Simultaneous Changepoint Detection in Temperature Data

Steven Moen and Yuhan Liu

The University of Chicago, Department of Statistics

Abstract

- Motivation for this work came from monthly temperature readings taken across the United States, which seem to be more tightly linked to each other in the mid-2000s than they were in prior decades
- To analyze this trend more carefully, we have created a test statistic that allows for inference about whether distinct time series have a simultaneous change point, which for our purposes means a change in the first derivative in two different series at the same point in time
- This statistic can allow us to measure the likelihood that two temperature time series are changing at the same time or whether the change is simply caused by random chance
- This statistic does not have a closed-form analytical density, so Monte Carlo simulation is required for hypothesis testing. Developing a proper simulation study to precisely measure the noise in the underlying data is beyond the scope of this work, but the statistic can be used to measure change in many areas of earth science in addition to temperature

Introduction and Motivation

Using average temperature data gathered from 15 locations across the U.S. from the U.S. Historical Climate Network [1] and fitting local quadratic models shows some evidence of an increase in the rate of warming in the mid-2000s:

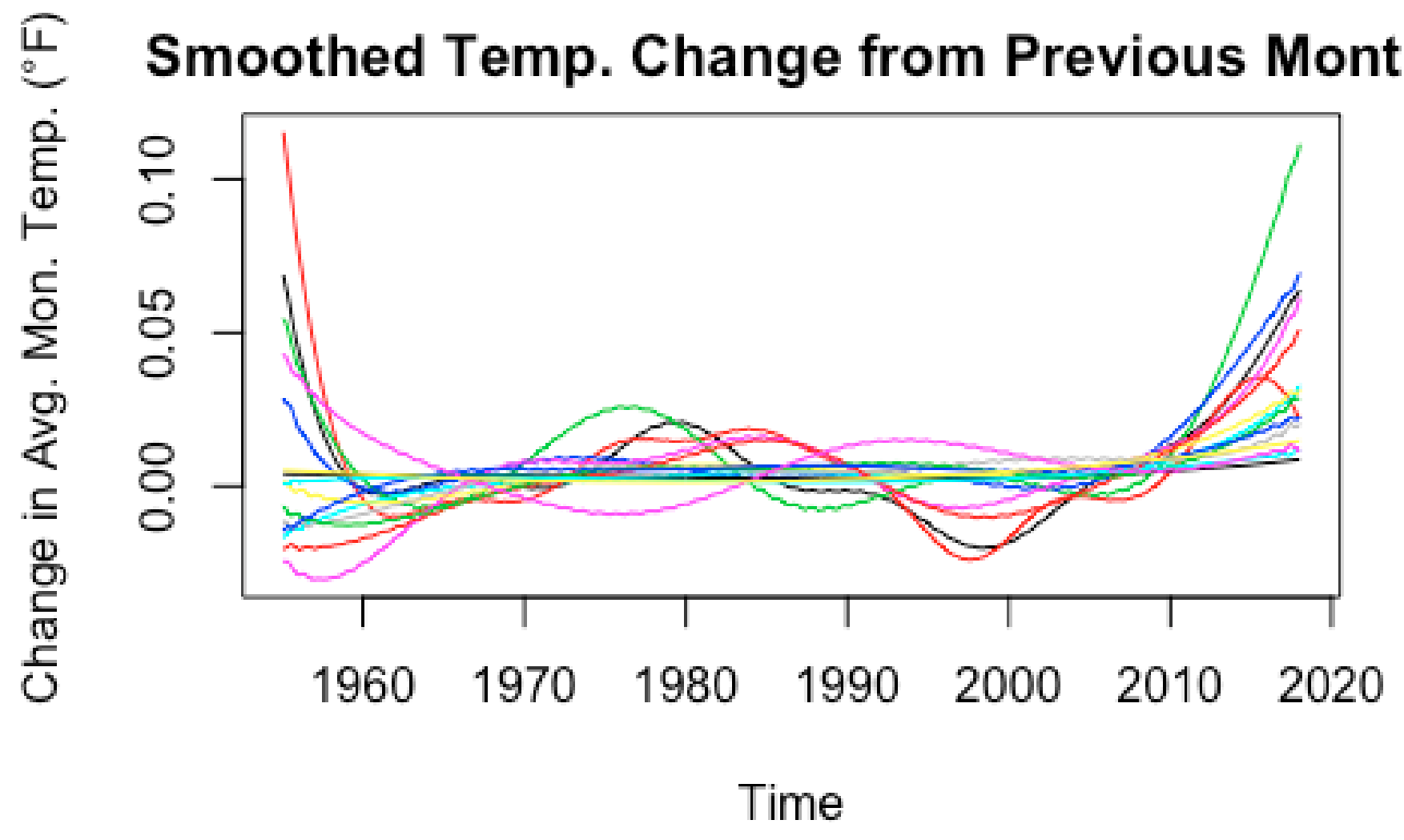

Figure 1:Many temperature series are changing simultaneously in the mid-2000s

Test Statistic Derivation

Consider two sets of observations indexed by time:

$$(X_1, Y_1), \cdots, (X_T, Y_T)$$

To measure whether they have a similar changepoint in their first derivative, the first step is to split each time series at a hypothesized changepoint and fit two separate quadratic regression models to each series:

$$X_{t,1} = a_2 t^2 + a_1 t + a_0 \qquad Y_{t,1} = b_2 t^2 + b_1 t + b_0$$

$$X_{t,2} = \hat{a}_2 t^2 + \hat{a}_1 t + \hat{a}_0 \qquad Y_{t,2} = \hat{b}_2 t^2 + \hat{b}_1 t + \hat{b}_0$$

From these models, their analytical first derivatives can be calculated as follows:

$$X'(t,1) = 2a_2 t + a_1 \qquad Y'(t,1) = 2b_2 t + b_1$$

$$X'(t,2) = 2\hat{a}_2 t + \hat{a}_1, \qquad Y'(t,2) = 2\hat{b}_2 t + \hat{b}_1$$

After calculating these above first derivatives, the average of two different slopes connecting three consecutive points would be calculated as follows:

$$\hat{X}_t = \frac{X_{t+1} + X_{t-1} - 2X_t}{2}, \qquad \hat{Y}_t = \frac{Y_{t+1} + Y_{t-1} - 2Y_t}{2}$$

After finding these two sets of values, they would be compared in using the residual sum of squares (RSS) statistic for every possible change point. The calculation necessary to find the RSS is given below:

$$RSS_{X'} = \sum_{t=1}^{\lfloor T_X \rfloor} (\hat{X}_t - X'(t,1))^2 + \sum_{t=\lfloor T_X \rfloor + 1}^{T} (\hat{X}_t - X'(t,2))^2$$

$$RSS_{Y'} = \sum_{t=1}^{\lfloor T_Y \rfloor} (\hat{Y}_t - Y'(t,1))^2 + \sum_{t=\lfloor T_Y \rfloor + 1}^{T} (\hat{Y}_t - Y'(t,2))^2$$

Test Statistic Derivation

After the time point that minimizes RSS is found for each series separately, then the RSS is found under the constraint that the value of t must be the same for both X and for Y, as described below:

$$RSS_{X' \cup Y'} = \sum_{t=1}^{\lfloor T_{X \cup Y} \rfloor} (\hat{X}_t - X'(t))^2 + \sum_{t=\lfloor T_{X \cup Y} \rfloor}^{T} (\hat{X}_t - X'(t))^2 +$$
$$\sum_{t=1}^{\lfloor T_{X \cup Y} \rfloor} (\hat{Y}_t - Y'(t))^2 + \sum_{t=\lfloor T_{X \cup Y} \rfloor}^{T} (\hat{Y}_t - Y'(t))^2$$

Distribution of RSS

To understand the distribution of each RSS term, two lemmas are required:

1. $\mathbf{X} \sim \mathbf{N}_k(\mu, \Sigma) \iff \exists \mu \in \mathbb{R}^k, M \in \mathbb{R}^{k \times l}$ such that $\mathbf{X} = M\mathbf{Z} + \mu$ for $\mathbf{Z} \sim \mathbf{N}(0, I_l)$. [2]
2. Let $\mathbf{Y} \sim N_n(O, I_n)$ and let A be a symmetric matrix. Then $\mathbf{Y}'A\mathbf{Y}$ is $\mathcal{X}_r^2$ if and only if A is idempotent of rank r. [3]

Suppose our model is $\mathbf{Y} = X\alpha + \boldsymbol{\varepsilon}$, where the noise vector $\boldsymbol{\varepsilon} \sim \mathbf{N}_n(0, \Sigma_n)$. Thus, by Lemma 1, there exists a $M \in \mathbb{R}^{k \times l}$, and a $\boldsymbol{\delta} \sim \mathbf{N}_p(0, I_p)$ such that $\boldsymbol{\varepsilon} = M\boldsymbol{\delta}$.

Using the above fact, and noting that $P = X(X^T X)^{-1} X^T$

$$\begin{aligned} RSS &= Y'(I_n - P)Y \\ &= (Y - X\alpha)'(I_n - P)(Y - X\alpha) \\ &= \boldsymbol{\varepsilon}'(I_n - P)\boldsymbol{\varepsilon} = \boldsymbol{\delta}' M'(I_n - P)M\boldsymbol{\delta} \end{aligned}$$

Then by Lemma 2, RSS is distributed according to the Chi-Square distribution.

Test Statistic

The RSS statistics described above are combined into one test statistic, which can be used for hypothesis testing. **Assuming that the underlying stochastic process is multivariate Gaussian, all of the terms in the statistic below are distributed as Chi-Square random variables, though K is not.** Larger values of K support rejecting the null of the same changepoint.

$$K = RSS_{X' \cup Y'} - (RSS_{X'} + RSS_{Y'})$$

Comments on the Test Statistic

- The test statistic K **does not** have a closed-form analytical density without the highly unrealistic assumption of independence between the terms used to calculate K
- $K \geq 0$ because the model assuming a changepoint at the same time is inherently simpler than a model assuming different changepoints
- Despite an extensive literature search, it looks like Monte Carlo testing is required for statistical inference, which is beyond the scope of this poster

Future Work and Acknowledgements

- Detecting changepoints in real data is difficult, and Monte Carlo simulation is necessary to construct a null distribution for hypothesis testing
- To conduct proper Monte Carlo analysis. a realistic theoretical model is required, so combining this test statistic with existing, sound theoretical models will allow for powerful insights
- The test statistic could be generalized to conduct more than just pairwise comparisons

Special thanks to Dr. Wei Biao Wu (Professor, Department of Statistics and the College at the Univeristy of Chicago) for his guidance and for the invaluable comments from people affilated with the Center for Spatial Data Science at UChicago.

References

[1] Ron; Applequist Scott; Korzeniewski Bryant; Menne Matthew J. Lawrimore, Jay H.; Ray.
Global summary of the month (gsom), 2016.
NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5QV3JJ5. Accessed June 2019.

[2] Allan Gut.
An Intermediate Course in Probability.
Springer, 2nd edition, 2009.

[3] George A.F. Seber and Alan J. Lee.
Linear Regression Analysis.
John Wiley & Sons, Inc., 2nd edition, 2003.