Bayesian Multiple Changepoint Detection: Mixing Documented and Undocumented Changepoints

Robert Lund
Clemson Math Sciences
Lund@Clemson.edu

Joint work with Shanghong Li, Yingbo Li, and Hewa Priyadarshani

Oaxaca Workshop
Statistics, Data Mining, and Environmental Sciences

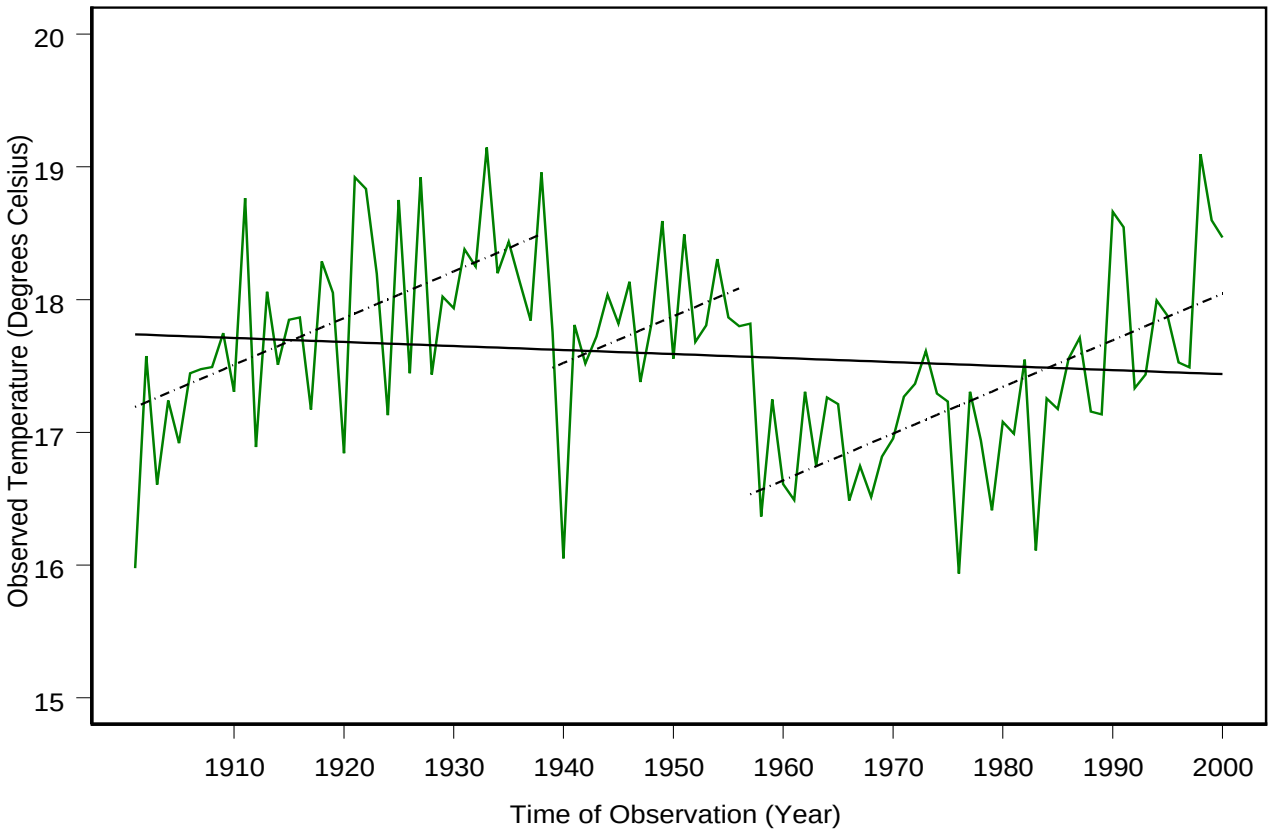
The Need to Detect Changepoints.

Changepoints are discontinuity times (inhomogeneities) in a time series. In climate settings, these can be induced from changes in observation locations, equipment, measurement techniques, environmental changes, etc.

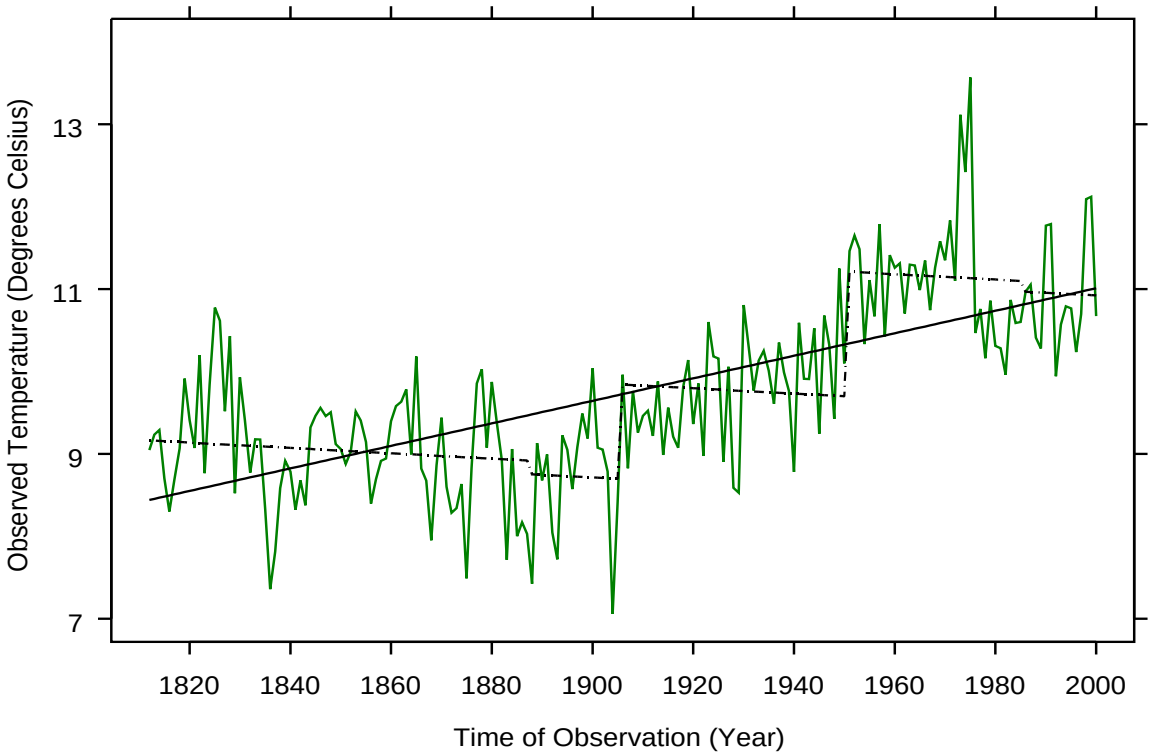
In this talk, a changepoint is a time where the mean of the series first undergoes a structural pattern change.

- Changepoint issues are critical when estimating trends.
- Many changepoints are undocumented.
- Changepoint techniques can help calibrate new gauges.

Tuscaloosa, AL Annual Temperatures



New Bedford, MA Annual Temperatures



Yearly Temperatures at New Bedford MA With Least Squares Trends

The Key Questions

- How many changepoints are there?
- At what times do the changepoints occur?

Some recent penalized likelihood references:

- Davis, Lee, and Rodriguez-Yam, *Journal of the American Statistical Association*, (2006).
- Lu, Lund, and Lee, *Annals of Applied Statistics*, (2010).
- Li and Lund, *Journal of Climate*, (2012).

The Crux

- Once changepoint times are identified, most statistical inference procedures are relatively straightforward.
- As such, this talk focuses on changepoint time identification.
- Changepoint problems are notoriously easy to statistically botch.

Underlying Changepoint Regression Model

For annual data $\{X_t\}_{t=1}^{N}$, our model is a simple linear time series regression with multiple mean level shifts:

$$X_t = \mu + \alpha t + \delta_t + \epsilon_t.$$

- Location parameter: μ.
- Linear time trend: αt. One can have other trend forms if desired.
- Piecewise constant mean shifts: $\{\delta_t\}$.
- Stationary but correlated errors: $\{\epsilon_t\}$.

Underlying Model

$$X_t = \mu + \alpha t + \delta_t + \epsilon_t.$$

The mean shifts are parametrized in $\{\delta_t\}$:

$$\delta_t = \begin{cases} \Delta_1 = 0, & 1 \leq t < \tau_1, \\ \Delta_2, & \tau_1 \leq t < \tau_2, \\ \vdots & \vdots \\ \Delta_{m+1}, & \tau_m \leq t \leq N \end{cases}.$$

Underlying Model

$$X_t = \mu + \alpha t + \delta_t + \epsilon_t.$$

The errors $\{\epsilon_t\}$ are a zero mean causal autoregressive process of known order p obeying

$$\epsilon_t = \sum_{k=1}^{p} \phi_k \epsilon_{t-k} + Z_t.$$

The process $\{Z_t\}$ is IID white noise with variance σ_Z^2.

Underlying Periodic Model

For monthly ($T = 12$) or daily ($T = 365$) data, our model uses a periodic time series regression with multiple level shifts:

$$X_{nT+\nu} = \mu_\nu + \alpha(nT + \nu) + \delta_{nT+\nu} + \epsilon_{nT+\nu}.$$

- The seasonal index $\nu \in \{1, \ldots, T\}$.
- μ_ν is the seasonal mean at season ν.
- α is a linear trend parameter.
- $\{\epsilon_{nT+\nu}\}$ is a causal periodic autoregression with period T.

Periodic Autoregressions

A zero-mean series $\{\epsilon_{nT+\nu}\}$ is called a periodic autoregression of order p (PAR(p)) and period T if it satisfies the periodic linear difference equation

$$\epsilon_{nT+\nu} = \sum_{k=1}^{p} \phi_k(\nu)\epsilon_{nT+\nu-k} + Z_{nT+\nu}.$$

Here, $\{Z_{nT+\nu}\}$ is zero-mean periodic white noise with $\mathrm{Var}(Z_{nT+\nu}) = \sigma_Z^2(\nu) > 0$ for all seasons ν.

$\phi_1(\nu), \ldots, \phi_p(\nu)$ are the PAR coefficients during season ν.

Such series are indeed "periodically stationary".

Penalized Likelihood Methods

A penalized likelihood for our model has form

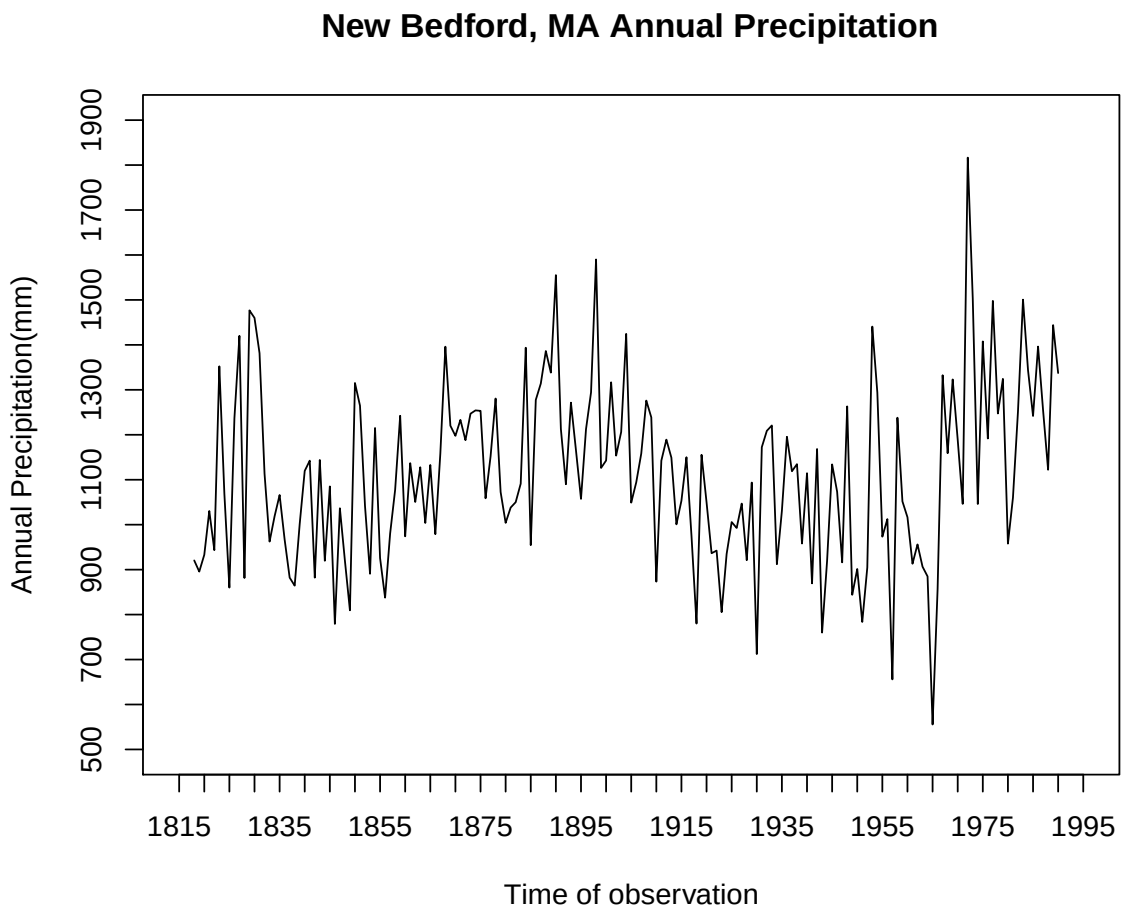
$$-2\log(L^*(m;\tau_1,\ldots,\tau_m)) + \text{Penalty}(m;\tau_1,\ldots,\tau_m).$$

- $L^*(m;\tau_1,\ldots,\tau_m)$ is an optimized model likelihood given m changepoints occurring at the times $\tau_1 < \cdots < \tau_m$.
- $\text{Penalty}(m;\tau_1,\ldots,\tau_m)$ is a penalty for the changepoint configuration.

Common $\text{Penalty}(m;\tau_1,\ldots,\tau_m)$ terms used:

- $\text{AIC} = 2m$.
- $\text{BIC} = m\log(N)$.
- $\text{MDL} = \sum_{i=1}^{m+1} \frac{\log(\tau_i - \tau_{i-1})}{2} + \log(m+1) + \sum_{i=2}^{m} \log(\tau_i)$.

New Bedford, MA Annual Precipitations



Lognormal Annual Precipitation Setup

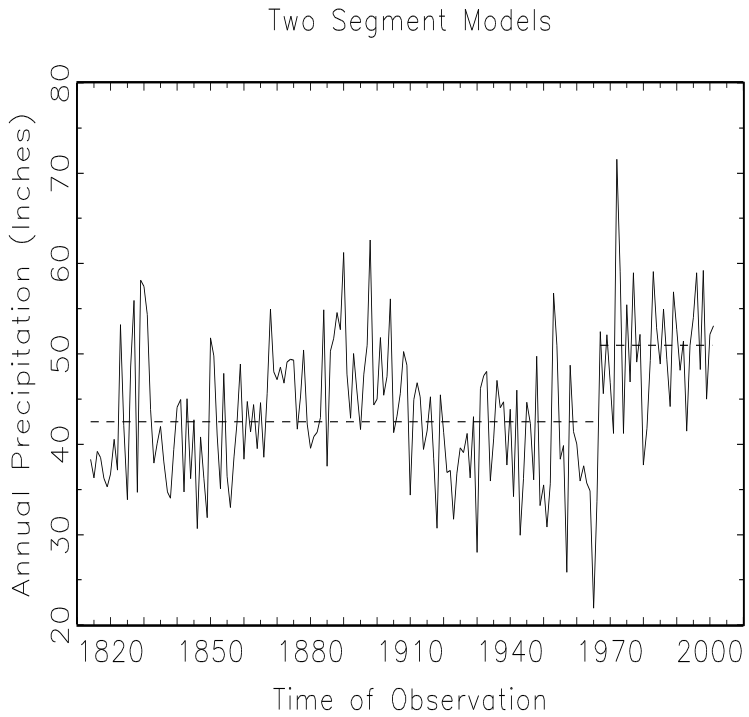
The logarithm of $\{X_t\}$ is modeled as a Gaussian time series (this sets the likelihood) with no trend, multiple mean shifts, and autoregressive errors (AR(p)). Here, $T = 1$: no periodicities.

For each changepoint configuration $(m; \tau_1, \ldots, \tau_m)$, we must

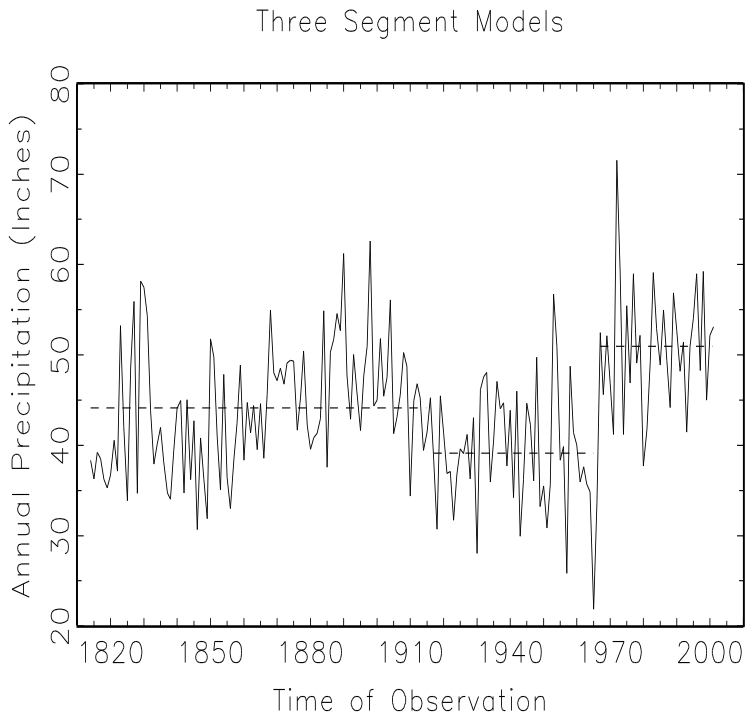
- Fit a Gaussian time series model with optimal time series parameters and mean shift sizes;
- Compute the penalty

$$\text{MDL}(m; \tau_1, \ldots, \tau_m) = \sum_{i=1}^{m+1} \frac{\log(\tau_i - \tau_{i-1})}{2} + \log(m+1) + \sum_{i=2}^{m} \log(\tau_i).$$

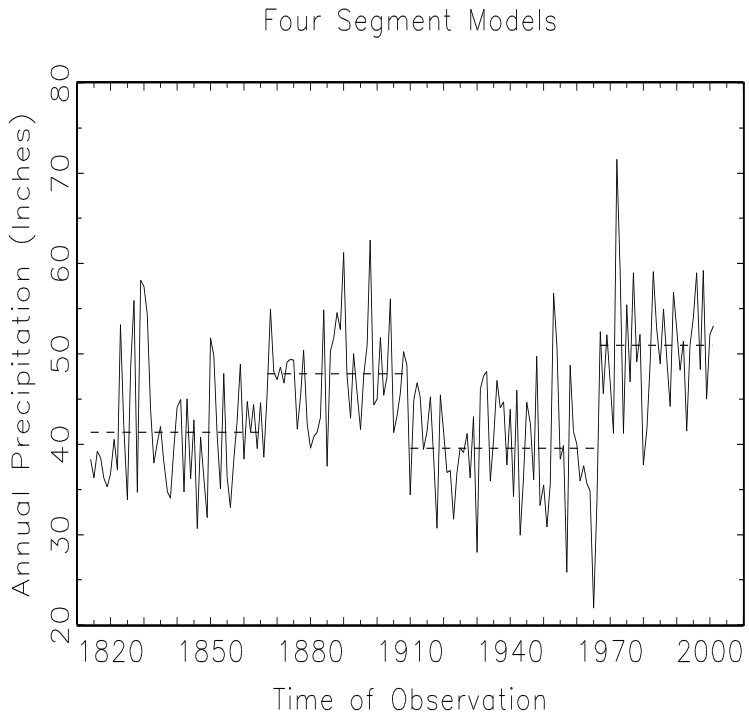
Two Segment Models



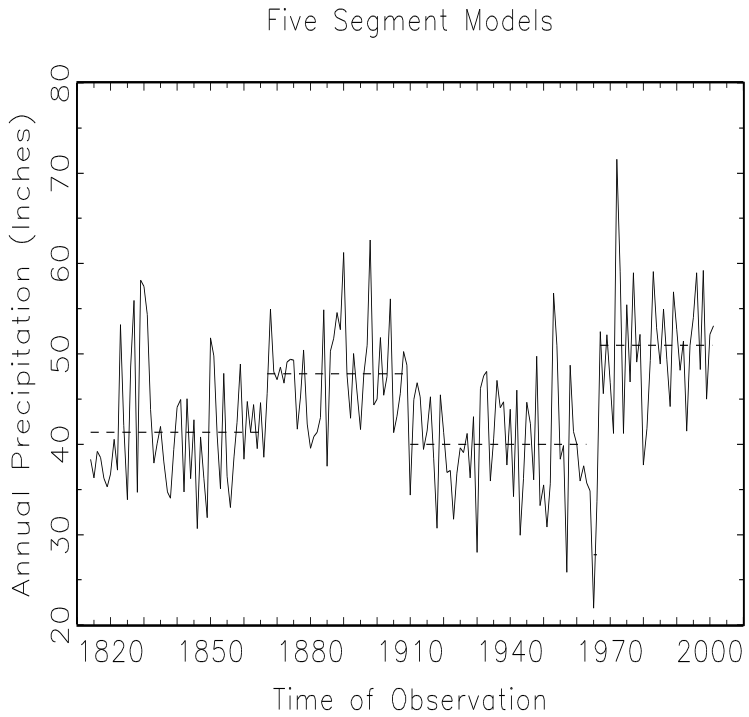
Three Segment Models



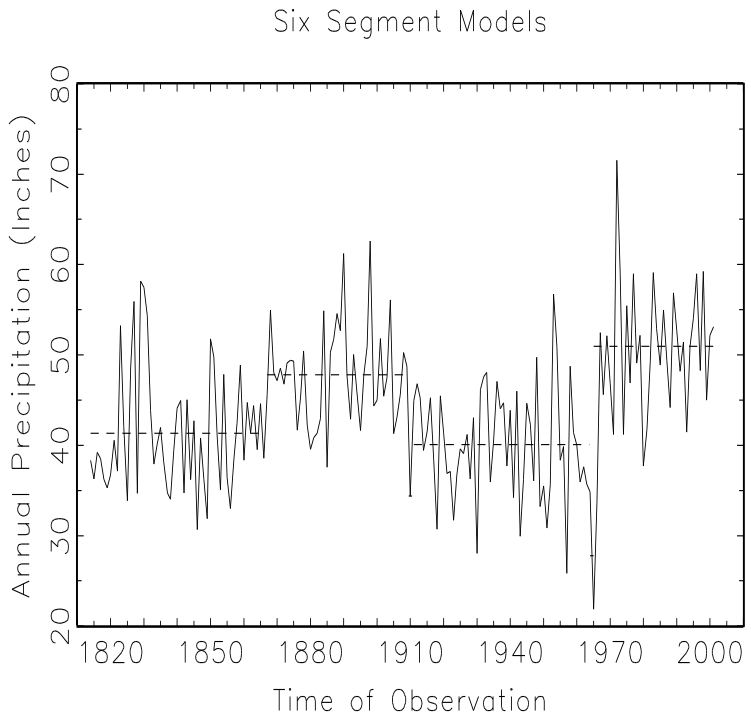
Four Segment Models



Five Segment Models



Six Segment Models



Summary

The table below shows optimum MDL scores for various numbers of model segments. These values were obtained by exhaustive search and are exact.

Table: Optimum MDL Scores

# Segments	Changepoint Times	MDL Score
1	—	-296.7328
2	1967	-303.8382
3	1917, 1967	-306.6359
4	1867, 1910, 1967	-309.2878
5	1867, 1910, 1965, 1967	-309.8570
6	1829, 1832, 1867, 1910, 1967	-308.2182

The Combinatorial Wall

We need to minimize

$$-2\log(L^*(m;\tau_1,\ldots,\tau_m)) + \mathrm{MDL}(m;\tau_1,\ldots,\tau_m)$$

over all m and $\tau_1,\ldots,\tau_m$.

An exhaustive search over all models with m changepoints requires evaluation of $\binom{N-1}{m}$ MDL scores.

Summing this over $m = 0, 1, \ldots, N-1$ shows that an exhaustive optimization requires 2^{N-1} different MDL evaluations.

We now devise a genetic algorithm for this. A genetic algorithm is an intelligent random walk search.

Genetic Algorithms (GAs)

Chromosome Representation. Each changepoint configuration has the form $(m; \tau_1, \ldots, \tau_m)$.

Selection. Give mating preference to the fittest individuals, allowing them to pass their genes on to the next generation. Chromosome fitness is determined by the objective function

$$-2\log(L^*(m; \tau_1, \ldots, \tau_m)) + \text{MDL}(m; \tau_1, \ldots, \tau_m).$$

GA Details

Starting the GA. An initial generation of size 200 chromosomes is simulated. Each chromosome is simulated by allowing each admissible time to be a changepoint with a small probability, set to average six changepoints per century.

GA Details

Chromosome Crossover. To produce each child in each successive generation of size 200, two current chromosomes are chosen as parents from the current generation to breed — call these $(m; \tau_1, \ldots, \tau_m)$ and $(k; \eta_1, \ldots, \eta_k)$. Better fit chromosomes are more likely to be selected as parents. We choose parents proportionally to fitness ranks $\{1, 2, \ldots, 200\}$.

A new child chromosome $(\ell; \xi_1, \ldots, \xi_\ell)$, having traits of both parents, is created by combining all changepoint times of both parents and then thinning these with fair coin flips.

GA Details

Mutation. Increases the diversity of the population, preventing premature convergence to suboptimal solutions. Our mutation mechanism allows a small portion of generated children to have extra changepoints. After each child is formed, each and every non-changepoint time is independently allowed to become a changepoint time with probability p_m. Typically, p_m is very small.

GA Details

Algorithm Termination. Successive generations are simulated until a termination condition has been reached. The best solution to the problem is the chromosome in the current generation with the smallest penalized likelihood.
Common terminating conditions are:

- A solution is found that satisfies a minimum criteria.
- A fixed number of generations is reached.
- The generation's fittest ranking member is peaking (successive iterations no longer produce better results).

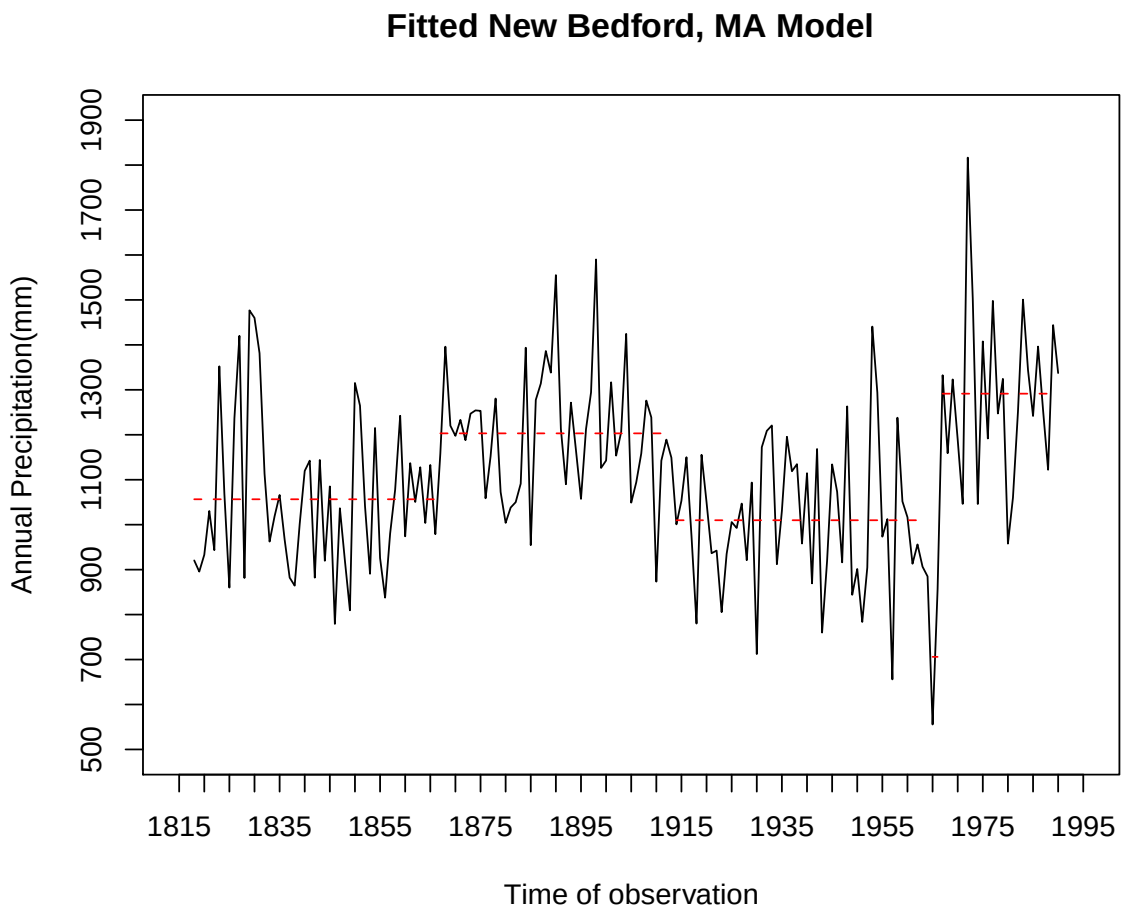
Optimal Model

The GA algorithm converged to a model with four changepoints at times $1867, 1910, 1965,$ and 1967.

The minimum MDL score achieved was -309.8570.

This segmentation is graphed against the data and appears visually reasonable.

Optimal Model Has Four Changepoints!



Simulations — Set I

Mimics the New Bedford Data with lognormal distributions:

1000 series of length $N = 200$ with no trend, seasonality, or changepoints: $\mu_t \equiv 6.8$; AR(1) errors $\{\epsilon_t\}$ with $\phi = 0.2$ and $\sigma^2 = 0.025$.

Table: Empirical proportions of estimated changepoint numbers. The correct value of m is zero.

m	Percent
0	99.0 %
1	0.4 %
2	0.5 %
3+	0.1 %

Simulations — Set II

$$\mu_t = \begin{cases} 6.8 & 1 \leq t \leq 49 \\ 7.0 & 50 \leq t \leq 99 \\ 7.2 & 100 \leq t \leq 149 \\ 7.4 & 150 \leq t \leq 200 \end{cases} .$$

Table: Empirical proportions of estimated changepoint numbers ($m = 3$)

m	Percent
0	0.0 %
1	3.6 %
2	28.8 %
3	63.1 %
4	4.3 %
5+	0.2 %

Count Detection Histogram

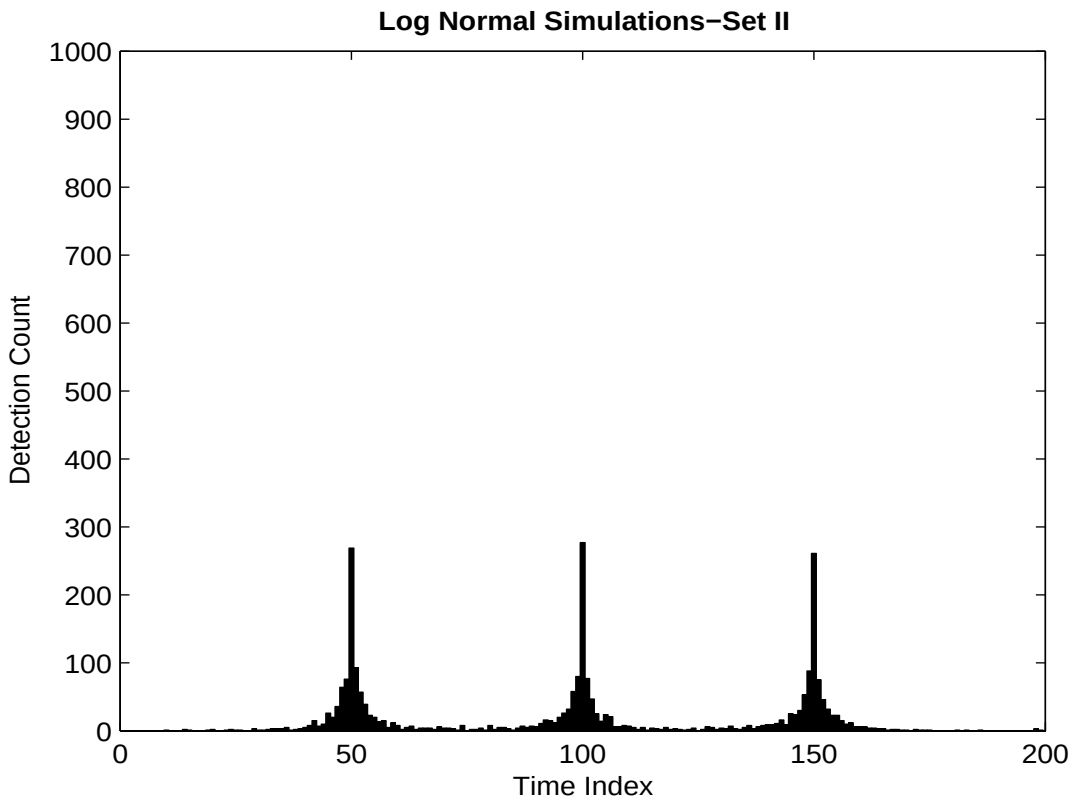


Figure: The detected changepoint times cluster around their true times of 50, 100, and 150.

Simulations — Set III

$$\mu_t = \begin{cases} 6.8 & 1 \leq t \leq 24 \\ 7.0 & 25 \leq t \leq 74 \\ 6.6 & 75 \leq t \leq 99 \\ 6.8 & 100 \leq t \leq 200 \end{cases} .$$

Table: Empirical proportions of estimated changepoints ($m = 3$)

m	Percent
0	0.0 %
1	6.0 %
2	19.5 %
3	69.2 %
4	5.1 %
5+	0.2 %

Count Detection Histogram

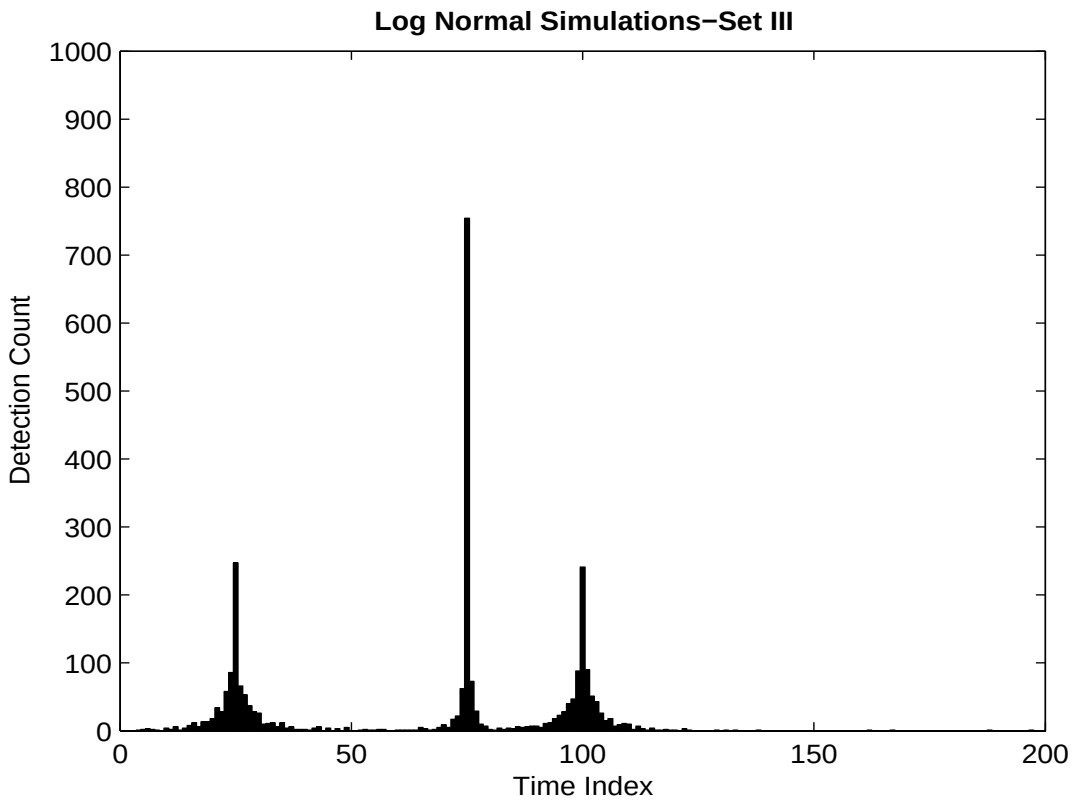


Figure: The detected changepoint times cluster around their true times of 25, 75, and 100.

Metadata

- In climate applications, metadata is a record listing times at which the station moves or instrumentation changes.
- Some stations have metadata, some do not.
- Mitchell (1953): US stations average six station and/or gage changes per century.
- Not all change times are necessarily listed in the metadata — metadata is notoriously incomplete.
- Not all metadata times necessarily impart a true mean shift.

MDL Modifications

We revisited the original MDL information theory derivation and modified it to accommodate prior information.

Our prior on the changepoint configuration imposes:

- Every time listed in the metadata has probability p_1 of inducing a mean shift.
- Every time not listed in the metadata has probability p_2 of inducing a mean shift.
- Metadata times are more likely to induce mean shifts: $p_1 \geq p_2$.
- Changepoint declarations at distinct times are assumed mutually independent.

Miscellaneous

Other model specifications:

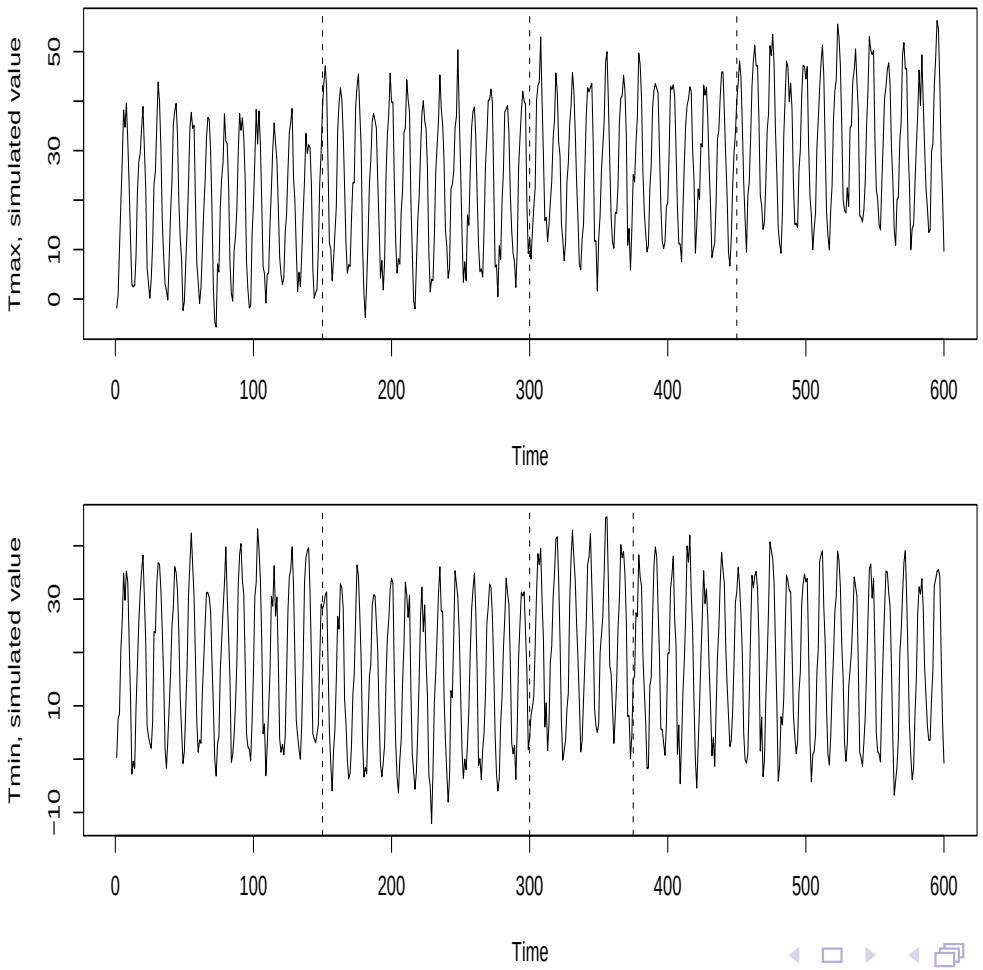
- Normally distributed prior mean shift sizes are imposed.
- Beta hyper-priors are placed on p_1 and p_2.

We integrated out everything possible.

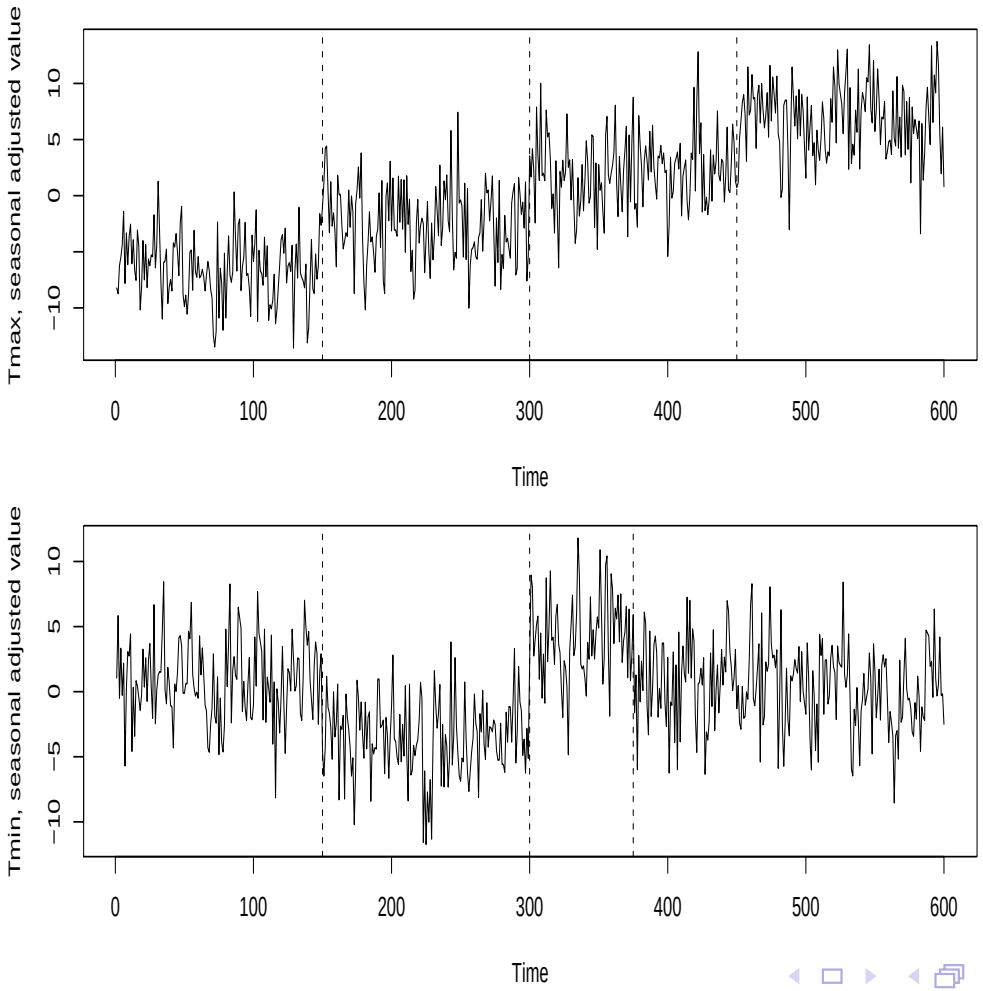
BMDL$(m; \tau_1, \ldots, \tau_m)$ is the resulting penalized likelihood with m changepoints at the times $\tau_1, \ldots, \tau_m$ — form is messy!

The best model is taken as the one that minimizes the BMDL; this is equivalent to finding the model that maximizes the Bayesian posterior distribution. Optimization proceeds as before.

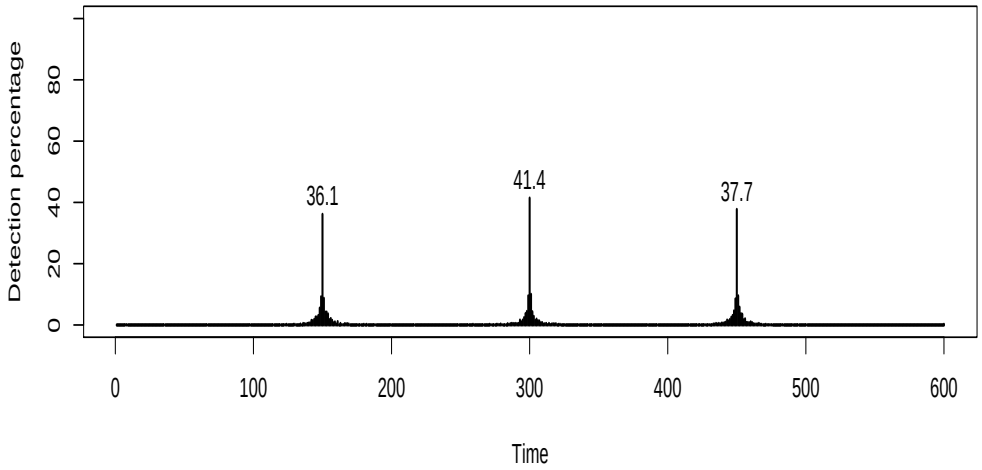
A Simulated Monthly Series, Maximums and Minimums

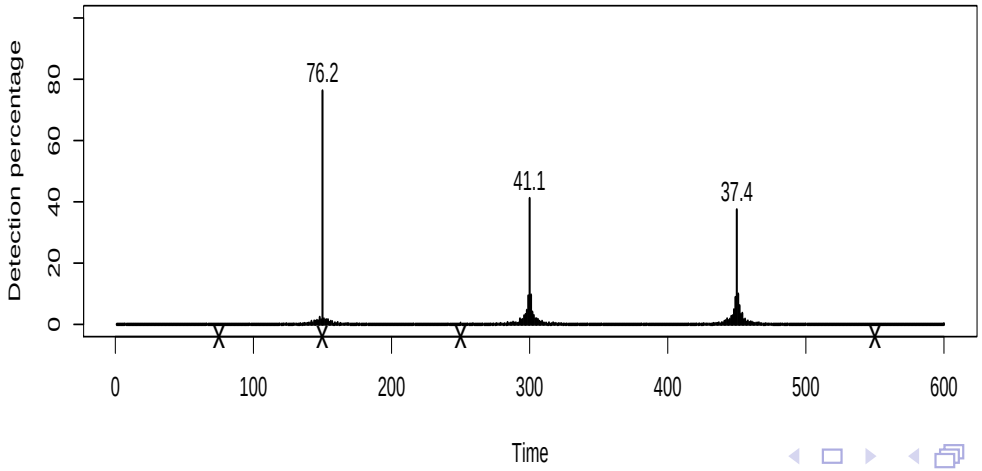


Last Graph after Subtracting Seasonal Means



Detection Rates over 1000 Independent $T_{\max}$ Simulations





Estimates of the Changepoint Count m

Table: Statistics for the estimated changepoint count of the maximum series averaged over the 1000 runs. The correct changepoint count is $m = 3$. Metadata times: 75, 150, 250, and 550. Signal to noise ratio = 1.5.

Method	$\hat{m}$	Standard Error
BIC	3.12	0.38
BMDL	3.02	0.13

- Other simulations: the methods select $m = 0$ over 99% of the time when there are no changepoints.
- Overall, the methods seem to work very well!

Asymptotics

- Infill asymptotics literature: Davis et al. (2006), Davis and Yau (2013), and Du et al. (2016).
- Relative changepoint configuration: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)'$: $\lambda_i = \tau_i / N$. This scales changepoint times to lie in [0,1].
- Put a superscript of zero on all true parameters. Here, m^0, the true number of changepoints, is unknown.
- Consider all changepoint configurations in

$$\boldsymbol{\Lambda} = \left\{ \boldsymbol{\lambda} : 0 \leq m \leq M, \min_{1 \leq r \leq m+1} \lambda_r - \lambda_{r-1} > \epsilon \right\}$$

Here, $M > m^0$ is a large positive integer and $\epsilon > 0$ is small.

Asymptotic Consistency

- The estimated relative changepoint configuration:

$$\hat{\boldsymbol{\lambda}}_N = \arg\min_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \mathrm{BMDL}(\boldsymbol{\lambda}).$$

- The estimated number of changepoints is $\hat{m}_N = |\hat{\boldsymbol{\lambda}}_N|$.

Theorem: In the stationary case ($T = 1$), as $N \to \infty$,

$$\hat{m}_N \xrightarrow{\mathcal{P}} m^0, \quad \hat{\boldsymbol{\lambda}}_N \xrightarrow{\mathcal{P}} \boldsymbol{\lambda}^0.$$

Furthermore, for each $r \in \{1, 2, \ldots, m^0\}$,

$$|\hat{\lambda}_r - \lambda_r^0| = O_P\left(\frac{1}{N}\right).$$

Consistency of Parameter Estimators

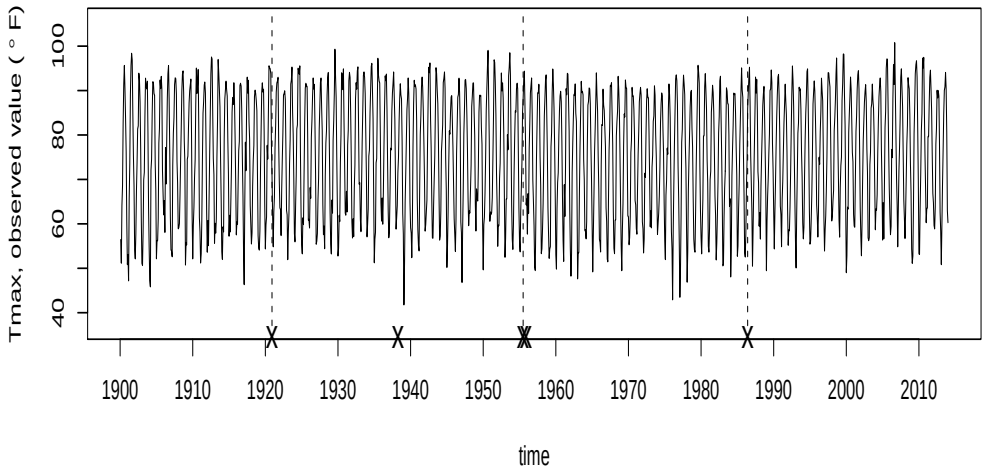
A theorem under the estimated changepoint model $\hat{\boldsymbol{\lambda}}_N$ with

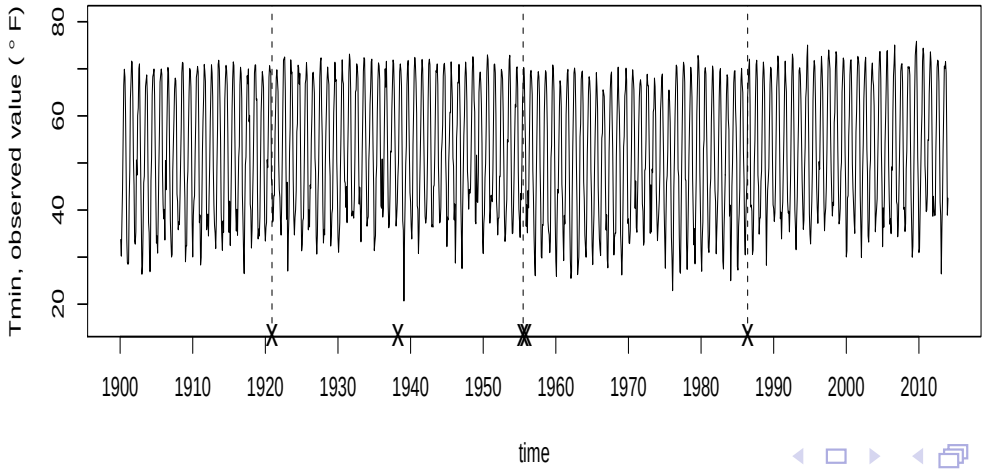
- Stationary errors ($T = 1$) and Yule-Walker estimators for the AR(p) time series parameters;
- BMDL optimizers for the seasonal means;
- Conditional posterior mean for $\boldsymbol{\Delta}$.

Theorem: As $N \to \infty$, all parameters converge to their true values:

$$\hat{\boldsymbol{\Delta}}_N \xrightarrow{\mathcal{P}} \boldsymbol{\Delta}^0, \quad \hat{\mathbf{s}}_N \xrightarrow{\mathcal{P}} \mathbf{s}^0, \quad \hat{\boldsymbol{\phi}}_N \xrightarrow{\mathcal{P}} \boldsymbol{\phi}^0, \quad \hat{\sigma}^2_N \xrightarrow{\mathcal{P}} (\sigma^2)^0.$$

Tuscaloosa Monthly Analysis: Tmax and Tmin





References

The metadata material was taken from:

- Li, Y. and R.B. Lund (2015). Multiple Changepoint Detection using Metadata, *Journal of Climate*, 28, 4199-4216.
- Priyadarshani, H., Y. Li, Y., R.B. Lund and J. Rennie (2017). Homogenization of Daily Temperature Data, *Journal of Climate*, 30, 4199-4216.
- Li, Y., Lund, R.B., and H. Priyadarshani (2017+). Multiple Changepoint Detection with Partial Information on the Changepoint Times, Submitted to, *Journal of the Royal Statistical Society*.

Thank you! ☺