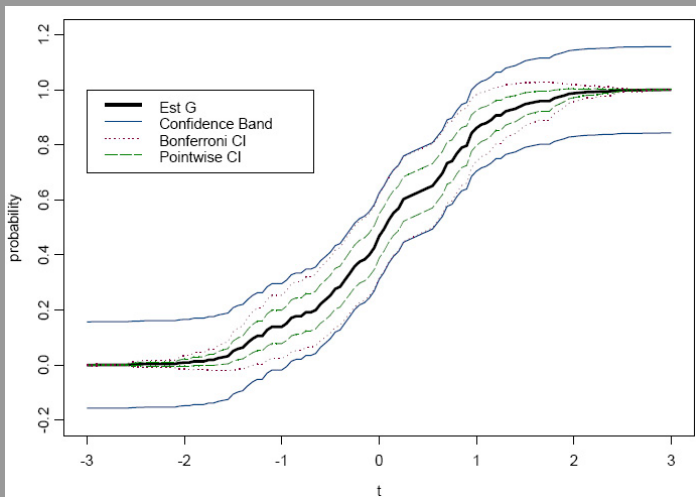


Figure: \hat{G} (black thick curve), 95% CB (blue curve), 95% Bonferroni simultaneous CI (red dotted curve), 95% pointwise CI (green dashed curve).

- **Data:** Hosmer & Lemeshow[1989, Ch. 1] used the logistic model to analyse the relationship between age and the status of coronary heart disease based on 100 subjects participating in a study. Let x denote age and $y = 1$ or 0 represent the presence or absence of coronary heart disease. $(x_i, y_i), i = 1, \dots, 100$.

- **Model:**

$$\frac{P(x|y = 1)}{P(x|y = 0)} = \exp(\alpha + \beta x).$$

- **Estimation:** $(\hat{\alpha}, \hat{\beta}) = (-5.0276, 0.1109)$. The value of the proposed test statistic $\Delta_n = 0.2199$, and the observed p -value is $P(\Delta_n^* > 0.2199) = 0.970$ based on 1000 bootstrap replications of Δ_n^* .

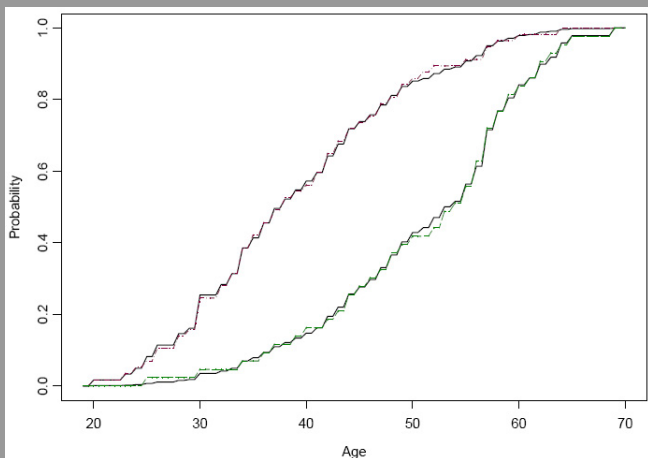


Figure: Solid and dash (red) curves on the upper left represent estimated cdf and empirical cdf for $P(x|y = 0)$, respectively; solid and dash (green) curves on the lower right represent estimated cdf and empirical cdf for $P(x|y = 1)$, respectively.

- Fit a More Complex Model:

$$\frac{P(x|y = 1)}{P(x|y = 0)} = \exp(\alpha + \beta x + \gamma x^2). \quad (28)$$

- Wald Test:** When $H_0 : \gamma = 0$ is true under model (28),

$$\sqrt{n}\hat{\gamma} \rightarrow N(0, \sigma_\gamma^2),$$

where σ_γ^2 is the asymptotic variance of $\hat{\gamma}$. Let $\hat{\sigma}_\gamma^2$ be the empirical version of σ_γ^2 on the basis of \hat{G} , we can use the statistic

$$T = \sqrt{n} \frac{\hat{\gamma}}{\hat{\sigma}_\gamma^2}$$

to test $H_0 : \gamma = 0$ is true under model (28).

- Estimation:** $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = (-3.9589, 0.0613, 0.0006)$. The Wald statistic $T = 0.2557$, and the observed p -value is 0.798 . This suggests to accept $H_0 : \gamma = 0$.

- **Odds-linear Model:**






$$\frac{P(y = 1|x)}{1 - P(y = 1|x)} = \alpha_0 + \beta_0 x. \quad (29)$$






This model is equivalent to







$$\frac{P(x|y = 1)}{P(x|y = 0)} = \exp\{\alpha + r(x; \beta)\}, \quad (30)$$

where $r(x; \beta) = \log(1 + (\beta_0/\alpha_0)x)$. The asymptotic results can be easily extended to model (30).

- **Estimation:** $(\hat{\alpha}, \hat{\beta}) = (-7.62, 47.47)$. The value of the proposed test statistic $\Delta_n = 1.55$, and the observed p -value is $P(\Delta_n^* > 1.55) = 0.005$ based on 1000 bootstrap replications of Δ_n^* .

-  Billingsley, P. (1999). *Convergence of Probability Measures*, New York: Wiley.
-  Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems, *Annals of Mathematical Statistics*, Vol. 23, 277-281.
-  Ethier, S. N. and Kurtz, T. G. (1985). *Markov Processes: Characterization and Convergence*, New York: Wiley.
-  Fokianos, K., Kedem, B., Qin, J., and Short, D.A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, Vol. 43, 56-65.
-  Gilbert, P.B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*, Vol. 28, 151-194.

-  Gilbert, P.B., Lele, S.R. and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, Vol. 86, 27-43.
-  Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distribution in biased sampling models. *Annals of Statistics*, Vol. 16, 1069-1112.
-  Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, Vol. 66, 403-411.
-  Qin, J. (1993). Empirical likelihood in biased sample problems. *Annals of Statistics*, Vol. 21, 1182-1196.
-  Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, Vol. 85 , 619-630

-  Qin, J. and Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, Vol. 22, 300-325.
-  Qin, J. and Zhang, B. (1997). A Goodness of Fit Test for Logistic Regression Models Based on Case-Control Data. *Biometrika*, Vol 84, 609-618.
-  Van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press.
-  Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*, New York: Springer.
-  Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, Vol. 10, 616-20.
-  Vardi, Y. (1985). Empirical distribution in selection bias models. *Annals of Statistics*, Vol. 13, 178-203.



Wellner, J. A. (1992). Empirical Processes in Action: A Review. *International statistical Review*, Vol 60, 247-269.



Zhang, B. (2000). A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica*, 10, 839-865.

Thank You!

Happy Brithday to Dr.
Kedem!