A Bayesian Method for General Change Point Problems

Xiaofei (Susan) Wang

Lecturer in Statistics
Department of Mathematics and Statistics
Amherst College

November 5, 2015

TERRAS IRRADIENT

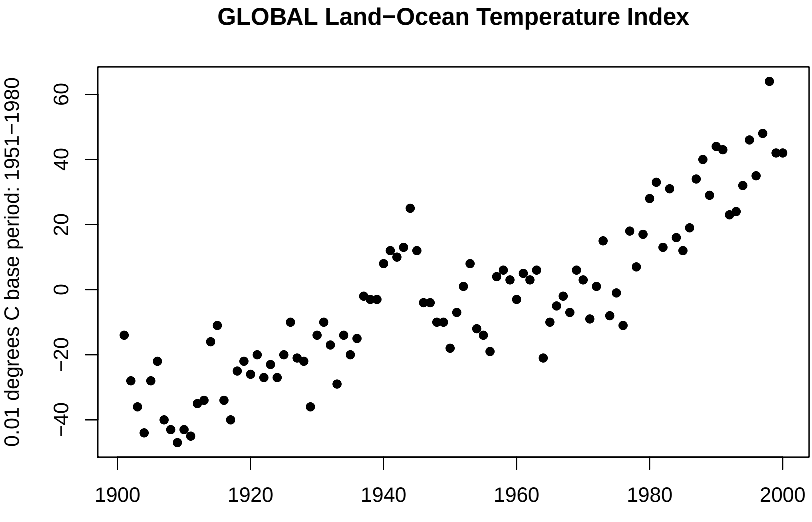
SIGILL. COLL. AMHERST. MASS. NOV. ANG. MDCCCXXV

Outline

Introduction to Change Points (Serial Data)

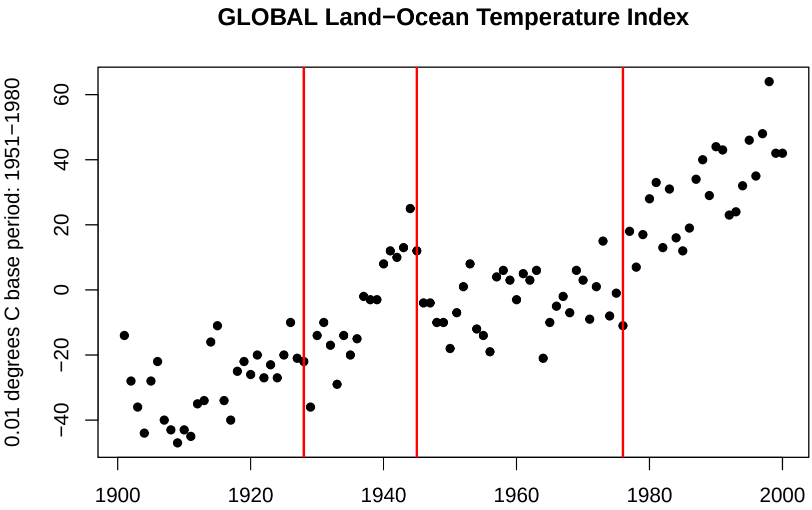
Example: Global Warming

What model is appropriate for these data?



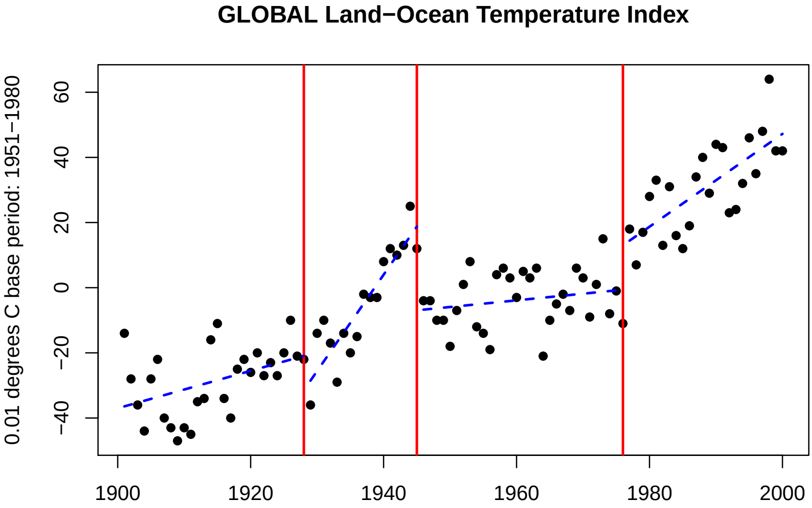
Example: Global Warming

What model is appropriate for these data?



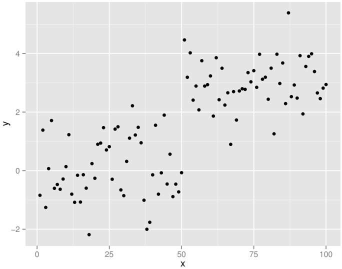
Example: Global Warming

What model is appropriate for these data?



What are change points?

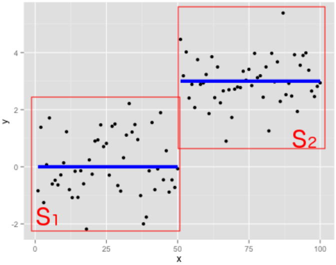
Change points partition observations into blocks. Within each block, observations share a common distribution.



- Observations: $y_1, \ldots, y_{100}$

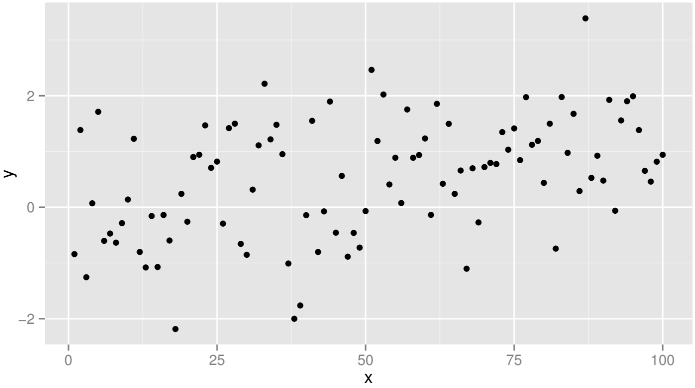
What are change points?

Change points partition observations into blocks. Within each block, observations share a common distribution.

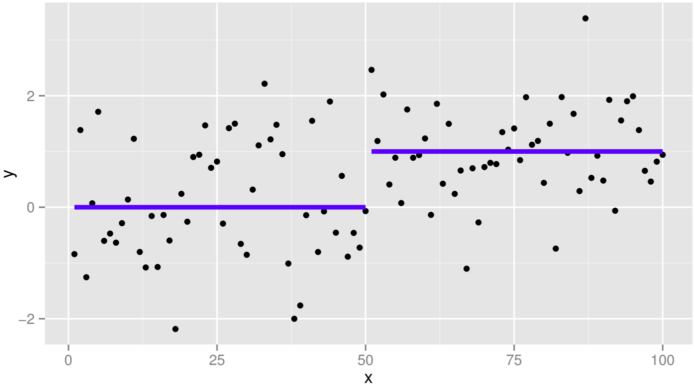


- Observations: $y_1, \ldots, y_{100}$
- Partition: $\rho = (S_1, S_2)$
  $S_1 = \{1, \ldots, 50\}$
  $S_2 = \{51, \ldots, 100\}$
- $y_{i:i \in S_1} \sim N(0, 1)$
  $y_{i:i \in S_2} \sim N(3, 1)$
- Change point at location 50.

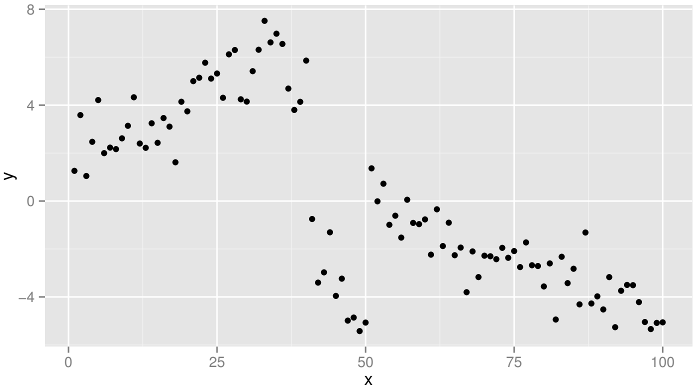
Example 1



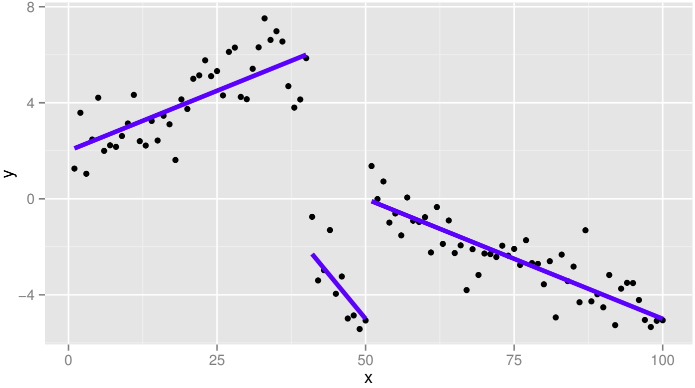
Example 1



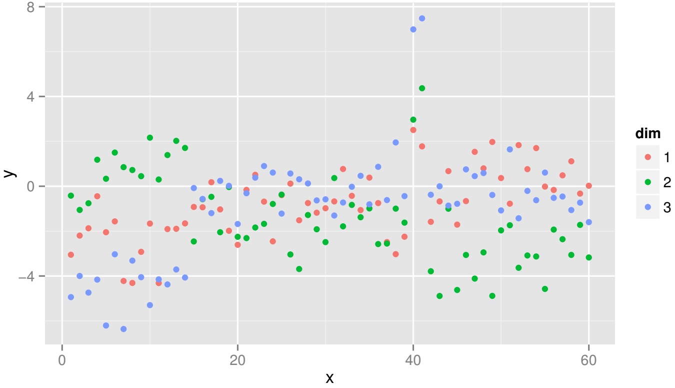
Example 2



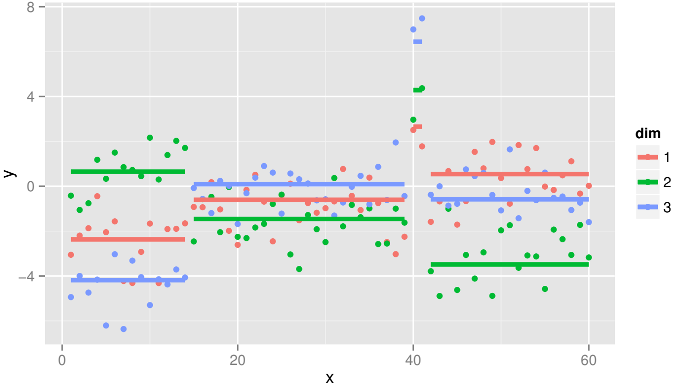
Example 2



Example 3



Example 3



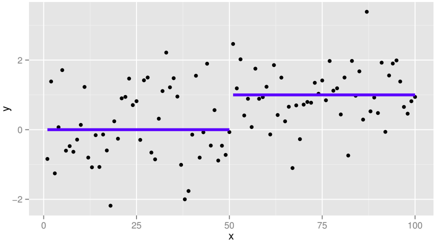
Related Work

Related Work

- Barry and Hartigan (1993, 1994), Erdman and Emerson (2007, 2008) - univariate, Bayesian (`library(bcp)`)
- Bai and Perron (2003), Zeileis, et al. (2001) - regression (`library(strucchange)`)
- Muggeo (2003) - regression (`library(segmented)`)
- Olshen, et al. (2004) - univariate, using circular binary segmentation
- Fearnhead (2005), Loschi, et al. (2010) - regression, Bayesian
- Killick and Eckley (2011) - univariate mean/variance (`library(changepoint)`)
- Matteson and James (2013) - multivariate (`library(ecp)`)
- ... and others

Classical Bayesian Change Point Analysis

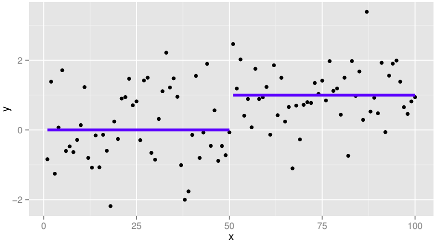
Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma)$ (e.g. Example 1)



- No need to specify number of blocks

Classical Bayesian Change Point Analysis

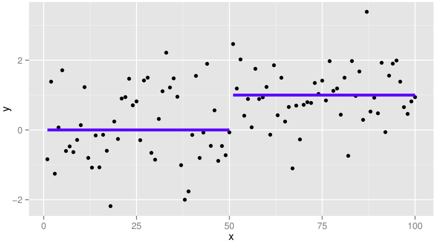
Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma)$ (e.g. Example 1)



- No need to specify number of blocks
- Models the partition ρ as a random variable

Classical Bayesian Change Point Analysis

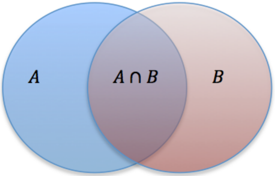
Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma)$ (e.g. Example 1)



- No need to specify number of blocks
- Models the partition ρ as a random variable
- Estimates means and probability of change point occurring at each location

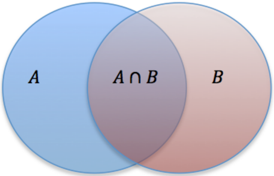
Bayes' Rule

Two events A and B:



Bayes' Rule

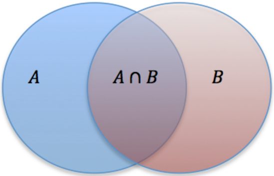
Two events A and B:



$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Bayes' Rule

Two events A and B:

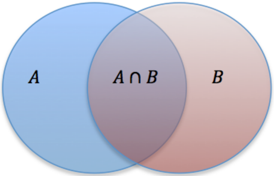


$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(A|B)P(B)$$

Bayes' Rule

Two events A and B:



$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

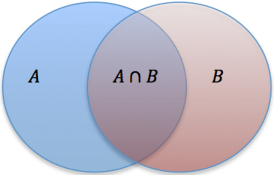
$$P(A \cap B) = P(A|B)P(B)$$

Combining these, we get

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayes' Rule

Two events A and B:



$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(A|B)P(B)$$

Combining these, we get

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Alternate form:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

More Generally

Prior: $\pi(\theta)$

Likelihood: $f(x|\theta)$

More Generally

Prior: $\pi(\theta)$

Likelihood: $f(x|\theta)$

Posterior:

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}$$

More Generally

Prior: $\pi(\theta)$

Likelihood: $f(x|\theta)$

Posterior:

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Classical Bayesian Change Point Analysis

Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma)$ (e.g. Example 1)

Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma)$ (e.g. Example 1)

Partition $\rho = (S_1, \ldots, S_b)$

Prior on the partition:

$$\pi(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b}dp$$

Classical Bayesian Change Point Analysis

Barry and Hartigan (1993): Univariate serial observations $y_i \sim N(\mu_i, \sigma)$ (e.g. Example 1)

Partition $\rho = (S_1, \ldots, S_b)$

Prior on the partition:

$$\pi(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b} dp$$

Likelihood:

$$y_{i:i\in S} | \theta_S, \sigma^2 \sim N(\theta_S, \sigma^2)$$

$$\theta_S | \mu_0, \sigma_0^2 \sim N\left(\mu_0, \frac{\sigma_0^2}{n_S}\right)$$

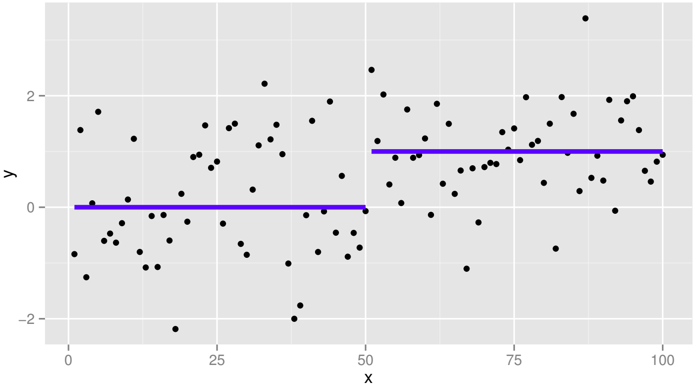
$$\mu_0 \sim U(-\infty, \infty)$$

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad \sigma^2 \in (0, \infty)$$

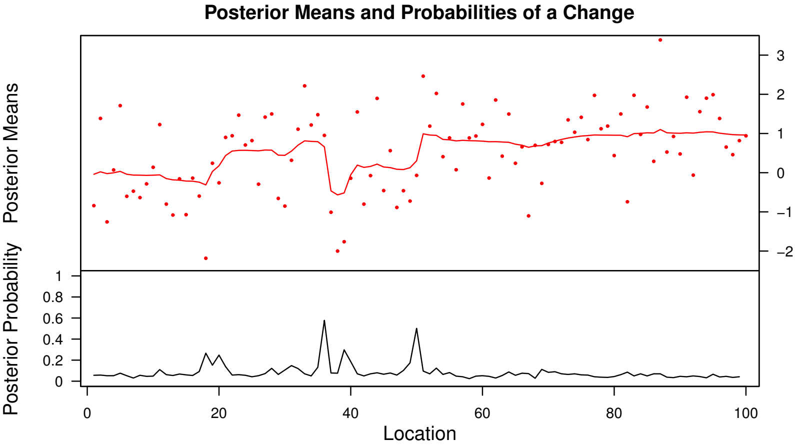
$$\pi(w) = \frac{1}{w_0'} \quad w \in (0, w_0')$$

where $w = \sigma^2/(\sigma_0^2 + \sigma^2)$

Example 1 (Revisited)



Example 1 (Revisited)



Generalized Bayesian Change Point Analysis (Serial Data)

Generalized Bayesian Change Point Analysis (BCP)

Features of Wang and Emerson (2015):

- multiple observations may be recorded at each location
- allows regressions within blocks, if predicting variables are available (e.g. Example 2)
  - includes multivariate/univariate normal observations (e.g. Examples 1 and 3) as special cases

Generalized Bayesian Change Point Analysis (BCP)

Features of Wang and Emerson (2015):

- multiple observations may be recorded at each location
- allows regressions within blocks, if predicting variables are available (e.g. Example 2)
  - includes multivariate/univariate normal observations (e.g. Examples 1 and 3) as special cases

Allows the data to reside on nodes of a connected graph (more on this later!)

Generalized BCP: Serial Data

Wang and Emerson (2015) - setup for univariate data

- Observations: $\{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{n}$
- $\boldsymbol{y_i}$ is $m_i \times 1$ (m_i independent observations at location i)
- $\boldsymbol{x_i}$ is $m_i \times k$ (k predictors)

Generalized BCP: Serial Data

Wang and Emerson (2015) - setup for univariate data

- Observations: $\{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{n}$
- $\boldsymbol{y_i}$ is $m_i \times 1$ (m_i independent observations at location i)
- $\boldsymbol{x_i}$ is $m_i \times k$ (k predictors)

Prior on the partition:

$$\pi(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b}\,dp$$

Likelihood:

$$\boldsymbol{y_{i:i\in S}} | \boldsymbol{x}, \beta_S, \sigma^2 \sim N_{m_i}(\boldsymbol{\widetilde{x}_i^S} \boldsymbol{\beta_S}, \sigma^2 \boldsymbol{I})$$

Prior on the intercept:

$$\beta_{S0} | \mu_0, \sigma_0^2 \sim N\left(\mu_0, \frac{\sigma_0^2}{n_S}\right)$$

Prior on other coefficients:

$$(\beta_{Sj} | \tau_S = 0) = 0 \quad \text{w.p. } 1$$

$$(\beta_{Sj} | \tau_S = 1, \sigma_j^2) \sim N\left(0, \frac{\sigma_j^2}{\sum_{i\in S}(x_{ij} - \bar{x}_{\cdot j}^S)^2}\right)$$

Generalized BCP: Serial Data

Wang and Emerson (2014): other priors

$$
\begin{aligned}
P(\tau_S = 0) &= \frac{d}{n_S + d}\mathbb{1}\{n_S \geq 2k\} + \mathbb{1}\{n_S < 2k\} \\
P(\tau_S = 1) &= \frac{n_S}{n_S + d}\mathbb{1}\{n_S \geq 2k\} \\
\mu_0 &\sim U(-\infty, \infty) \\
\pi(\sigma^2) &\propto \frac{1}{\sigma^2} \quad \sigma^2 \in (0, \infty) \\
\pi(w_j) &= \frac{1}{w_j} \quad w \in (0, w_j')
\end{aligned}
$$

where $w_j = \sigma^2/(\sigma_j^2 + \sigma^2)$

Bayesian Inference

We are interested in:

- $f(\rho|\boldsymbol{y}, \boldsymbol{x})$
- Posterior means for each location i

Bayesian Inference

We are interested in:

- $f(\rho|\boldsymbol{y}, \boldsymbol{x})$
- Posterior means for each location i (this depends on $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$)

Bayesian Inference

We are interested in:

- $f(\rho|\boldsymbol{y}, \boldsymbol{x})$
- Posterior means for each location i (this depends on $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$)

But... there is no closed form for the posterior $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$.

Bayesian Inference

We are interested in:

- $f(\rho|\boldsymbol{y}, \boldsymbol{x})$
- Posterior means for each location i (this depends on $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$)

But... there is no closed form for the posterior $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$.
So we use a form of Gibbs sampling:

Bayesian Inference

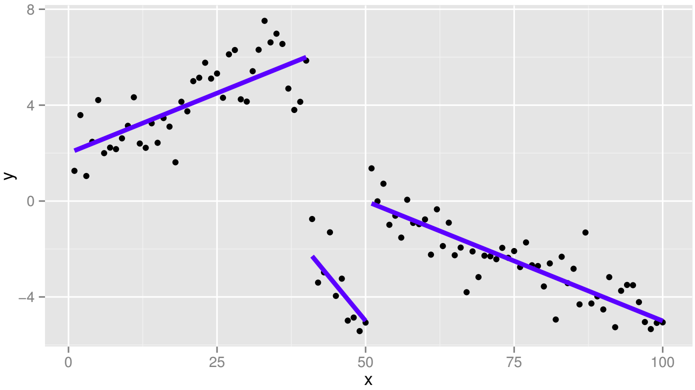
We are interested in:

- $f(\rho|\boldsymbol{y}, \boldsymbol{x})$
- Posterior means for each location i (this depends on $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$)

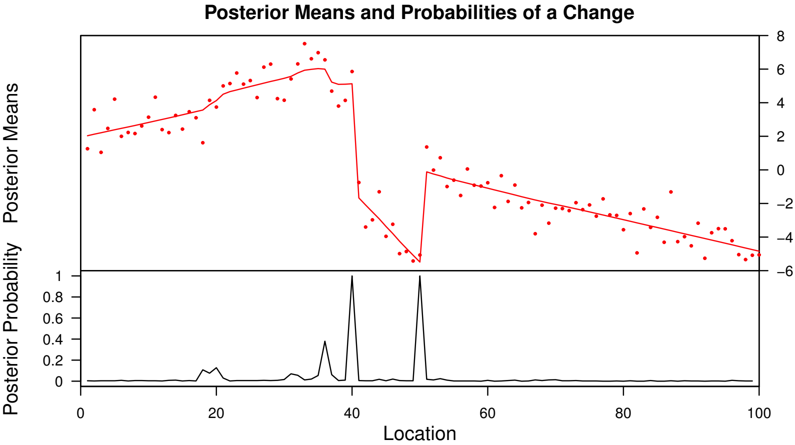
But... there is no closed form for the posterior $f(\rho, \tau, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})$.
So we use a form of Gibbs sampling:

- Start with any arbitrary $(\rho^{(0)}, \tau^{(0)}, \boldsymbol{w}^{\boldsymbol{(0)}})$
- For $t = 1, \ldots, T$:
  1. Sample $(\rho^{(t)}, \tau^{(t)})$ from $f(\rho, \tau|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}^{\boldsymbol{(t-1)}})$
  2. Sample $\boldsymbol{w}^{\boldsymbol{(t)}}$ from $f(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x}, \rho^{(t)}, \tau^{(t)})$

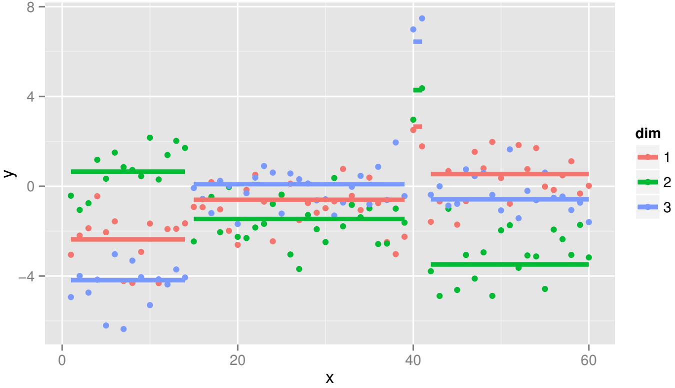
Example 2 (Revisited)



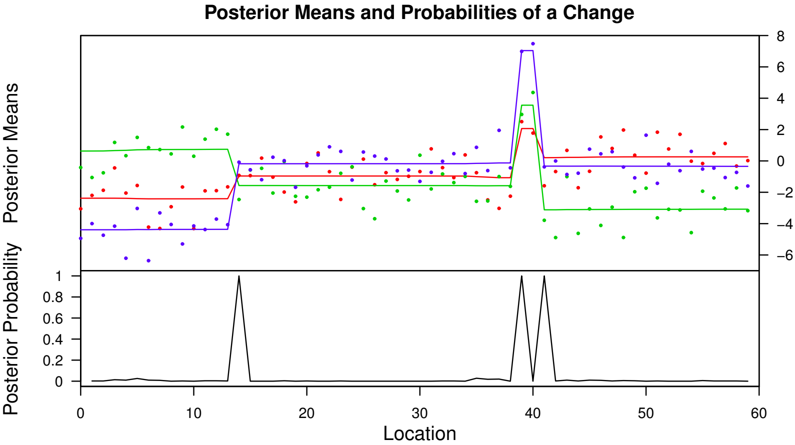
Example 2 (Revisited)



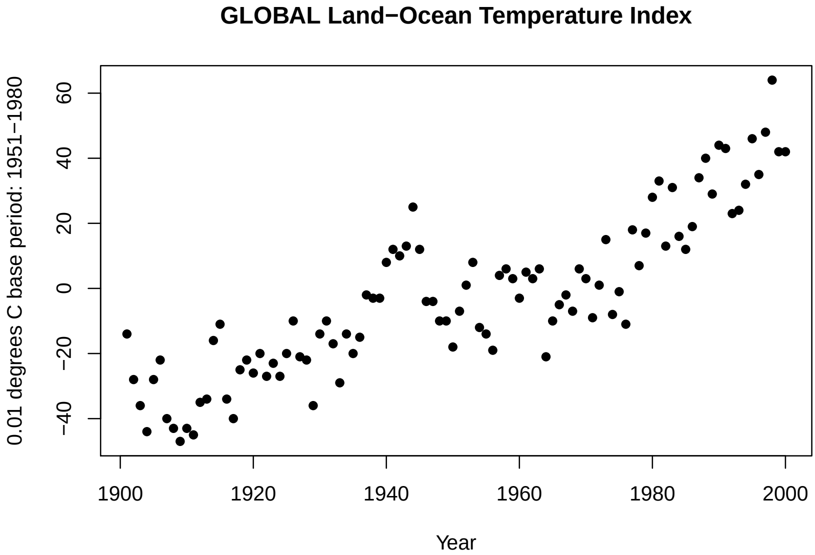
Example 3 (Revisited)



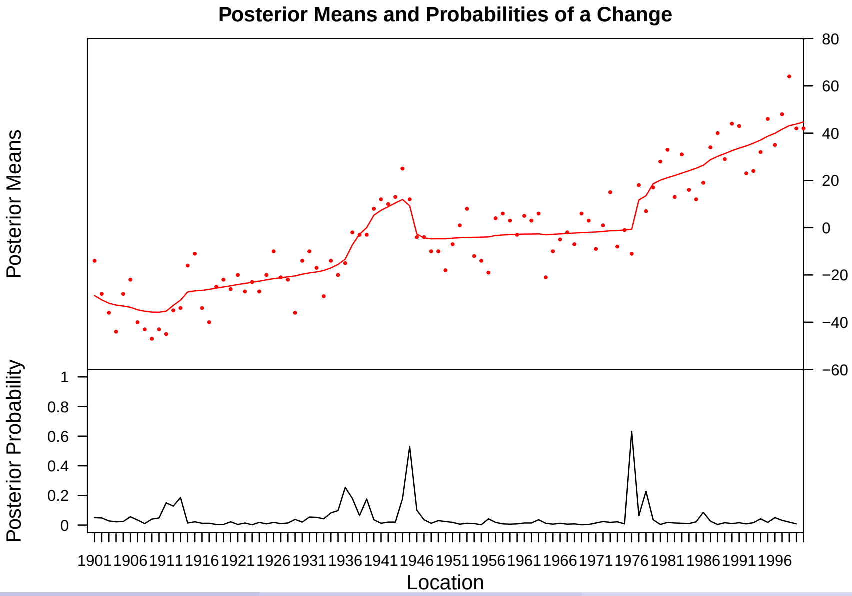
Example 3 (Revisited)



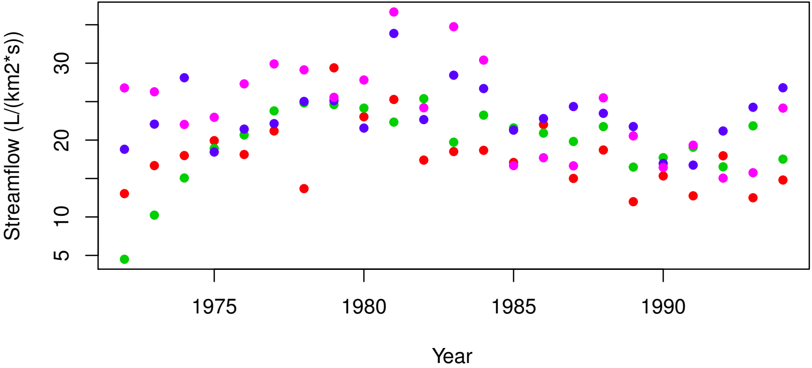
Example: Global Warming



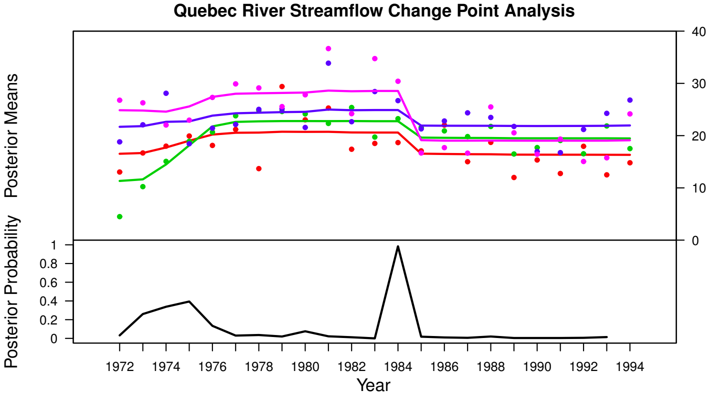
Example: Global Warming



Example: Quebec Rivers (Perreault et al. (2000))



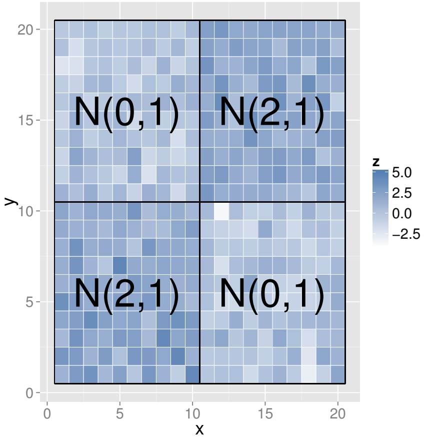
Example: Quebec Rivers (BCP Result)



Generalized Bayesian Change Point Analysis (General Graphs)

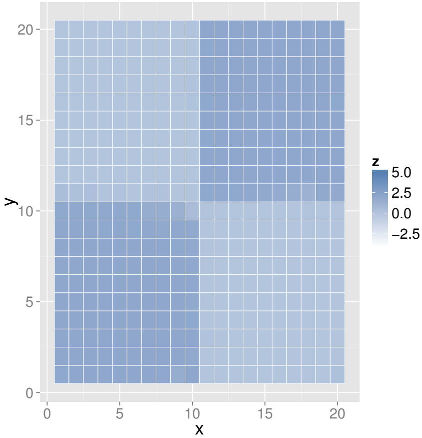
Change Points on a Grid Graph

Simulated Data



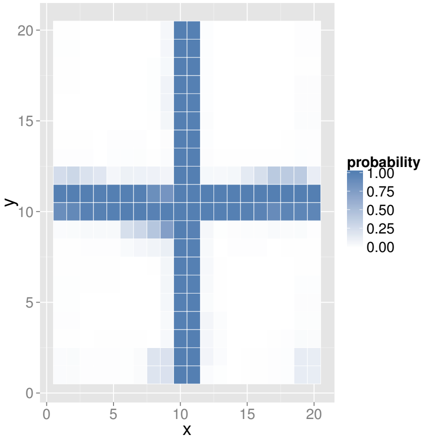
Change Points on a Grid Graph

Posterior Means



Change Points on a Grid Graph

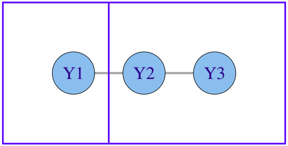
Posterior Boundary Probabilities



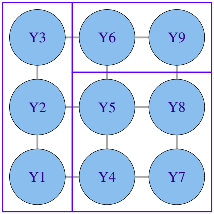
Change Points on a Grid Graph

What is a change point on a grid?

Before: (serial data $\Rightarrow$ path graph)



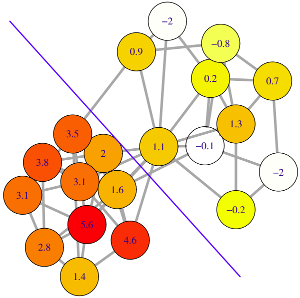
Now: (grid graph)



More Generally

Simulated Data

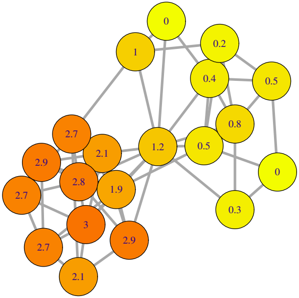
- Numbers and colors reflect **observed values**.
- 2 blocks:
    - $N(3,1)$
    - $N(0,1)$



More Generally

Posterior Means

- Numbers and colors reflect **posterior means**.



Prior on Partition

Originally (path graph):

$$f(\rho) \propto \int_0^{p_0} p^{b-1}(1-p)^{n-b}dp$$

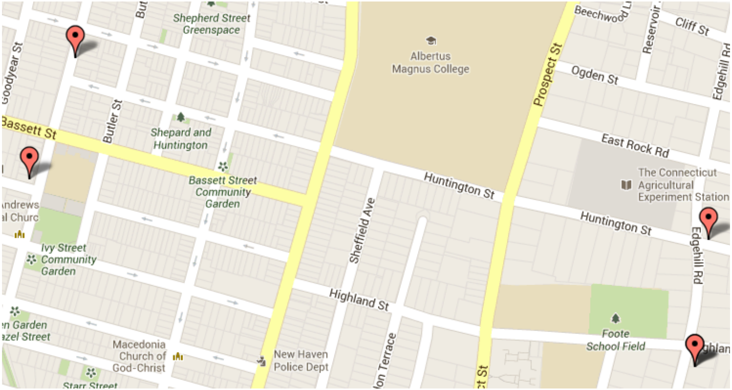
New (general graph):

$$f(\rho) \propto \alpha^{l(\rho)}$$

where $0 < \alpha < 1$ and $l(\rho)$ is a measure of boundary length

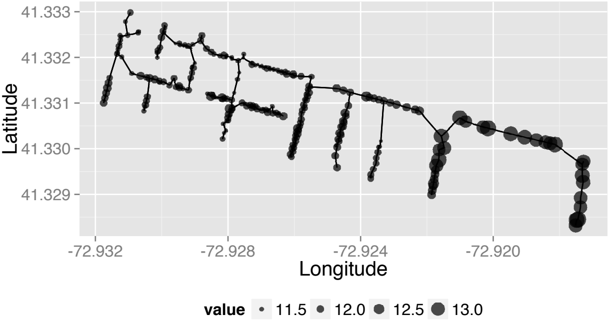
Example: New Haven Housing

$n = 244$ houses



Example: New Haven Housing

$n = 244$ houses

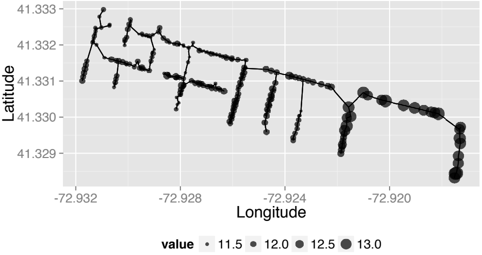


Circle size corresponds to log(2011 assessed values).

Visualizing the Properties

Available characteristics:

- # of bedrooms
- lot size
- living area size



Goal: To model 2011 assessed values using the other variables.

A Naive Approach

Model 1: Plain Old Linear Regression

```
lm(assessed values ∼ number of bedrooms + lot size +
                     living area)
```

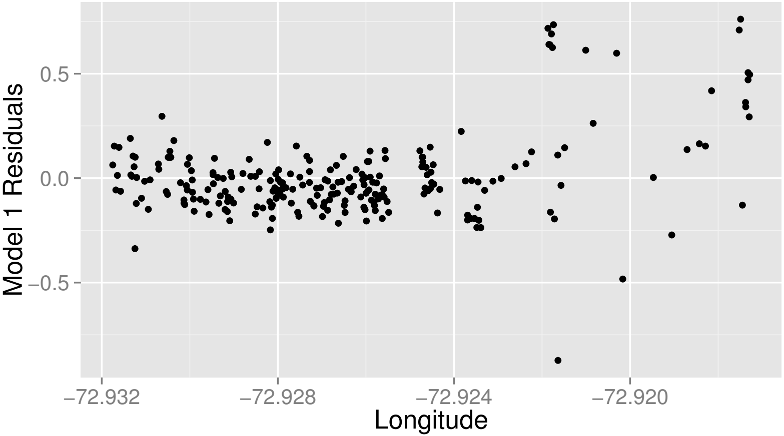
Model 1 Summary

	Estimate	Std. Error	t value	p value
(Intercept)	10.4953	0.0674	155.72	$<$2e-16
living area	0.0281	0.0022	12.89	$<$2e-16
beds	-0.0364	0.0116	-3.13	0.0020
size	1.5266	0.1079	14.15	$<$2e-16

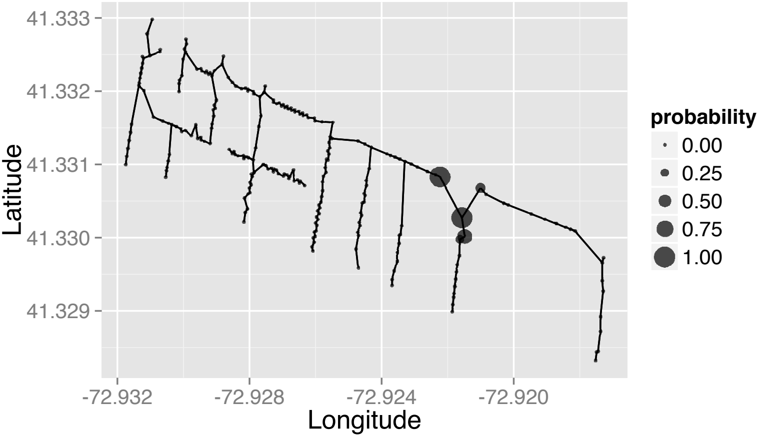
Residual standard error: 0.1998 on 240 degrees of freedom

Multiple R-squared: 0.8225

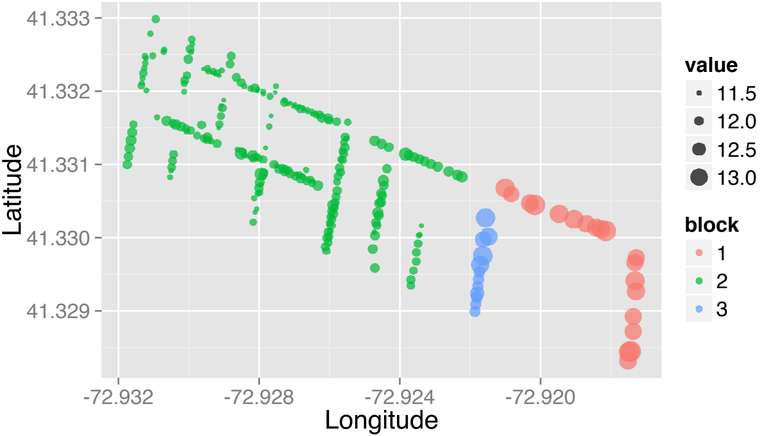
Model 1 Residuals



BCP: Posterior Boundary Probabilities

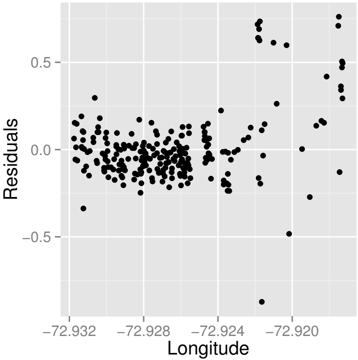


BCP: Modal Partition

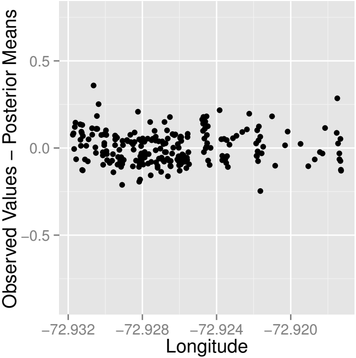


Model Comparison

Model 1: Regression
residual SE: 0.1998

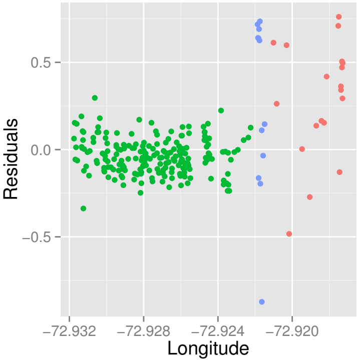


Model 2: BCP
SD("residuals"): 0.0920

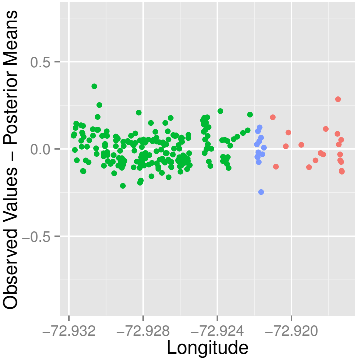


Model Comparison

Model 1: Regression
residual SE: 0.1998



Model 2: BCP
SD("residuals"): 0.0920



Summary

- BCP produces good results in simulation settings and credible results on real data examples

Summary

- BCP produces good results in simulation settings and credible results on real data examples
- In simulation cases where comparisons are possible,

Summary

- BCP produces good results in simulation settings and credible results on real data examples
- In simulation cases where comparisons are possible,
    - BCP usually performs much better than other methods

Summary

- BCP produces good results in simulation settings and credible results on real data examples
- In simulation cases where comparisons are possible,
  - BCP usually performs much better than other methods
  - in a few cases, BCP performs worse, but not by much

Summary

- BCP produces good results in simulation settings and credible results on real data examples
- In simulation cases where comparisons are possible,
  - BCP usually performs much better than other methods
  - in a few cases, BCP performs worse, but not by much
- Runtime can be slow, particularly for large datasets on a general graph

Preprint of article submitted to JASA:

- Wang and Emerson (2015). "Bayesian Change Point Analysis of Linear Models on Graphs." `http://arxiv.org/abs/1509.00817`.

Preprint of article submitted to JASA:

- Wang and Emerson (2015). "Bayesian Change Point Analysis of Linear Models on Graphs." `http://arxiv.org/abs/1509.00817`.

R package `bcp` is available on R CRAN:
`https://cran.r-project.org/web/packages/bcp/index.html`.

Preprint of article submitted to JASA:

- Wang and Emerson (2015). "Bayesian Change Point Analysis of Linear Models on Graphs." http://arxiv.org/abs/1509.00817.

R package bcp is available on R CRAN: https://cran.r-project.org/web/packages/bcp/index.html.

Thank you!