AN ABSTRACT OF THE THESIS OF

Satpreet Harcharan Singh for the degree of Master of Science in Computer Science presented on August 1, 2017.

Title: Visualization and Analysis of Sensor Data for Detecting Microclimate Cold Air Pools

Abstract approved: ____

Weng-Keen Wong

Cold air pools are spatiotemporal phenomena that occur when cold air from higher elevations roll down the slope to accumulate in lower elevations. Behaviors like this lead to microclimate anomalies such as the city of Corvallis (Oregon) experiencing persistent cold weather even on a sunny day. We analyze multivariate temperature time-series data and associated covariates from about 160 sensors from the HJ Andrews Research Forest (Oregon) through visualization and modeling to study this phenomenon. We develop detectors to localize cold air pools in both time and space, and carry out simulation studies to assess their performance under different microclimatic and sensor-performance conditions.

Visualization and Analysis of Sensor Data for Detecting Microclimate Cold Air Pools

by

Satpreet Harcharan Singh

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented August 1, 2017
Commencement June 2018

Master of Science thesis of Satpreet Harcharan Singh presented on August 1, 2017.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Satpreet Harcharan Singh, Author

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor, Dr. Weng-Keen Wong, for his immense help and support during these last two years, starting from his invitation to explore a project with him at the intersection of Machine Learning and Sustainability/Ecology, to his dedication to the advising role in the face of the many big changes that happened in his life during this time-period.

We thank Dr. Julia Jones, from the OSU Dept. of Forestry, who provided us the data and the domain-expertise that was essential to making progress on this project.

I would like to thank my committee: Dr. Raviv Raich, Dr. Stephen Ramsey and Dr. Rebecca Hutchinson (and Dr. Matthew Campbell, GCR), for their insightful questions and feedback.

Special thanks to Dr. Stephen Ramsey for hosting me for two terms in his lab working on some very interesting Computational Genomics projects.

I am grateful to many instructors in the EECS department at OSU for their wonderful courses. Thank you to my friends and collaborators at OSU for making the graduate-school experience even more enjoyable! Special thanks to Michael Slater for helping me polish my defense slides, and for being an all-round great colleague during my time here!

Last but not the least, I am grateful to my parents and other well-wishers for their love and support!

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

Chapter 1: Introduction

The collection and analysis of large-scale environmental-data via sensor networks has become popular in the Ecological Sciences community and has brought with it new data-processing and interpretation related challenges [5]. This thesis deals with one such dataset collected for studying, besides other things, a spatiotemporal phenomenon called Cold Air Pools (CAPs) [13, 10, 2, 3, 8, 7, 6].

1.1 HJ Andrews Experimental Forest

The HJ Andrews Experimental Forest (HJA) is a 16,000-acre ecological research site in Oregons western Cascades Mountains, and the subject of several long-term ecological research studies at the Oregon State University (OSU). **Dr. Julia Jones** [17] from the OSU College of Earth, Ocean, and Atmospheric Sciences, is our domain-expert collaborator.

1.2 Dataset

For our analyses, we use the temperature multivariate time-series data, sampled every 20 minutes, and associated covariates (Latitude, Longitude and Elevation) from about **160 sensor**s deployed across the HJA (Figure 1.2). Other covariates such as the aspect of the sensors, and time-series of solar-radiation, are also available, but have not been used in this analysis.

Our typical analysis considers all **72 time-steps** (sampled every 20 minutes, e.g. figure 1.3) of a day's 24-hour time-series , except in a small fraction of a analyses, where we only consider the nightly (4pm-7am) cooling curve consisting of 48 time-steps. Sensor elevations range from 469 meters to 1558 meters (above sea level).

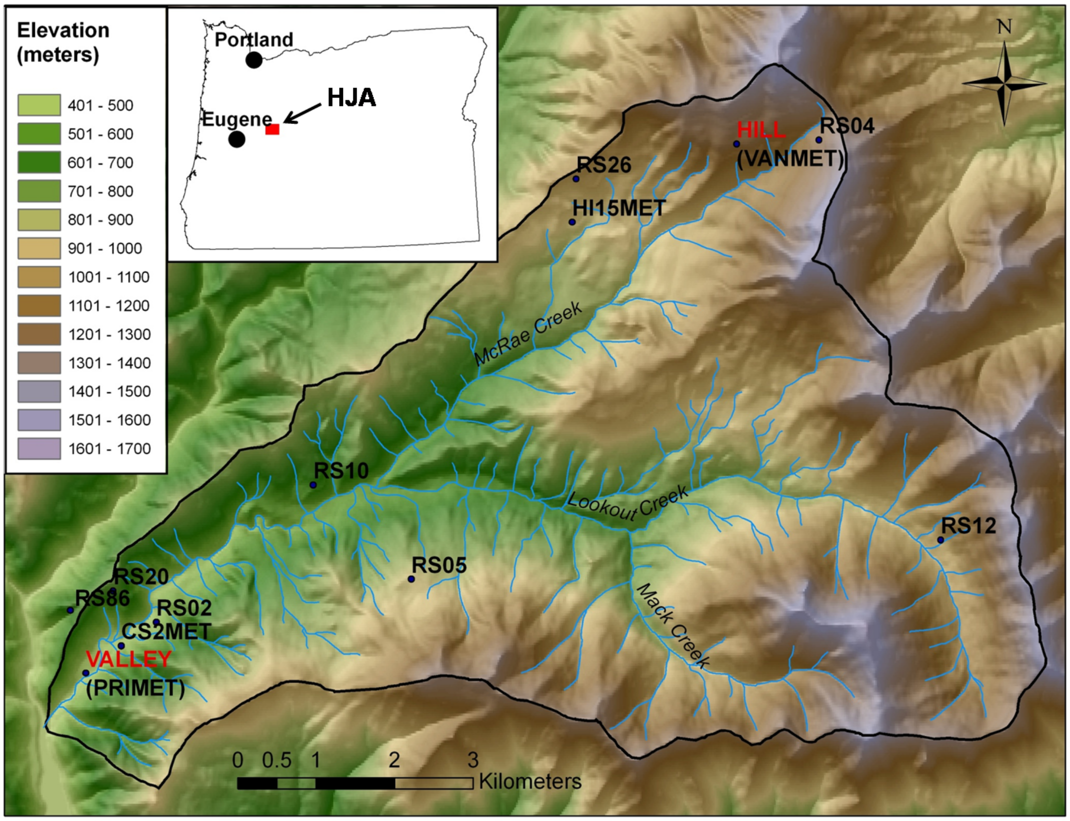

Figure 1.1: Topological Relief Map of the HJA. Inset shows rough location of HJA in the state of Oregon. (Used with author's permission from [2])

1.3 Cold Air Pools

Cold Air Pools occur when cold air from higher elevations roll down a slope to accumulate in lower elevations, or when atmospheric processes favor warming of higher elevations [12] or both. Persistent CAPs are associated with stagnation of air in the topographic depression in which they form, and with low insolation (exposure to sun's rays) [12]. As an example, microclimate anomalies such as the city of Corvallis (OR) experiencing persistent cold weather even on a sunny day, can be explained by Cold Air Pooling.

Since CAPs are formed when insolation is low, they are more frequent and more pronounced in **winter** than in summer. There are also transient CAPs formed on **cold**

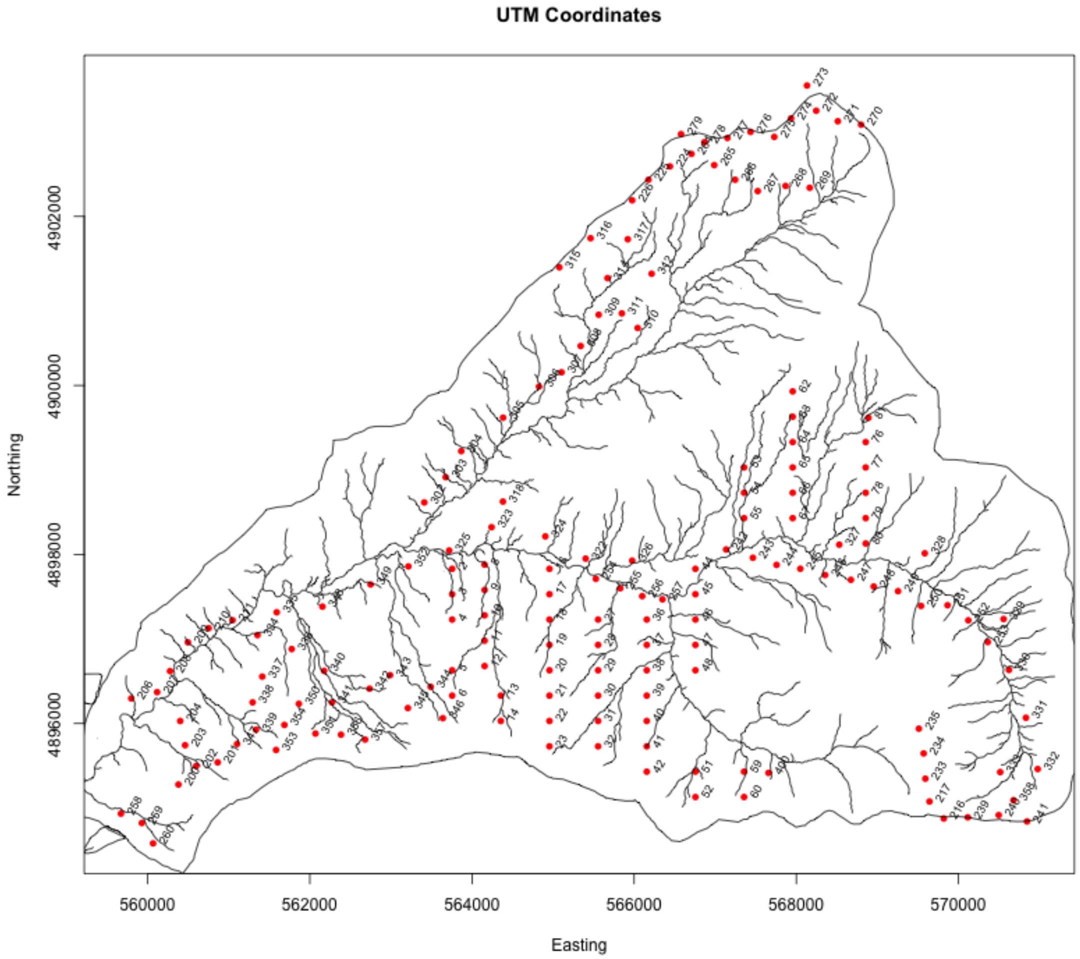

Figure 1.2: Map showing locations of around 160 sensors spread out over the HJA landscape. Map uses Universal Transverse Mercator (UTM) system instead of Latitude/Longitude to conform with the one used by the domain-expert.

nights regardless of season.

1.4 Lapse Rate and Inversions

The rate at which atmospheric temperature **decreases** with a rise in altitude is known as the lapse rate. It is typically around 5° F per 1,000 meters [13, 12] in dry weather. In

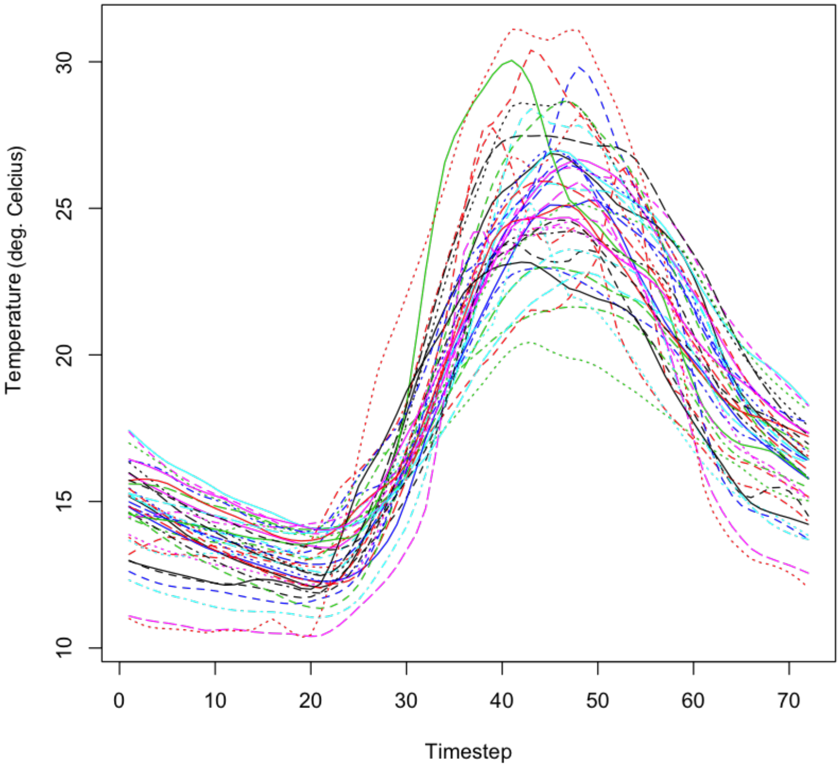

Figure 1.3: Multivariate Time Series for August 1st 2011, a summer day with almost no CAPs. 24-hours produce 72 time-steps, sampling at every 20 minutes. Each sensor's time-series is in a different color.

certain conditions, the temperature is known to increase with rise in altitude, and since this is the opposite of the typical behavior, it is called an “Inversion”.

Note that “lapse-rate” is a term associated with atmospheric (i.e. relatively high-altitude) phenomena. We reuse this terminology when we use the term lapse-rate in the context of rate of change of temperature using low-altitude (near surface) readings at different elevations (i.e. heights of the ground with respect to sea-level). This is not uncommon in the literature [13, 14, 18, 10]

The presence of CAPs and presence of Inversions are, by definition, correlated.

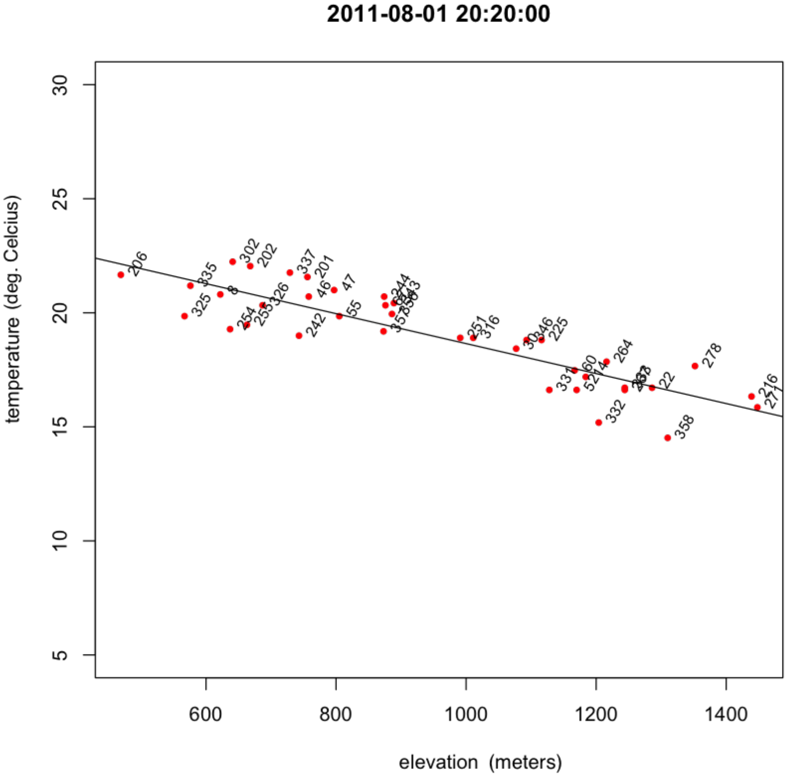

Figure 1.4: Temperature ($^\circ C$) vs. Elevation (meters above sea level) plot for a normal summer evening. Lapse Rate is $\approx -6.5^\circ C/1000$ meters. implying no inversions or CAPs

The rest of the manuscript is organized as follows: Chapter 2 describes our approaches to exploring and understanding the CAP phenomenon. Chapter 3 and 4 describe automated approaches to CAP detection and their performance on real-world data. Chapter 5 describes the design of a simulator and the performance of the detectors on synthetic datasets. Chapter 6 describes a graphical-model that we have created to explore further extensions of the aforementioned work, possibly incorporating more domain-knowledge. Chapter 7 concludes with a discussion of the work described and proposes directions for further exploration.

Chapter 2: Exploratory Data Analysis and Modeling Approaches

In this section we describe the three high-level approaches we took to exploring the presence of CAPs in our dataset, ranging from the *most-general and domain-agnostic, to the most domain-dependent.* This section also details the various visualization strategies we used to explore our multivariate time-series spatiotemporal dataset.

The first two approaches approach CAP detection problem as an **anomaly detection** problem: Dereszynski et al [5], who have previously studied the data streams received from the HJA, classify the anomalies possible in sensor-data into three degrees:

1. **Simple** Anomalies, which are readings outside the acceptable range of temperature and are introduced into the data-stream by the sensor itself e.g. a large negative value reading for a disconnected logger or an out-of-range voltage reading.

2. **Medium** Anomalies, which are associated with malfunctions in a sensor's hardware (e.g. a damaged sun-shield on a sensor results in positively biased readings during the day) or a change in a sensor's functionality (e.g. a sensor gets temporarily buried by snow).

3. **Complex** Anomalies, which can only be detected by considering *multiple sensors.*

In their paper, the authors address the first two types of anomalies in individual sensor readings using a Dynamic Bayesian Network model, but not the third type (Complex Anomalies, spanning multiple sensors). When viewed as an anomaly detection problem, the detection of CAPs necessarily requires multiple sensors to be considered as CAPs often cover large spatial expanses.

2.1 Approach 1: Time-series Changepoint Detection

2.1.0.1 Pairwise analysis

Since CAPs are associated with Inversions, the simplest form of changepoint to consider is the location of the change of the *covariance* of two adjacent sensors' time-series. Manual inspection of pairs of sensors on the same slope but having different elevations indeed revealed several such changepoints (see example in Fig. 2.1)

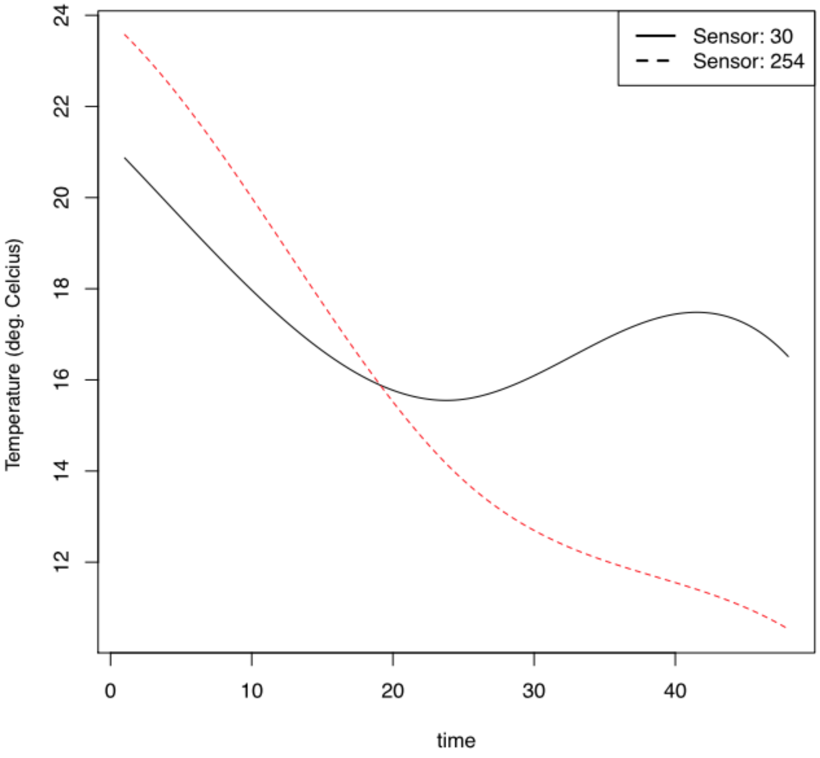

Figure 2.1: Nighttime cooling-curves (temperature over time) for a pair of sensors on the same slope. Sensor 254 is down-slope compared to Sensor 30, and starts off warmer as is expected from a normal lapse-rate. Later, after around timestep 20, the relationship between the temperature of the two sensors flips, due to cold air pooling at lower elevations. Curves were obtained by using a smoothed fit to a spline basis using the R package *fda* [20]

2.1.0.2 Multivariate analysis

If we visually inspect the (multivariate) time-series plots for a summer day with no (or minor night-time) CAPs (see figure 1.3), and compare them with a winter day with heavy persistent Cold Air Pooling (see figure 1.3), we can easily see differences in both the mean and the variance between time-series with normal and inverted behavior.

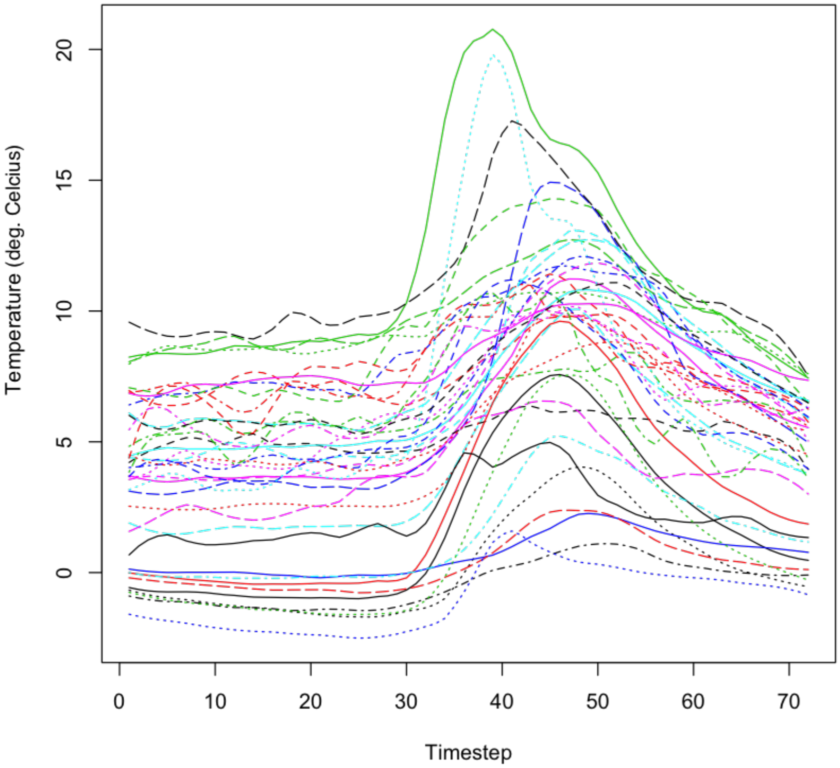

Figure 2.2: Multivariate Time Series for December 9th 2011, a winter day with heavy CAPs. Compare with equivalent plot for August 1st 2011, figure: 1.3

We experimented with automatic changepoint-detection algorithms for both pairwise and multivariate approaches. However, due to the widely-varying shapes and noise on individual sensor time-series, both approaches suffered the problem of excessive False Positives. Additionally, typical multivariate time-series changepoint detection algorithms [22] only detect changepoints in time, and do not easily support localizing changes to specific time-series components (i.e. sensors).

2.2 Approach 2: Low-dimensional manifold extraction

Since the data we are dealing with is from a natural physical phenomenon, and we have densely sampled it via our network of sensors, we suspect that it would be possible to find an underlying low-dimensional generative process that could explain the data. Additionally, it might be easier to separate 'normal' behavior from behavior involving CAPs by investigating such a low-rank representation.

To use a visual metaphor, one can think of the multivariate time-series data from the HJA to represent a temperature surface that oscillates over time. Normal behavior, i.e. without CAPs, implies a relatively 'smooth' surface with a 'regular' oscillation. Inversions and CAPs then imply that there 'bumps' of varying size that form on this oscillating surface.

We tried a variety of techniques for low-dimensional structure discovery, including Singular Value Decomposition (SVD) [21], Dynamic Mode Decomposition (DMD) [19] and Functional Principal Component Analysis (fPCA) [20]. As an example, the Singular Value Decomposition (SVD) of the multivariate time-series data indeed reveals (see Figures 2.3 & 2.4) such a low-rank structure. However, we were not able to see any clear separation between the so-called normal behavior and behavior involving CAPs, by examining the low-rank structure.

2.3 Approach 3: Lapse Rate Analysis

A recurring observation that came up during the aforementioned analyses, but has not been explicitly modeled yet, is the structure observed in the Temperature vs. Elevation relationship within the sensor readings. To investigate this further quantitatively, we fit linear regression models on all time-steps of several days where we expected few-to-no

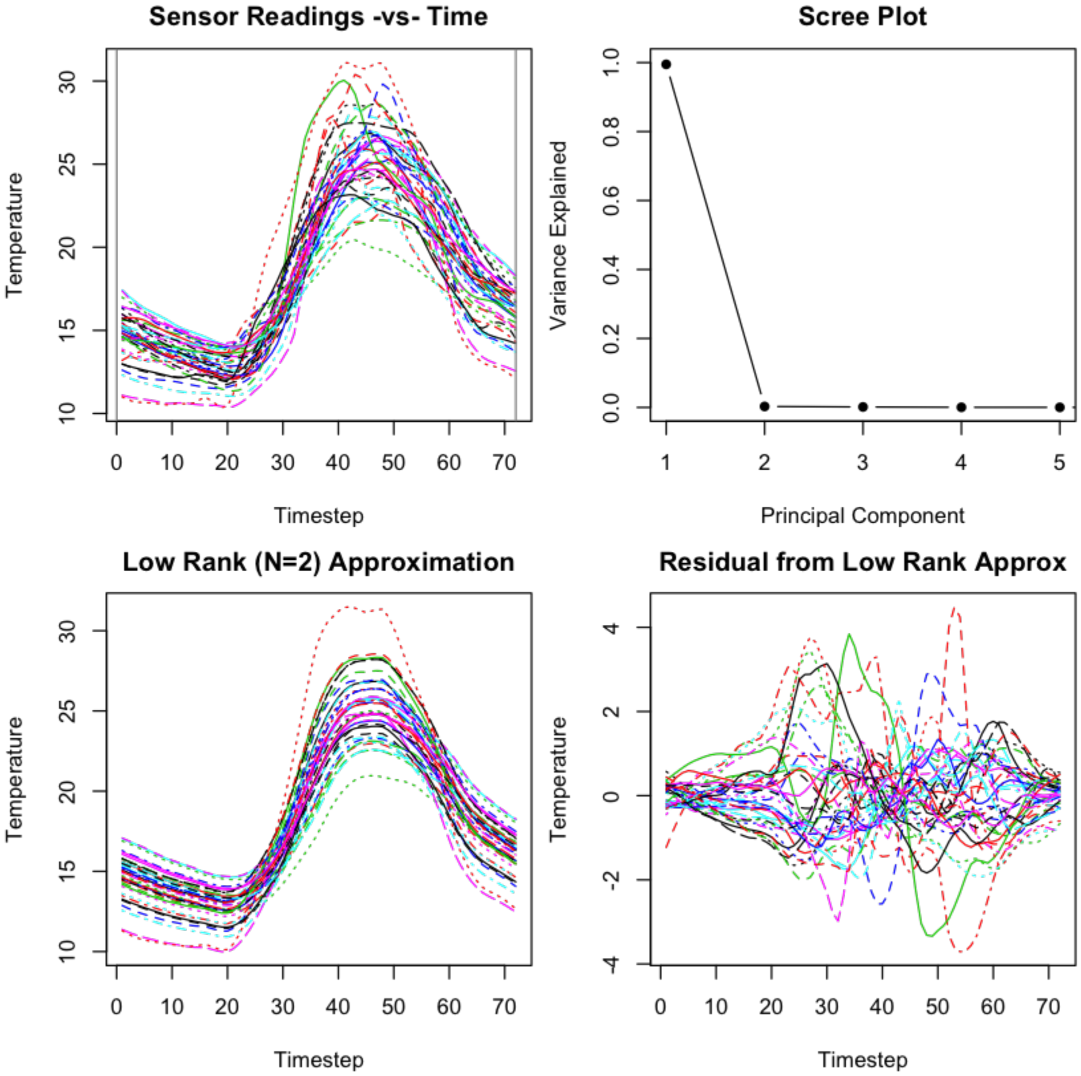

Figure 2.3: Separating Low-Rank Approximation from Residue for Aug. 1st 2011: (1) Top-left: Original Data, (2) Top-right: Scree plot of singular-values reveals low-rank structure (3) Bottom-left: Low-rank (N=2) Approximation and (4) Bottom-right: Residue after removing low-rank approximation from original.

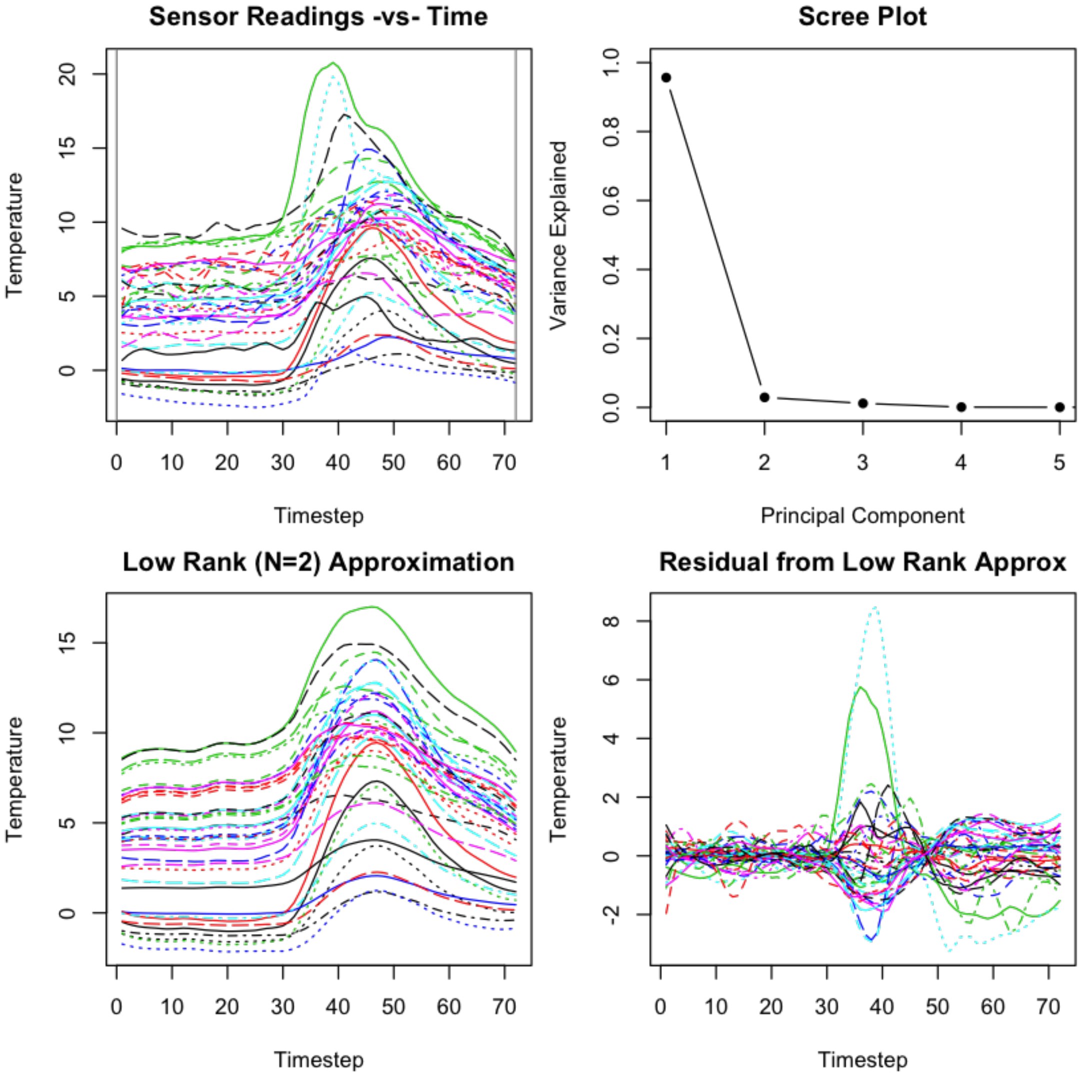

Figure 2.4: Separating Low-Rank Approximation from Residue for Dec. 9th 2011: (1) Top-left: Original Data, (2) Top-right: Scree plot of singular-values reveals low-rank structure (3) Bottom-left: Low-rank (N=2) Approximation and (4) Bottom-right: Residue after removing low-rank approximation from original.

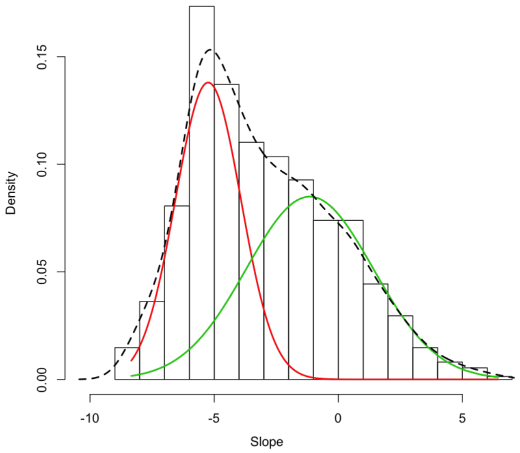

Figure 2.5: Two-component GMM fit to the distribution of lapse-rates for month of Aug. 2011. X-axis shows lapse-rate (slope) in units °C/1000-meters.

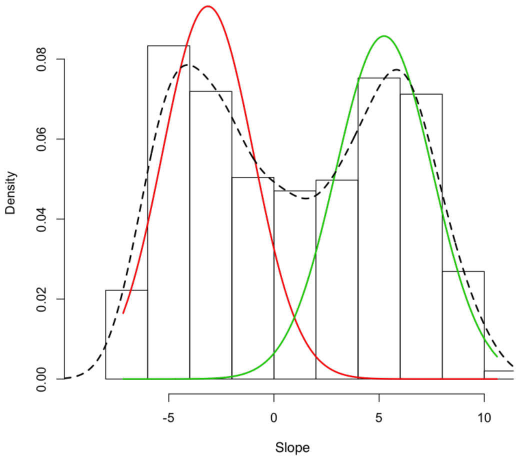

Figure 2.6: Two-component GMM fit to the distribution of lapse-rates for month of Dec. 2011. X-axis shows lapse-rate (slope) in units °C/1000-meters.

CAPs (August) and compared them with several days where CAPs are expected to be frequent (December). Fitting two-component Gaussian Mixture Models (GMMs) to the distribution of lapse rates (i.e. slope coefficient from regression fits) indeed revealed a clear difference between these two groups (see Figures 2.5 & 2.6).

This is the relationship that we will investigate further in the rest of this manuscript. Note that this is the *most domain-specific approach* among the three discussed, because it specifically takes into account the domain-knowledge that the Temperature vs Elevation relationship is important to the detection of CAPs.

Chapter 3: Linear Detector

3.1 A Naive Linear CAP Detector

Motivated by the observations noted in the last section, we further investigate the dynamics of the lapse rate by fitting a linear regression model:

$$\text{Temperature} = \beta_0 + \beta_1 * \text{Elevation} + \epsilon$$

for each time-step of the data and plotting the slope coefficient β_1 (i.e. lapse-rate) over time.

As shown in the figure 3.1, many days in August show brief periods of (night-time) inversion, while in December, only a few days are free of long periods of inversion.

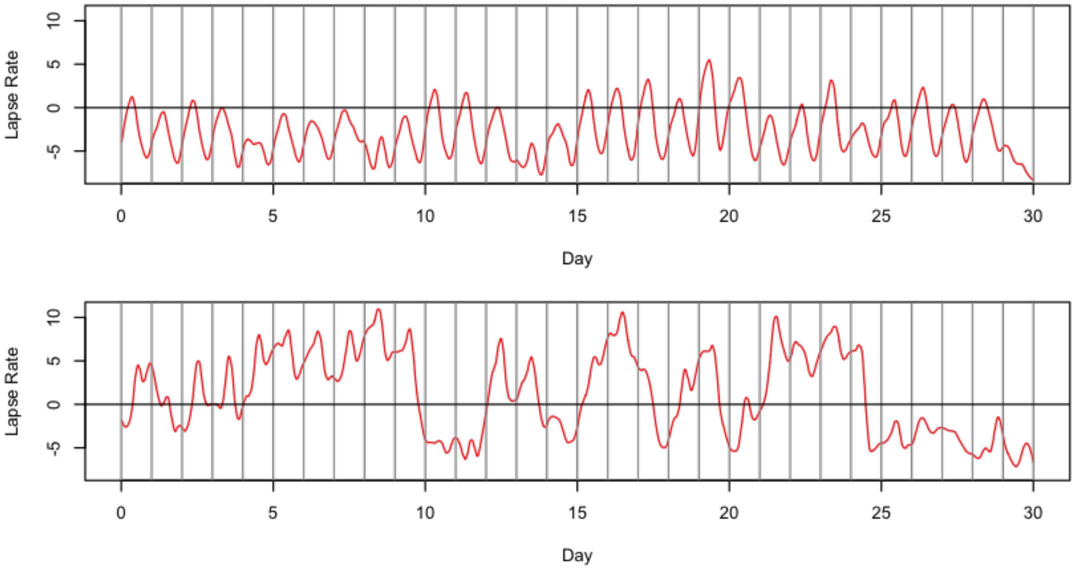

Figure 3.1: Lapse Rate Time Series for August (Above) and December (Below), reveals regular nightly CAPs in August, and persistent CAPs in December. Recall that positive lapse-rate values imply ‘inversions’.

As we know that inversions are correlated with CAPs, we now have a naive detector that can temporally localize CAPs. To verify that our detector is indeed working as it should, we sought out a qualitative assessment from our domain-expert.

3.2 Qualitative assessment

To facilitate the qualitative assessment, we created an **animation 'dashboard'** (see screenshot in Figure 3.2) that enabled a side-by-side evaluation, at each time-step, of (1) the linear detector and (2) a spatial visualization of the temperature surface of the HJA.

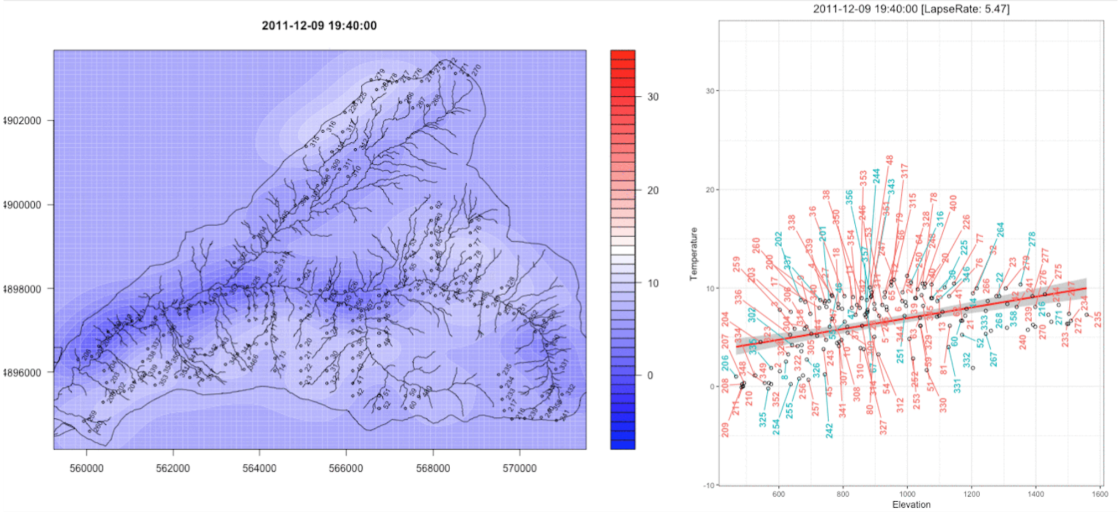

Figure 3.2: Dashboard for inspecting linear-detector fit. **Right panel** shows Temperature vs. Elevation plot with Linear Model. **Left Panel** shows the (interpolated) temperature surface over the HJA landscape (Northing × Easting Coordinates in UTM units). See *Supplemental Materials* 7.2 to access full-length animations.

The domain-expert found these animations useful and provided us with feedback that we could incorporate into the next version of the detector. In addition to being able to use the lapse-rate time-series to precisely localize in time the global state (inverted or not) of the HJA, she was also able to (visually, applying her own expert-heuristic) localize the CAP to specific sensors near the main stem of the stream (the darker-blue regions in Figure 3.2).

Chapter 4: Nonlinear Detectors: Quadratic + Piecewise

The Linear Detector allows us to localize a CAP in time but not in space. To localize a CAP in space (i.e. to specific sensors), we further analyze the relationship between Temperature and Elevation by considering two non-linear regression models.

4.1 Non-linear Models

4.1.1 Quadratic Model

The quadratic model is a simply a quadratic-regression [11] model:

$$\text{Temperature} = \beta_0 + \beta_1 * \text{Elevation} + \beta_2 * \text{Elevation}^2 + \epsilon$$

The addition of the extra (β_2) term allows for a non-linear curve, thereby improving the fit to the data in some cases (see Figure 4.1 for an example).

Additionally, we can calculate the location of the **inflection-point** of the curve as:

$$\begin{aligned} \frac{d\text{Temperature}}{d\text{Elevation}} &= \beta_1 + 2 * \beta_2 * \text{InflectionElevation} = 0 \\ \text{InflectionElevation} &= -\beta_1/2\beta_2 \end{aligned}$$

When the quadratic model estimates a **concave-downward** curve, the inflection-point separates an upward-sloping curve (to its left) from a downward-sloping curve (to its right). This can be thought of as approximating two connected linear models, having arisen from an inversion in only a part of the landscape. The upshot of this is that the inflection-point can be thought of as **estimating the height of the CAP**. The CAP can now be **spatially localized** to sensors with elevations below the inflection-point.

There are also situations in which the quadratic model estimates a concave-upward curve, where the inflection-point separates an downward-sloping curve (to its left) from a

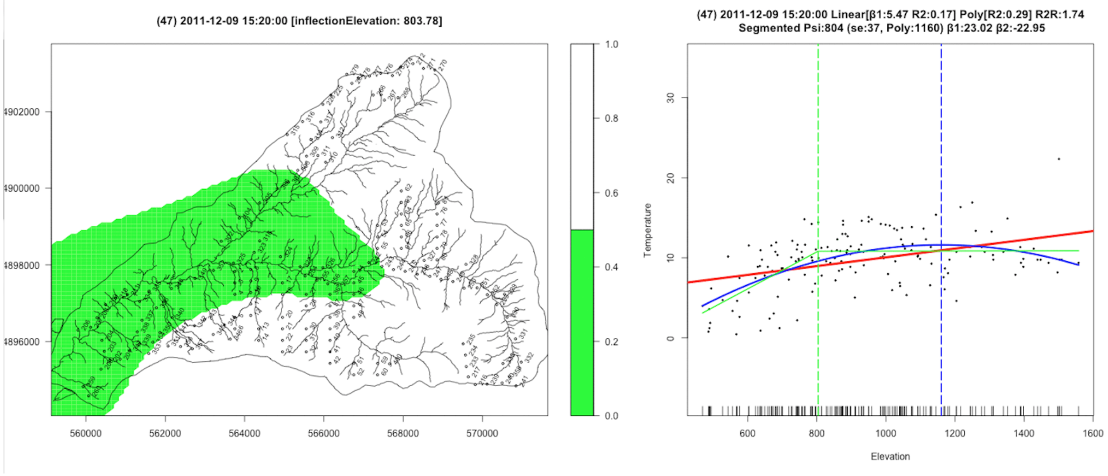

Figure 4.1: Dashboard for Nonlinear Detectors: **Right panel:** Temperature vs. Elevation plot, with Linear-detector (**solid-red**), Quadratic-detector (**solid-blue**) and Piecewise-Linear/Segmented detector (**solid-green**). **Dashed vertical lines** indicate location of inflection-point/breakpoint for Quadratic and Piecewise-Linear/Segmented detectors (respectively), and are colored when model is chosen by the hypothesis test, and grayed-out otherwise. **Left panel:** All sensors to the left of the inflection-point/breakpoint are regarded as in a CAP. Left panel shows this region (**green**) on the HJA map corresponding to the Piecewise-Linear/Segmented detector. (Northing × Easting Coordinates in UTM units). See *Supplemental Materials* 7.2 to access full-length animations.

upward-sloping curve (to its right). Such situations, as far as we know, do not correspond to a physical phenomenon, and are ignored in our analysis.

4.1.2 Piecewise Linear Model (or Segmented Model)

To directly model the data as a pair of linear-models connected at an inflection-point, we can use the Piecewise Linear Model (PLM) a.k.a. the Segmented Model.

We use the ‘segmented’ R package [16] to obtain our Piecewise Linear fits, as given by the model:

$$y_i = \beta_1 x_i + \beta_2 (x_i - \psi)_+$$

where $(x_i - \psi)_+ = (x_i - \psi) \times I(x_i > \psi)$ and $I(\cdot)$ is the indicator function. β_1 is the slope of the left arm, and β_2 is the slope of the right-arm, and ψ is the break-point (inflection-point).

As described in [15], an analytic solution is no longer possible, and so the resulting non-linear objective function is approximated via a linearization around ψ, and fitted via an iterative algorithm.

4.2 Model Selection

In many cases, the fit produced by the non-linear model is superior (in R^2 terms) to that produced by the linear model, and we should clearly choose the non-linear model. In some cases, the fit produced by the non-linear model and the linear model are similar (or the non-linear model has produced a degenerate fit, e.g. in figure 4.2). In such cases, by Occam's Razor, we prefer the simpler model. To make this decision in a principled manner, we perform model selection using a hypothesis test.

4.2.1 Linear model vs. Quadratic Model

The *ANOVA (F-test) for nested models* [11, p. 116] is applicable here. The Null Hypothesis is that the linear and quadratic models fit the data equally well, and the Alternative Hypothesis is that the quadratic model is superior.

4.2.2 Linear model vs. Piecewise-Linear Model

The *Chow test* [1], which is commonly used in econometrics literature to test for structural breakpoints, is applicable here. It tests whether the true coefficients of two different linear regressions on different data-sets are actually equal. In our case, we are interested in testing if β_{L1} is the same as β_{R1}, where

- Temperature $= \beta_{L0} + \beta_{L1} *$ Elevation $+ \epsilon$, is the model fit to data to the left of the InflectionElevation, and
- Temperature $= \beta_{R0} + \beta_{R1} *$ Elevation $+ \epsilon$, is the model fit to data to the right of the InflectionElevation

Specifically, the Null Hypothesis is that $\beta_{L1} = \beta_{R1}$, i.e. the simpler (linear) model is

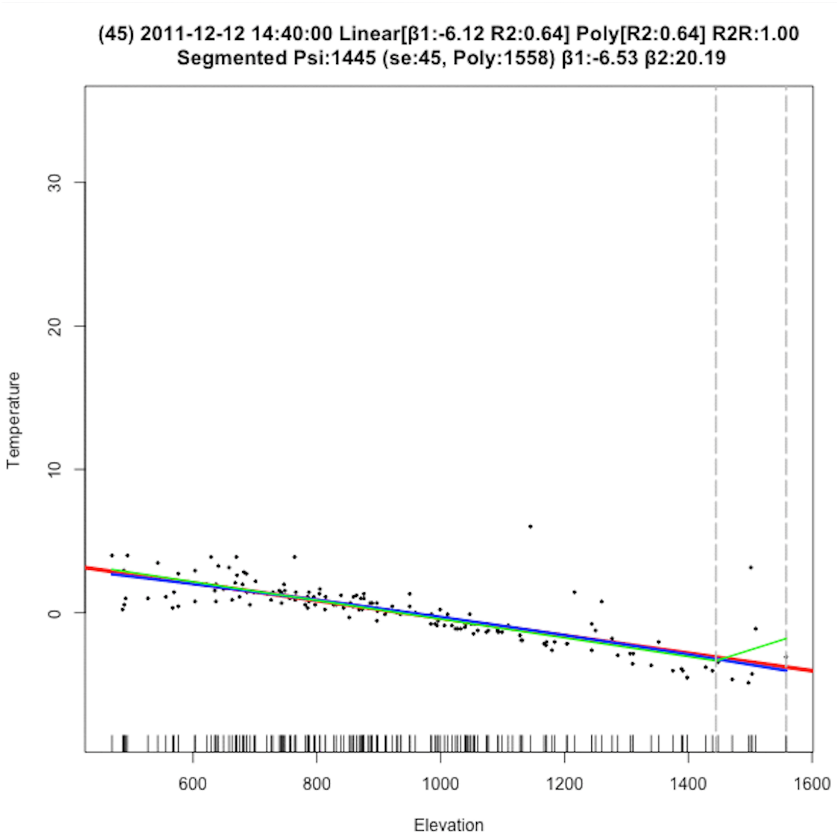

Figure 4.2: An example of a case when the data is not particularly non-linear in structure, and leads to a particularly pathological PLM fit. The vertical dashed-line associated with with both non-linear models have been grayed-out to indicate that the alternative hypothesis (Non-linear Model) was not accepted by the hypothesis test.

chosen, and the Alternative Hypothesis is that $\beta_{L1} \neq \beta_{R1}$, i.e. the Piecewise-Linear Model is more appropriate.

4.2.3 Multiple Hypothesis Error Correction

Since we carry out one test per time-step for each of 72 timesteps in a day, we need [9, p. 686] to apply a correction to account for multiple tests.

The p-value threshold (typically 0.05) that is used to decide when the the linear model is rejected over a more complex model (Quadratic or PLM) is made more stringent via the Bonferroni correction [9, p. 686], and set to 0.05/72.

4.3 Qualitative assessment

To facilitate qualitative assessment of the non-linear models, the **non-linear 'dashboard'** (screenshot in Figure 4.1), and an **auxiliary time-series** (example in figure 4.3) were provided to the domain-expert.

Both nonlinear models seem to fit the data better (in R^2 terms) than the linear model, during periods of suspected CAPs. In August, when CAPs are known to occur only on cool nights, the linear model dominates most of the day. In December, when CAPs are known to persist throughout the day, the more complex models get selected by the hypothesis test for most of the day.

Between the two linear models, it seems like the PLM provides a better estimate of the inflection-point (CAP elevation), while the Quadratic model's inflection point tends to often over-estimate it.

Non-linear models achieve the goal of localizing CAPs in both space and time. The shape of the region indicated as under a CAP in the non-linear dashboard (e.g. figure 4.1), is consistent with the topology of the HJA (figure 1.1) that indicates that the lowest elevations are towards the bottom-left of the map, and extend along the river.

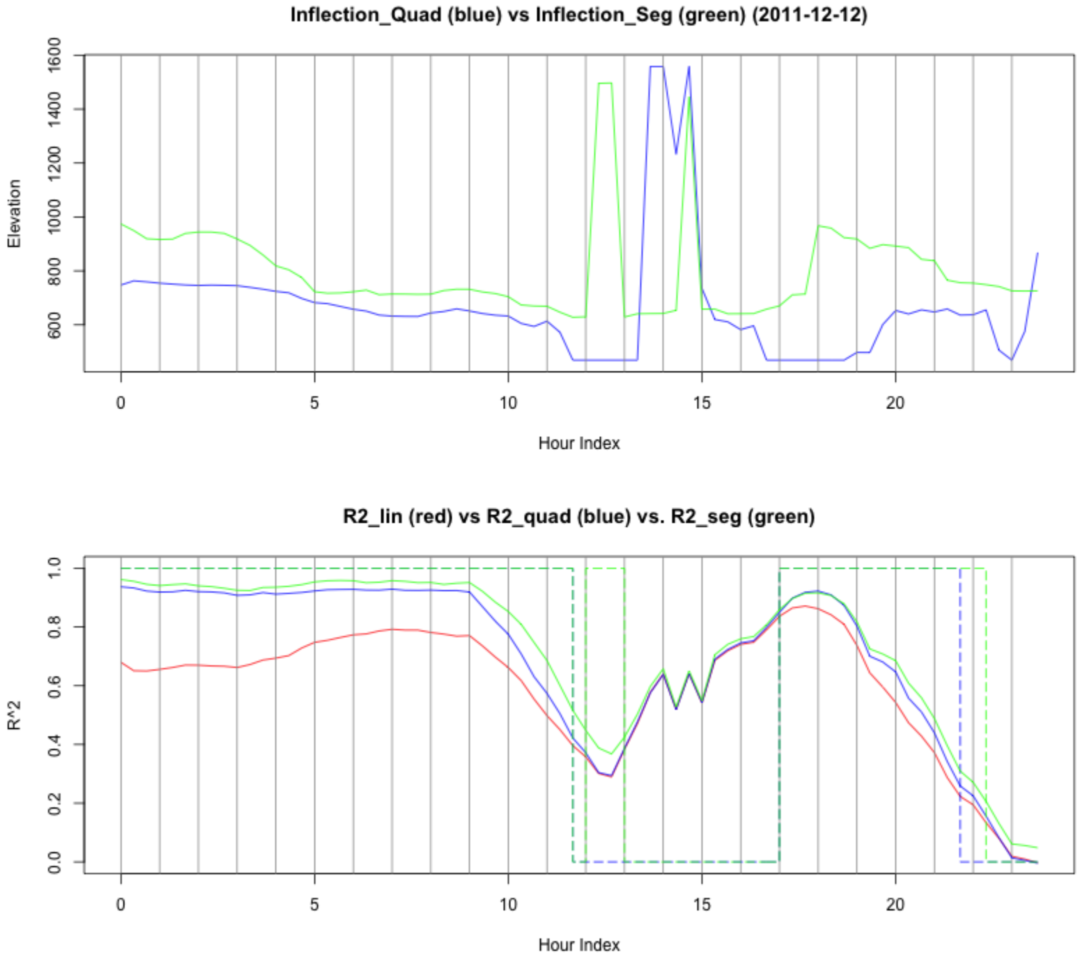

Figure 4.3: **Auxillary time-series**: **Lower panel:** R^2 time series for Linear (**solid-red**), Quadratic (**solid-blue**) and PLM/Segmented (**solid-green**) detectors. Overlayed **dashed step-function** curves, when high, indicate when the respective detector was chosen by the (multiple-testing corrected) hypothesis test. **Upper Panel:** Inflection-point (CAP elevation) estimates from the Quadratic (**solid-blue**) and PLM/Segmented (**solid-green**) detectors.

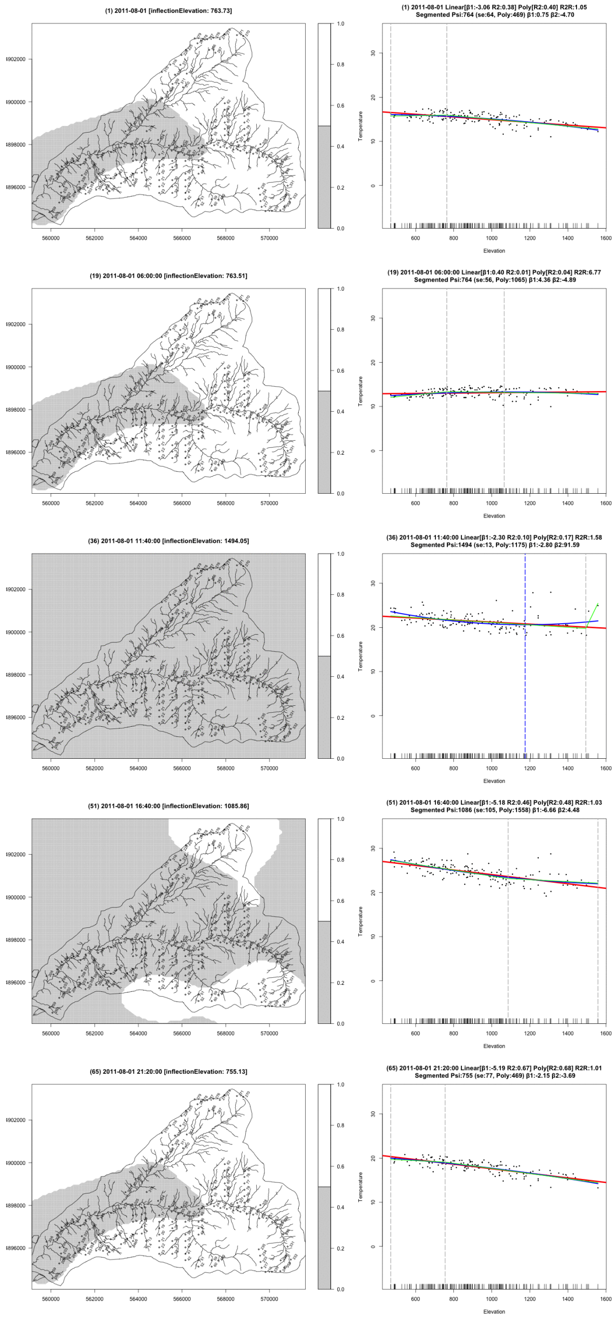

Figure 4.4: Dashboard animation frames from August 1st 2011, a day with normal "summer" behavior (with a minor night-time CAP). Frames from top to bottom are for (1) 12:00 midnight, (2) 6:00 am, (3) 11:40 am, (4) 4:40pm and (5) 9:20pm. Full animations available in *Supplemental Materials*. Grayed-out area on the left panels indicate that the Segmented detector did not pass hypothesis test for that timestep. Similarly, grayed-out dashed vertical lines indicate respective nonlinear classifier did not pass hypothesis test for that timestep.

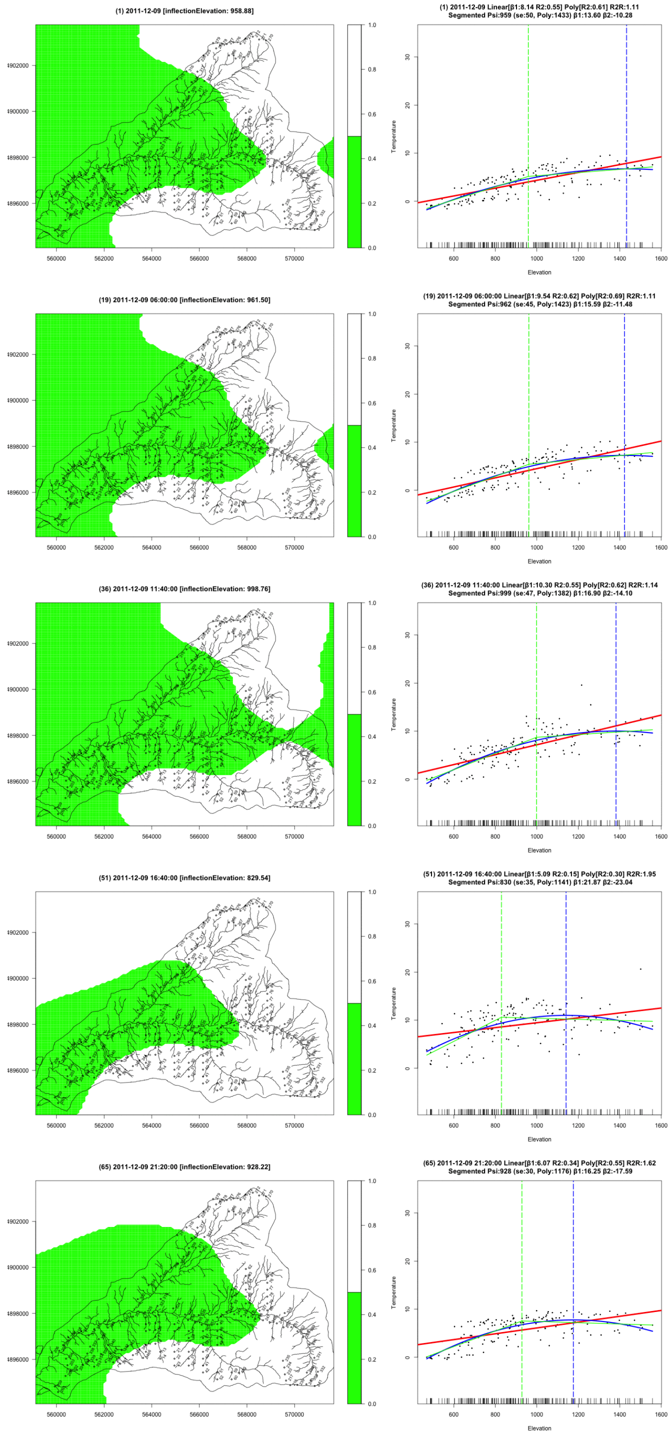

Figure 4.5: Dashboard animation frames from December 9th 2011, a "winter" day with heavy cloud-cover and persistent inversion and CAPs. Frames from top to bottom are for (1) 12:00 midnight, (2) 6:00 am, (3) 11:40 am, (4) 4:40pm and (5) 9:20pm. Full animations available in *Supplemental Materials*

Chapter 5: Simulation Study

To evaluate the performance of the various CAP detectors, we test their performance on synthetic data by (1) varying input-noise conditions and (2) varying the elevation of the inflection-point.

5.1 Simulator Design

To generate synthetic data, we wrote a simulator that met the following criteria:

- A (Temperature) time-series is created for each sensor that is simulated and bears the same **sinusoid-like shape** found in real data.
- An adjustable amount of normally distributed random **noise** is added to the **amplitude** at each time-step. A small amount of random noise is also added to the **phase** of the sinusoid (does not vary per time-step), and was manually adjusted to make the output look like a real sensor's output.
- The **number** of time-series generated is adjustable, and set equal to the number of sensors (about 160) being studied.
- The real-data values for the **covariates** (Elevation, Latitude & Longitude) associated with each sensor are used for the synthetic data.
- The mean-value for a simulated time-series is adjusted such that it is proportional to a **user-set lapse-rate** (plus a small amount of random normal noise).
- Since the lapse-rate itself varies with time, an adjustable magnitude **lapse-rate time-series** can be superimposed on the data. The **shape** of this superimposable time-series can either be (1) **Gaussian** or (2) **Sinusoid**, both with adjustable peak-magnitude and peak-location.
- In addition to being able to superimpose a lapse-rate time-series on top of the data, we can choose an elevation (**simulated inflection-point**) above which the lapse-rate adjustment will not be applied.

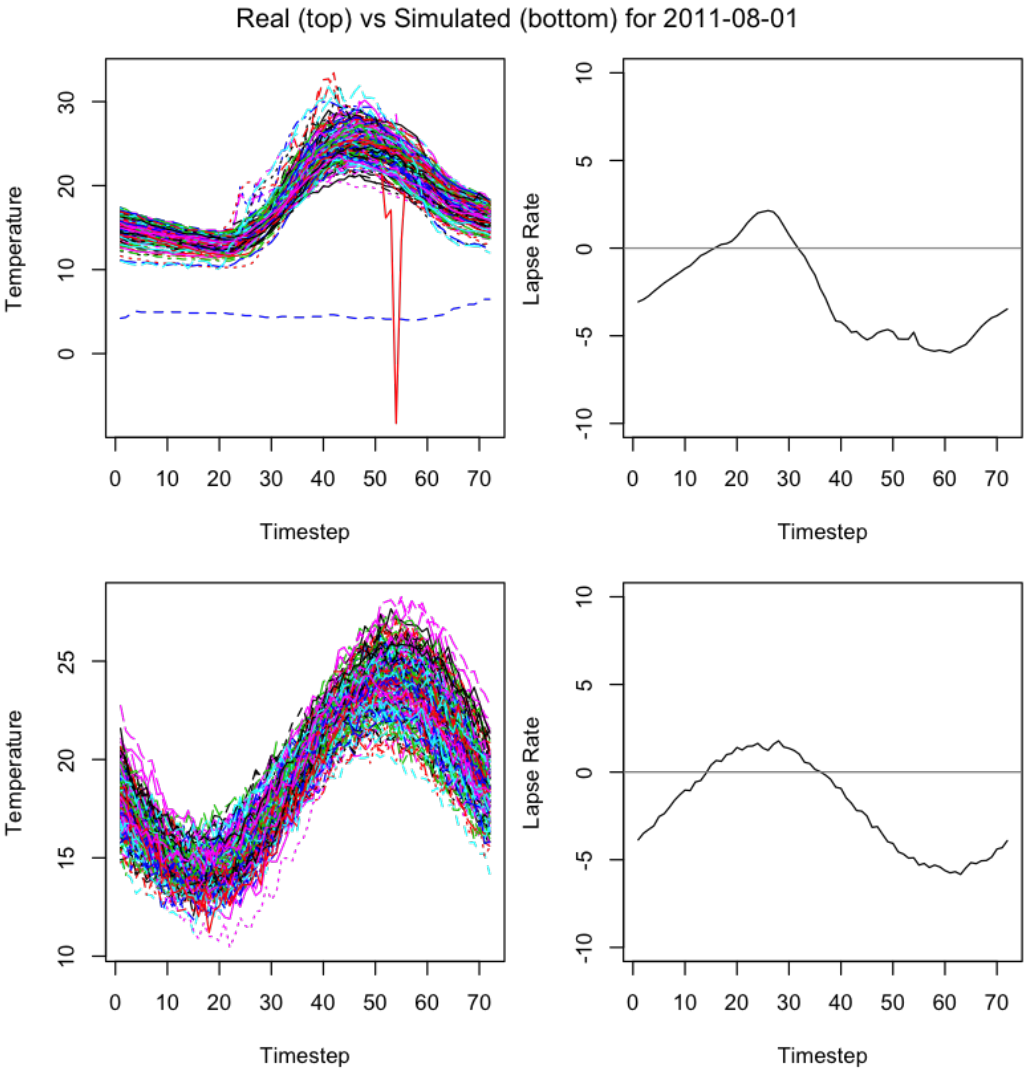

Figure 5.1: **Synthetic Data Example 1: Top Left:** Real data Multivariate time-series for August 2011-08-01, a typical summer day. **Top right:** Lapse-rate time series showing a short period of early-morning cold air pooling (time-steps 25-35). **Bottom Row:** Bottom left and right panels show the equivalent time-series and lapse-rate plots using simulated-data. No attempt was made to replicate the (rare) outliers in real-data.

Two examples of synthetic-data produced by the simulator is shown in Figures 5.1 and 5.2. These synthetic-data examples were generated in such a way so as to match real-data from different days, in order to show the flexibility of the simulator to model different microclimatic conditions.

5.2 Experiment: Detection Robustness

We would like to know how well the detectors can detect CAPs for varying levels of input (amplitude) noise. Synthetic Input noise, in the form of additive random normal noise, is varied from 0% to 20% (”error-fraction” in figure 5.3) of the amplitude dynamic-range of the noiseless baseline curve. F1-scores were calculated by comparing against the linear detector with no noise added, and averaging over 35 randomly-initialized trials.

Detection of an inversion is done as follows for each detector:

- **Linear**: Inversion detected when $\beta_1 >= 0$
- **Quadratic**: Inversion detected when $\beta_1 >= 0$
- **Piecewise Linear/Segmented**: Inversion detected when $\beta_L >= 0$

As shown in figure 5.3, we found that the Linear and Quadratic detectors were quite robust to added noise. The Piecewise-Linear/Segmented detector performed consistently poorly because of its tendency towards degenerate fits when the data is inherently mostly linear (similar to the one shown in figure 4.2).

Hybrid-detectors: We expect that the performance of the non-linear detectors can be made more stable by using them in combination with the linear detector. We introduce two Hybrid-detectors that implement this reasoning: (1) Linear+Quadratic (2) Linear+Segmented.

For the Hybrid-detectors, the non-linear detector is only consulted if two conditions pass: (1) the linear-detector indicates inversion (to filter out concave-upward fits) and (2) the hypothesis test passes i.e. indicates that the data is significantly non-linear to justify a more complex model. As can be seen in figure 5.3, the performance-curves for the hybrid-detectors basically hug the performance-curve for the linear-detector.

5.3 Experiment: Inflection Point Estimation

We would like to know how well the non-linear detectors estimate the location of the inflection-point. This would additionally help quantify any systematic biases that the detectors might have in terms of under or over-estimating the inflection point.

We generate synthetic-data, similar to that in figure 5.2, and vary its inflection point from 600-meters to 1300-meters. RMSE and Mean Error values are recorded and averaged over 35 trials to summarize the overall-error and any systematic-bias respectively. Note that the Quadratic detector can generate inflection-point locations that are outside the range of sensor-elevations and thus has been trimmed to be within that range.

As shown in Figure 5.4, the Piecewise-Linear/Segmented detector consistently estimates the true inflection-elevation with small error.

The Quadratic detector, consistently over-estimates the true inflection-elevation. This is consistent with expectations, and follows directly from the geometry of the resulting curve when a inflection-point is introduced by artifically lowering the temperatures of simulated sensors below a certain elevation.

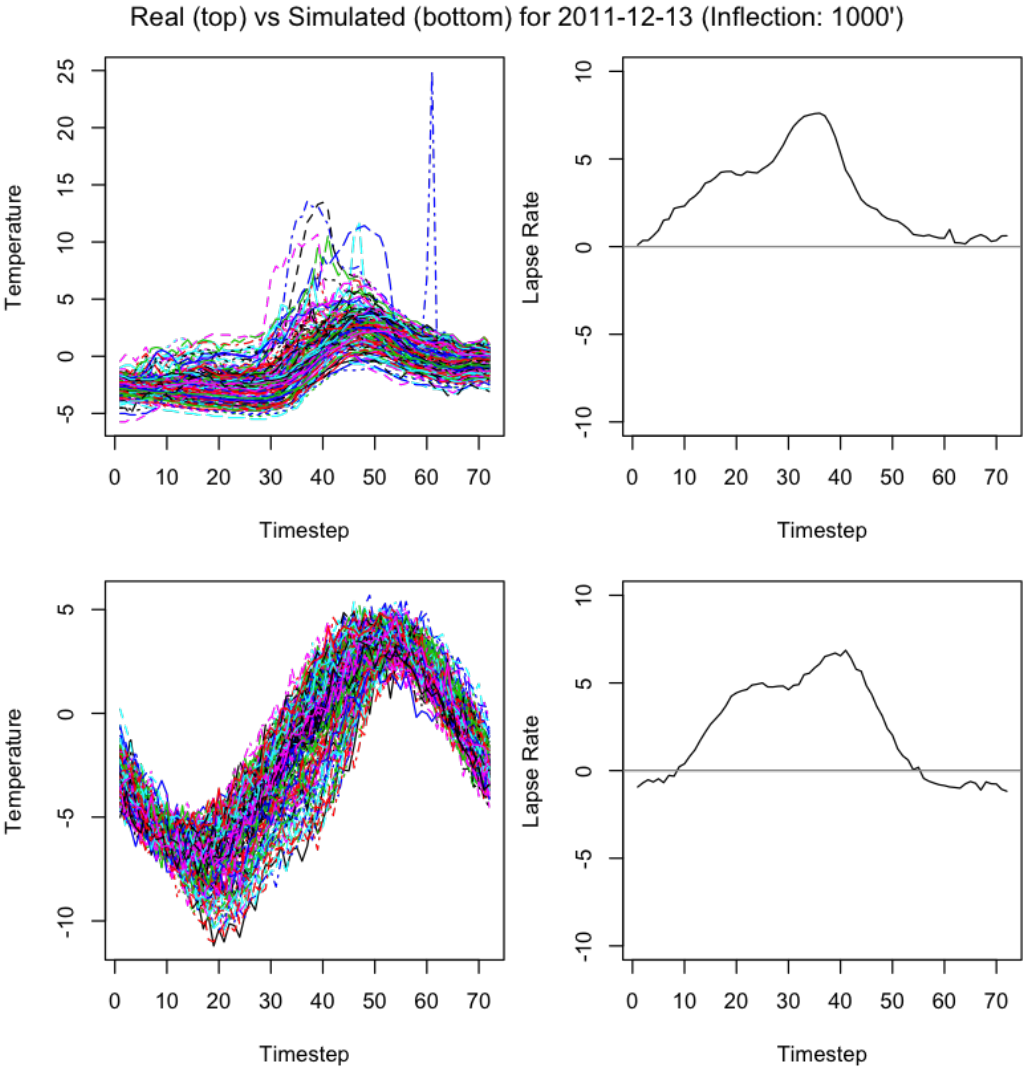

Figure 5.2: **Synthetic Data Example 2: Top Left:** Real data Multivariate time-series for December 2011-12-13, a winter day with heavy cold air pooling. **Top right:** Lapse-rate time series showing persistent cold air pooling. **Bottom Row:** Bottom left and right panels show the equivalent time-series and lapse-rate plots using simulated-data. This simulation required repeated superimposition of a Gaussian-shaped lapse-rate time-series. Note that an adjustable inflection-point was introduced at 1000-ft. No attempt was made to replicate the (rare) outliers in real-data.

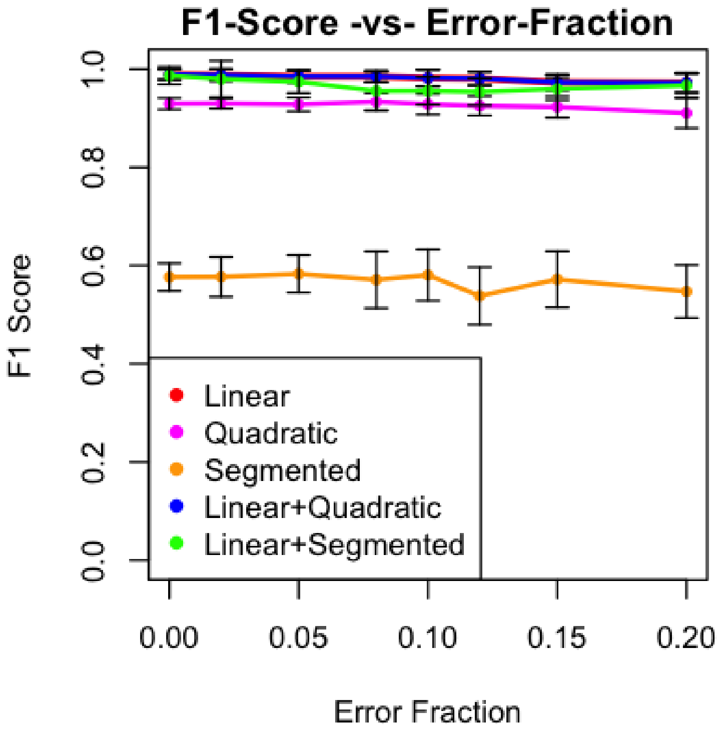

Figure 5.3: Results of experiment to compare detection of inversions: *Error-fraction* denotes the amount of synthetic Input noise, in the form of additive random normal noise, ranging from 0% to 20% of the amplitude dynamic-range of the noiseless baseline curve. *F1-scores* were calculated by comparing against the linear detector with no noise added, and were averaged over 35 randomly-initialized trials. Linear and Quadratic detectors, being very similar in nature, perform equally well. Piecewise-Linear/Segmented detector performed consistently poorly because of its tendency towards degenerate fits (e.g. in figure 4.2) when the data is inherently mostly linear. The curves for the hybrid-detectors almost overlap the linear-detector's curve.

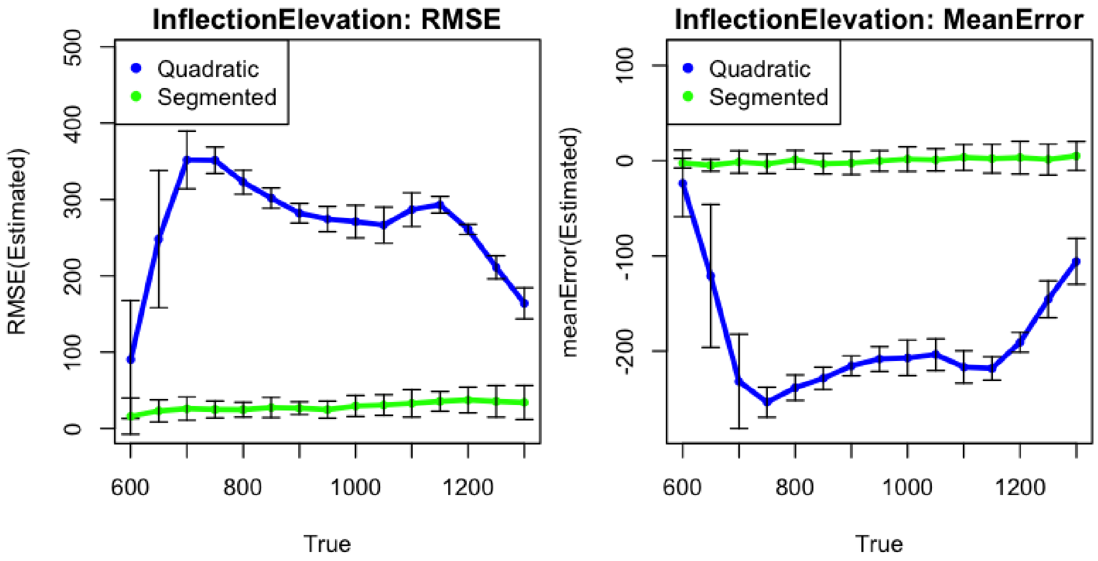

Figure 5.4: Results of experiment to compare estimation of inflection-elevation: The Piecewise-Linear/Segmented detector consistently estimates the true inflection-elevation with small error, while the Quadratic detector, consistently over-estimates the true inflection-elevation. Note that the Quadratic detector can generate inflection-point locations that are outside the range of sensor-elevations and thus has been trimmed to be within that range. This trimming explains the reduction in the error close to the edges, for the Quadratic-detector

Chapter 6: Nonlinear Detectors: Probabilistic Graphical Model

We would like to explore further extensions of the nonlinear detectors that model the data better, possibly by incorporating more domain knowledge. With this in mind, we have created a probabilistic graphical model for Piecewise-Linear Regression, that could serve as a starting-point for future extensions.

6.1 Graphical Model

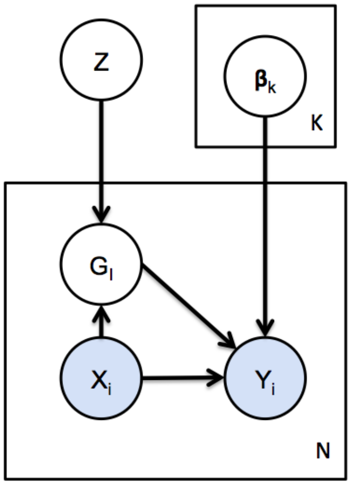

Figure 6.1: Plate Diagram for Graphical Model:

Figure 6.1 shows a plate-diagram for the graphical-model. Z represents the latent-variable for the inflection-point, and can take one of $z_1, \cdots, z_K$ values. $\boldsymbol{\beta_k}$ represents the vector of regression coefficients for two linear regressions, one to each side of the inflection-point. There are K such vectors, one associated with each of K values that the inflection-point latent-variable Z can take. X_i and Y_i represent Elevation and Temperature readings respectively, for each of $i = 1 \cdots N$ observation-pairs. G_i is a binaryvariable that indicates the assignment of this observation-pair to either the left-hand ($G_i = 0$) or the right-hand ($G_i = 1$) linear-regression.

The model assumes that the observed-data originates from the following generative-process: First, we select an inflection-elevation $Z = z_k$, by drawing from a multinomial distribution with parameter π. The expected-value (Temperature) of an observation with elevation lesser than (or equal to) the inflection-elevation is $\boldsymbol{\beta_{L,k}} X_i I[G_i = 0]$ while if higher than the inflection-elevation is $\boldsymbol{\beta_{R,k}} X_i I[G_i = 1]$. Finally, the observed Temperature (Y_i) is generated by drawing from a Gaussian distribution with mean $\boldsymbol{\beta_{j,k}} X_i I[G_i = g_j]$ and variance σ_k^2, where $j \in \{0,1\}$ indexing the left-hand and right-hand regressions respectively. We assume that σ_k is a common variance parameter between the left-hand and right-hand regressions for a given inflection-point z_k.

The complete-data log-likelihood corresponding to the above generative-process is given by the following equation:

$$\begin{aligned}\mathcal{L}_c(\theta) = \sum_{k=1}^{K} \Bigg[I[Z = z_k] \sum_{i=1}^{N} \Bigg(\sum_{j \in \{0,1\}} \Bigg(& I[G_i = g_j] logP(Y_i | X_i, G_i = g_j, Z = z_k, \theta) \\ & + I[G_i = g_j] logP(G_i = g_j | X_i, Z = z_k) \Bigg) + logP(X_i) \Bigg) \\ & + I[Z = z_k] logP(Z = z_k | \theta) \Bigg]\end{aligned}$$

6.2 Inference

We use the Expectation-Maximization (EM) [4] algorithm to estimate the model parameters ($\theta = \{\pi, \beta, \sigma\}$). The $P(G_i = g_j | X_i, Z = z_k)$ term is actually deterministic, and is simply equal to $I[X_i > z_k]$. The EM algorithm iterates between the E-step and M-step till a convergence condition has been satisfied.

The E-step comprises of calculating the expected complete-data log-likelihood function $Q(\boldsymbol{\theta}, \boldsymbol{\theta'})$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \mathbb{E}_{\boldsymbol{G}, Z | \boldsymbol{X}, \boldsymbol{Y}} \mathcal{L}_c$$

where θ and θ' are the current iteration and previous iteration's parameter values. In this process, we calculate two quantities, $\gamma_{i,j,k}$ and δ_k, defined below:

$$\begin{aligned}\gamma_{i,j,k} &= P(Z = z_k, G_i = g_j|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta'}) = \frac{P(Z = z_k, G_i = g_j, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'})}{\sum_Z \sum_{\boldsymbol{G}} P(Z, \boldsymbol{G}, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'})} \\ \delta_k &= P(Z = z_k|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta'}) = \sum_{i,j} \gamma_{i,j,k}\end{aligned}$$

In the M-Step, we estimate the parameter values using the expected-values of the latent-variables from the E-step. Given $Q(\theta, \theta^{t-1})$ from the E-Step, the M-step calculation is:

$$\boldsymbol{\theta^{t+1}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ Q(\boldsymbol{\theta}, \boldsymbol{\theta^t})$$

The complete derivation of both the E- and M-steps is provided in the Appendix B. We summarize the parameter estimators below:

$$\begin{aligned}\pi_k &= \frac{\delta_k}{\sum_{k=1}^K \delta_k} \\ \beta_{0_{j,k}} &= \frac{\sum_{i=1}^{N_j} \gamma_{i,j,k}(Y_i - \beta_{1_{j,k}} X_i)}{\sum_{i=1}^{N_j} \gamma_{i,j,k}} \\ \beta_{1_{j,k}} &= \frac{\sum_{i=1}^{N_j} \gamma_{i,j,k} X_i (Y_i - \beta_{0_{j,k}})}{\sum_{i=1}^{N_j} \gamma_{i,j,k}(X_i^2)} \\ \sigma_k^2 &= \frac{\sum_{j\in\{0,1\}} \sum_{i=1}^{N_j} \gamma_{i,j,k}(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}} X_i)^2}{\sum_{j\in\{0,1\}} \sum_{i=1}^{N_j} \gamma_{i,j,k}}\end{aligned}$$

6.3 Nonlinear Detector

To use the graphical model as non-linear detector, the inflection point is obtained by getting the *MAP estimate* of the Z variable, while the $\boldsymbol{\beta_k}$ associated with that inflection-point provide the estimates of the slopes for the left-hand and right-hand side regressions.

To initialize the algorithm, we choose $z_1, \cdots, z_K$ to correspond to 21 equally spaced quantiles of the sensor elevations, and set $\pi_k = 1/K$. $\boldsymbol{\beta}s$ and σs are initialized with coefficients from linear-regressions fit with an assumption of a median-elevation inflection-point.

For convergence, we iterate the EM-algorithm till the change in the $Q(\boldsymbol{\theta^t}, \boldsymbol{\theta^{t-1}})$

function becomes lesser than an arbitrary small-threshold ϵ. In practice, this was almost always less than 20 iterations.

6.4 Quantitative Assessment

We update the results from the simulation-study in Chapter 5, by re-running the experiments for 35 trials each. As shown in experiment results, figures 6.2 and 6.3, the performance of the graphical-model is close to that of the Segmented detector.

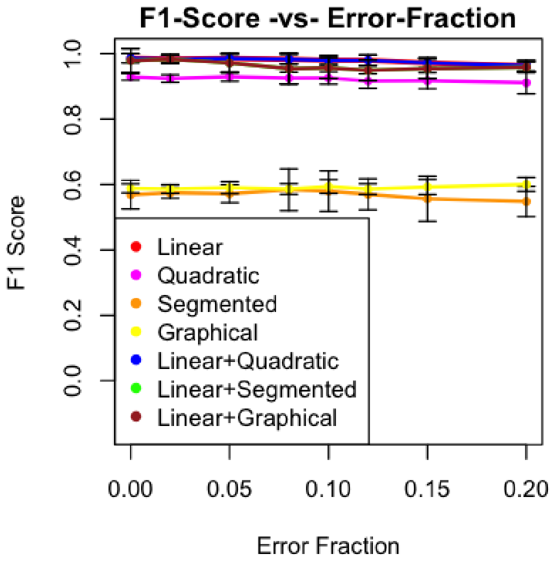

Figure 6.2: Inversion Detection experiment (N=35 trials) results with the Graphical-Model detector (and hybrid Linear+Graphical detector added)

6.5 Qualitative Assessment

We found that the fits provided by the graphical-model are quite similar to those provided by the Segmented model introduced in Chapter 5. Since no attempt was made to learn a continuous function, there is often a ‘gap’ (i.e. discontinuity) between the value of the function at the inflection-point, unlike in the case of the Segmented model.

Though the graphical-model does not have any specific advantages over other detec-

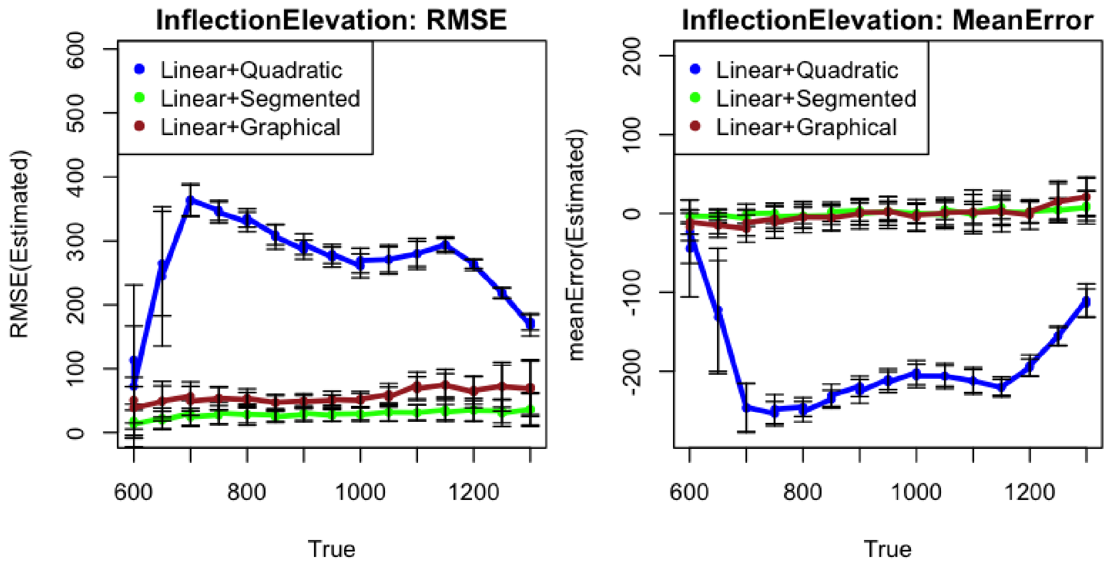

Figure 6.3: Inflection Elevation estimation experiment (N=35 trials) results with the Graphical-Model detector (and hybrid Linear+Graphical detector) added

tors currently, it, due to its extensibility, serves as a starting-point for future explorations.

Chapter 7: Conclusion

We achieve our goal of temporally and spatially localizing CAPs in our multivariate time-series dataset. Key to achieving this outcome was using a domain-driven approach where we explicitly took into account the Temperature vs. Elevation (”lapse-rate”) relationship. Visualizations were essential at each stage to decide on next-steps in modeling.

7.1 Key takeaways

- To *detect* the presence of cold-air-pools in a landscape at a specific timestep, a linear model is often sufficient. To *localize* cold-air-pools spatially in addition to temporally, a non-linear model is required.
- Between the proposed non-linear detectors, the Piecewise-Linear/Segmented model outperforms the Quadratic detector in detecting lapse-rate inflection-points resulting from cold-air-pools. However, the Piecewise-Linear/Segmented model should be used in conjunction with a Linear detector, because it easily produces spurious fits when given data that is not inherently non-linear.
- The left-panel in the non-linear dashboard (ref. figure 4.1, and full animation linked to in the *Supplemental Materials*) that shows spatial extent of CAP reveals that, for the most part, only one large CAP forms in HJA. This is consistent with the topological map of the HJA (figure 1.1) that shows one primary depression along the stem of the river extending to the bottom-left end of the map.

7.2 Future work

We envision the following directions of modeling and exploration would be useful in the future:

- **Temporal correlation:** Specifically modeling the temporal correlation of the sensor data (Autoregressive in the observations), or the estimate of the inflection-point (Autoregressive in the latent-variable) could lead to more realistic dynamics for the estimated inflection-point.

- **Spatial correlation:** Similarly, we could also model the tendency of nearby areas to behave more like one another, and possibly take into account additional covariates such as the aspect (i.e. the direction that the ground slope faces) of spatial locations.
- **Spatially Disjoint CAPs:** In the case of a landscape where multiple disjoint CAPs form over time, our current models will not correctly estimate different CAP inflection-elevations (if they are indeed different) for different locations. To achieve this, extensions to the current models need to be developed that first identify the disjoint spatial CAP locations, and then estimate CAP inflection-elevations separately for each.

Bibliography

[1] Gregory C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[2] Christopher Daly, David R. Conklin, and Michael H. Unsworth. Local atmospheric decoupling in complex topography alters climate change impacts. *International Journal of Climatology*, 30(12):1857–1864, 2010.

[3] Christopher Daly, Michael Halbleib, Joseph I. Smith, Wayne P. Gibson, Matthew K. Doggett, George H. Taylor, Jan Curtis, and Phillip P. Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International Journal of Climatology*, 28(15):2031–2064, 2008.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[5] Ethan W. Dereszynski and Thomas G. Dietterich. Probabilistic models for anomaly detection in remote sensor data streams. *CoRR*, abs/1206.5250, 2012.

[6] SOLOMON Z. DOBROWSKI. A climatic basis for microrefugia: the influence of terrain on climate. *Global Change Biology*, 17(2):1022–1035, 2011.

[7] Solomon Z. Dobrowski, John T. Abatzoglou, Jonathan A. Greenberg, and S.G. Schladow. How much influence does landscape-scale physiography have on air temperature in a mountain environment? *Agricultural and Forest Meteorology*, 149(10):1751 – 1758, 2009.

[8] Sarah J. K. Frey, Adam S. Hadley, Sherri L. Johnson, Mark Schulze, Julia A. Jones, and Matthew G. Betts. Spatial models reveal the microclimatic buffering capacity of old-growth forests. *Science Advances*, 2(4), 2016.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2 edition, 2009.

[10] Zachary A. Holden, John T. Abatzoglou, Charles H. Luce, and L. Scott Baggett. Empirical downscaling of daily minimum air temperature at very fine resolutions in complex terrain. *Agricultural and Forest Meteorology*, 151(8):1066 – 1073, 2011.

[11] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[12] Neil P. Lareau, Erik Crosman, C. David Whiteman, John D. Horel, Sebastian W. Hoch, William O. J. Brown, and Thomas W. Horst. The persistent cold-air pool study. *Bulletin of the American Meteorological Society*, 94(1):51–63, 2013.

[13] Jessica D. Lundquist, Nicholas Pepin, and Caitlin Rochford. Automated algorithm for mapping regions of cold-air pooling in complex terrain. *Journal of Geophysical Research: Atmospheres*, 113(D22):n/a–n/a, 2008. D22107.

[14] Justin R. Minder, Philip W. Mote, and Jessica D. Lundquist. Surface temperature lapse rates over complex terrain: Lessons from the cascade mountains. *Journal of Geophysical Research: Atmospheres*, 115(D14):n/a–n/a, 2010. D14122.

[15] Vito M.R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22:3055–3071, 2003.

[16] Vito M.R. Muggeo. segmented: an r package to fit regression models with broken-line relationships. *R News*, 8(1):20–25, 2008.

[17] Ocean OSU College of Earth and Atmospheric Sciences. Prof. Julia Jones, Faculty Profile. `http://ceoas.oregonstate.edu/profile/jones/`, 2017. [Online; accessed 06-July-2017].

[18] N. C. Pepin and J. D. Lundquist. Temperature trends at high elevations: Patterns across the globe. *Geophysical Research Letters*, 35(14):n/a–n/a, 2008. L14701.

[19] Joshua L Proctor and Philip A Eckhoff. Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International health*, 7(2):139–145, 2015.

[20] J. O. Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[21] Gilbert Strang. *Linear algebra and its applications*. Thomson, Brooks/Cole, 2006.

[22] Xiang Xuan. *Bayesian inference on change point problems*. PhD thesis, University of British Columbia, 2007.

APPENDICES

Appendix A: Supplemental Materials

The following supplemental material for this thesis can be found online at the URL `https://github.com/satpreetsingh/osu-cap`

Two days each in August (Summer) and December (Winter) have been chosen as exemplars for the behavior of the HJA:

- Normal (or minor night-time inversion): Aug 1st and Dec 12th
- Inversions: Aug 21st and Dec 9th (heavy inversion)

Artifacts provided:

- Linear Detector Dashboard Animations [LinearDashboard_2011-MM-DD.fullday.gif]
- Nonlinear Detector Dashboard Animations [NonlinearDashboard_2011-MM-DD.fullday.gif]
- Nonlinear Detectors Inflection-Elevation and R^2 time-series: [NonlinearAuxillary_2011-MM-DD.fullday.png]
- Graphical Model Animations: [Nonlinear_GraphicalModel_2011-MM-DD.fullday.png]

Appendix B: EM Derivations for Graphical Model

B.1 Complete-data log-likelihood:

$$
\begin{aligned}
\mathcal{L}_c(\theta) &= logP(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{G}, Z|\boldsymbol{\theta}) \\
&= log\Big(\prod_{i=1}^{N} P(Y_i, X_i, G_i|Z, \theta)\Big)P(Z|\theta) \\
&= log\prod_{k=1}^{K}\Big[\prod_{i=1}^{N}\Big(P(Y_i|X_i, G_i, Z = z_k, \theta)P(G_i|X_i, Z = z_k)P(X_i)\Big)P(Z = z_k|\theta)\Big]^{I[Z=z_k]} \\
&= log\prod_{k=1}^{K}\Big[\prod_{i=1}^{N}\Big(\prod_{j=1}^{J}\Big\{P(Y_i|X_i, G_i = g_j, Z = z_k, \theta)P(G_i = g_j|X_i, Z = z_k)\Big\}^{I[G_i=g_j]}P(X_i)\Big) \\
&\qquad P(Z = z_k|\theta)\Big]^{I[Z=z_k]} \\
&= \sum_{k=1}^{K}\Big[log\prod_{i=1}^{N}\Big(\prod_{j=1}^{J}\Big\{P(Y_i|X_i, G_i = g_j, Z = z_k, \theta)P(G_i = g_j|X_i, Z = z_k)\Big\}^{I[G_i=g_j]}P(X_i)\Big) \\
&\qquad P(Z = z_k|\theta)\Big]^{I[Z=z_k]} \\
&= \sum_{k=1}^{K}\Big[I[Z = z_k]log\prod_{i=1}^{N}\Big(\prod_{j=1}^{J}\Big\{P(Y_i|X_i, G_i = g_j, Z = z_k, \theta)P(G_i = g_j|X_i, Z = z_k)\Big\}^{I[G_i=g_j]}P(X_i)\Big) \\
&\qquad P(Z = z_k|\theta)\Big] \\
&= \sum_{k=1}^{K}\Big[I[Z = z_k]\sum_{i=1}^{N}\Big(\sum_{j\in\{0,1\}} log\Big\{P(Y_i|X_i, G_i = g_j, Z = z_k, \theta)P(G_i = g_j|X_i, Z = z_k)\Big\}^{I[G_i=g_j]} \\
&\qquad P(X_i)\Big) + logP(Z = z_k|\theta)\Big] \\
&= \sum_{k=1}^{K}\Big[I[Z = z_k]\sum_{i=1}^{N}\Big(\sum_{j\in\{0,1\}}\Big(I[G_i = g_j]log\Big\{P(Y_i|X_i, G_i = g_j, Z = z_k, \theta) \\
&\qquad P(G_i = g_j|X_i, Z = z_k)\Big\}\Big) + logP(X_i)\Big)I[Z = z_k]logP(Z = z_k|\theta)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{K}\Bigg[I[Z=z_k]\sum_{i=1}^{N}\Bigg(\sum_{j\in\{0,1\}}\Bigg(I[G_i=g_j]logP(Y_i|X_i,G_i=g_j,Z=z_k,\theta)\\
&\quad + I[G_i=g_j]logP(G_i=g_j|X_i,Z=z_k)\Bigg)+logP(X_i)\Bigg)\\
&\quad + I[Z=z_k]logP(Z=z_k|\theta)\Bigg]
\end{aligned}
$$

B.2 E-Step:

For $Q(\boldsymbol{\theta},\boldsymbol{\theta'})$, take $\mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y}}\mathcal{L}_c$. Let θ' be the previous iteration's value of the parameters.

$$
\begin{aligned}
&Q(\theta,\theta^{t-1})\\
&= \mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}}\sum_{k=1}^{K}\Bigg[I[Z=z_k]\sum_{i=1}^{N}\Bigg(\sum_{j\in\{0,1\}}\Bigg(I[G_i=g_j]logP(Y_i|X_i,G_i=g_j,Z=z_k,\theta)\\
&\quad + I[G_i=g_j]logP(G_i=g_j|X_i,Z=z_k)\Bigg)+logP(X_i)\Bigg)\\
&\quad + I[Z=z_k]logP(Z=z_k|\theta)\Bigg]\\
&= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}}\Bigg[I[Z=z_k]\sum_{i=1}^{N}\Bigg(\sum_{j\in\{0,1\}}\Bigg(I[G_i=g_j]logP(Y_i|X_i,G_i=g_j,Z=z_k,\theta)\\
&\quad + I[G_i=g_j]logP(G_i=g_j|X_i,Z=z_k)\Bigg)+logP(X_i)\Bigg)\\
&\quad + I[Z=z_k]logP(Z=z_k|\theta)\Bigg]\\
&= \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},,\boldsymbol{\theta'}}\Bigg[\sum_{i=1}^{N}\Bigg(\sum_{j\in\{0,1\}}\Bigg(I[Z=z_k,G_i=g_j]logP(Y_i|X_i,G_i=g_j,Z=z_k,\theta)\\
&\quad + I[Z=z_k,G_i=g_j]logP(G_i=g_j|X_i,Z=z_k)\Bigg)\\
&\quad + I[Z=z_k]logP(X_i)\Bigg)+I[Z=z_k]logP(Z=z_k|\theta)\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} \mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}}\Big[I[Z=z_k,G_i=g_j]logP(Y_i|X_i,G_i=g_j,Z=z_k,\theta)\Big] \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} \mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}}\Big[I[Z=z_k,G_i=g_j]logP(G_i=g_j|X_i,Z=z_k)\Big] \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}}\Big[I[Z=z_k]logP(X_i)\Big] \\
&\quad + \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{G},Z|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}}\Big[I[Z=z_k]logP(Z=z_k|\theta)\Big] \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} P(Z=z_k,G_i=g_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}')logP(Y_i|X_i,G_i=g_j,Z=z_k,\theta) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} P(Z=z_k,G_i=g_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}')logP(G_i=g_j|X_i,Z=z_k) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N} P(Z=z_k|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}')logP(X_i) \\
&\quad + \sum_{k=1}^{K} P(Z=z_k|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'})logP(Z=z_k|\theta)
\end{aligned}
$$

Next, to compute $P(Z=z_k,G_i=g_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'})$, we need:

$$
\begin{aligned}
P(Z=z_k,G_i=g_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta'}) &= \frac{P(Z=z_k,G_i=g_j,\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta'})}{P(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta'})} \\
&= \frac{P(Z=z_k,G_i=g_j,\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta'})}{\sum_Z\sum_{\boldsymbol{G}} P(Z,\boldsymbol{G},\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta'})}
\end{aligned}
$$

Thus, we need to compute the joint $P(Z = z_k, G_i = g_j, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'})$.

$$\begin{aligned}
&P(Z = z_k, G_i = g_j, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'}) \\
&= \sum_{G_1,\ldots,G_{i-1},Gi+1,\ldots,G_N} P(Z = z_k, G_1, \ldots, G_{i-1}, G_i = g_j, G_{i+1}, \ldots, G_N, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'}) \\
&= \sum_{\boldsymbol{G}_{-i}} P(Z = z_k, G_0, \ldots, G_{i-1}, G_i = g_j, G_{i+1}, \ldots, G_N, \boldsymbol{X}, \boldsymbol{Y}) \\
&= \sum_{\boldsymbol{G}_{-i}} \Bigg[P(Z = z_k) \Bigg(P(Y_i|X_i, G_i = g_j, Z = z_k, \theta')P(G_i = g_j|Z = z_k, X_i, \theta')P(X_i) \Bigg) * \\
&\qquad \Bigg(\prod_{l \neq i}^{N} P(Y_l|X_l, G_l, Z = z_k, \theta')P(G_l|Z = z_k, X_l, \theta')P(X_l) \Bigg) \Bigg] \\
&= P(Z = z_k) \Bigg(P(Y_i|X_i, G_i = g_j, Z = z_k, \theta')P(G_i = g_j|Z = z_k, X_i, \theta')P(X_i) \Bigg) * \\
&\qquad \sum_{\boldsymbol{G}_{-i}} \Bigg[\Bigg(\prod_{l \neq i}^{N} P(Y_l|X_l, G_l, Z = z_k, \theta')P(G_l|Z = z_k, X_l, \theta')P(X_l) \Bigg) \Bigg] \\
&= P(Z = z_k) \Bigg(P(Y_i|X_i, G_i = g_j, Z = z_k, \theta')P(G_i = g_j|Z = z_k, X_i, \theta')P(X_i) \Bigg) * \\
&\qquad \Bigg[\prod_{l \neq i}^{N} P(X_l) \Bigg(\sum_{G_l \in \{0,1\}} P(Y_l|X_l, G_l, Z = z_k, \theta')P(G_l|Z = z_k, X_l, \theta') \Bigg) \Bigg]
\end{aligned}$$

Substitutions:

Let

$$\gamma_{i,j,k} = P(Z = z_k, G_i = g_j|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta'}) = \frac{P(Z = z_k, G_i = g_j, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'})}{\sum_Z \sum_{\boldsymbol{G}} P(Z, \boldsymbol{G}, \boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta'})}$$

Let

$$\delta_k = P(Z = z_k|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta'}) = \sum_{i,j} \gamma_{i,j,k}$$

B.3 M-Step:

Given $Q(\theta, \theta^{t-1})$ from E-Step, get

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\ Q(\boldsymbol{\theta}, \boldsymbol{\theta^t})$$

B.3.1 $G_{i,j}$ and $P(X_i)$:

- $G_{i,j}$ is a deterministic quantity given $Z = z_k$ and observed quantity X_i, and is simply set as: $G_{i,0} = I[X_i < z_k]$ and $G_{i,1} = I[X_i \geq z_k]$
- $P(X_i)$ is a Uniform distribution with mean $1/N$ and does not need any parameters learned. (Here, N is the number of observations or sensors)

For other parameters, we take take partial derivatives and set to zero to find the maximum:

B.3.2 For π_k:

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta'})}{\partial \pi_k} = \frac{\partial}{\partial \pi_k} \Bigg[& \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \{0,1\}} \gamma_{i,j,k} log P(Y_i | X_i, G_i = g_j, Z = z_k, \theta) \\
& + \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \{0,1\}} \gamma_{i,j,k} log P(G_i = g_j | X_i, Z = z_k) \\
& + \sum_{k=1}^{K} \sum_{i=1}^{N} \delta_k log P(X_i) + \sum_{k=1}^{K} \delta_k log P(Z = z_k | \theta) \\
& + \lambda \Bigg(\sum_{k=1}^{K} \pi_k - 1 \Bigg) \Bigg] \\
= \frac{\partial}{\partial \pi_k} \Bigg[& \sum_{k=1}^{K} \delta_k log(\pi_k) + \lambda \Bigg(\sum_{k=1}^{K} \pi_k - 1 \Bigg) \Bigg] \\
= \frac{\delta_k}{\pi_k} & + \lambda
\end{aligned}$$

By setting to 0, we get:

$$\begin{aligned}
\frac{\delta_k}{\pi_k} + \lambda &= 0 \\
\delta_k + \lambda \pi_k &= 0 \\
\sum_{k=1}^{K} \delta_k + \lambda \sum_{k=1}^{K} \pi_k &= 0 \\
\lambda &= -\sum_{k=1}^{K} \delta_k \\
\therefore \pi_k &= \frac{\delta_k}{\sum_{k=1}^{K} \delta_k}
\end{aligned}$$

B.3.3 For β_0:

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta,\theta'})}{\partial \beta_{0_{j,k}}} &= \frac{\partial}{\partial \beta_{0_{j,k}}}\Bigg[\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}}\gamma_{i,j,k} logP(Y_i|X_i, G_i = g_j, Z = z_k, \theta) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}}\gamma_{i,j,k} logP(G_i = g_j|X_i, Z = z_k) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N}\delta_k logP(X_i) \\
&\quad + \sum_{k=1}^{K}\delta_k logP(Z = z_k|\theta)\Bigg] \\
&= \frac{\partial}{\partial \beta_{0_{j,k}}}\Bigg[\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}}\gamma_{i,j,k} logP(Y_i|X_i, G_i = g_j, Z = z_k, \theta)\Bigg] \\
&= \sum_{i=1}^{N}\gamma_{i,j,k}\frac{\partial}{\partial \beta_{0_{j,k}}}\Bigg[logP(Y_i|X_i, G_i = g_j, Z = z_k, \theta)\Bigg] \\
&= \sum_{i=1}^{N_j}\gamma_{i,j,k}\frac{\partial}{\partial \beta_{0_{j,k}}}\Bigg[log\mathcal{N}(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i, \sigma_k^2)\Bigg] \\
&= \sum_{i=1}^{N_j}\gamma_{i,j,k}\frac{\partial}{\partial \beta_{0_{j,k}}}\Bigg[-0.5log(2\pi\sigma_k^2) - \frac{(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i)^2}{2\sigma_k^2}\Bigg] \\
&= \sum_{i=1}^{N_j}\gamma_{i,j,k}\Bigg[\frac{2(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i)}{2\sigma_k^2}\Bigg]
\end{aligned}
$$

...where N_j is subset of N s.t. $G_i = g_j$

Setting to 0, we get

$$\beta_{0_{j,k}} = \frac{\sum_{i=1}^{N_j}\gamma_{i,j,k}(Y_i - \beta_{1_{j,k}}X_i)}{\sum_{i=1}^{N_j}\gamma_{i,j,k}}$$

B.3.4 For β_1:

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta'})}{\partial \beta_{1_{j,k}}} &= \frac{\partial}{\partial \beta_{1_{j,k}}} \Bigg[\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \{0,1\}} \gamma_{i,j,k} log P(Y_i | X_i, G_i = g_j, Z = z_k, \theta) \\
&\quad + \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \{0,1\}} \gamma_{i,j,k} log P(G_i = g_j | X_i, Z = z_k) \\
&\quad + \sum_{k=1}^{K} \sum_{i=1}^{N} \delta_k log P(X_i) \\
&\quad + \sum_{k=1}^{K} \delta_k log P(Z = z_k | \theta) \Bigg] \\
&= \frac{\partial}{\partial \beta_{1_{j,k}}} \Bigg[\sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in \{0,1\}} \gamma_{i,j,k} log P(Y_i | X_i, G_i = g_j, Z = z_k, \theta) \Bigg] \\
&= \sum_{i=1}^{N} \gamma_{i,j,k} \frac{\partial}{\partial \beta_{1_{j,k}}} \Big[log P(Y_i | X_i, G_i = g_j, Z = z_k, \theta) \Big] \\
&= \sum_{i=1}^{N_j} \gamma_{i,j,k} \frac{\partial}{\partial \beta_{1_{j,k}}} \Big[log \mathcal{N}(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}} X_i, \sigma_k^2) \Big] \\
&= \sum_{i=1}^{N_j} \gamma_{i,j,k} \frac{\partial}{\partial \beta_{1_{j,k}}} \Bigg[-0.5 log(2\pi\sigma_k^2) - \frac{(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}} X_i)^2}{2\sigma_k^2} \Bigg] \\
&= \sum_{i=1}^{N_j} \gamma_{i,j,k} \Bigg[\frac{X_i (Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}} X_i)}{\sigma_k^2} \Bigg]
\end{aligned}
$$

...where N_j is subset of N s.t. $G_i = g_j$

Setting to 0, we get

$$
\beta_{1_{j,k}} = \frac{\sum_{i=1}^{N_j} \gamma_{i,j,k} X_i (Y_i - \beta_{0_{j,k}})}{\sum_{i=1}^{N_j} \gamma_{i,j,k} (X_i^2)}
$$

B.3.5 For σ_k:

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta},\boldsymbol{\theta'})}{\partial \sigma_k} &= \frac{\partial}{\partial \sigma_k}\Bigg[\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} \gamma_{i,j,k} logP(Y_i|X_i, G_i = g_j, Z = z_k, \theta) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} \gamma_{i,j,k} logP(G_i = g_j|X_i, Z = z_k) \\
&\quad + \sum_{k=1}^{K}\sum_{i=1}^{N} \delta_k logP(X_i) \\
&\quad + \sum_{k=1}^{K} \delta_k logP(Z = z_k|\theta)\Bigg] \\
&= \frac{\partial}{\partial \sigma_k}\Bigg[\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j\in\{0,1\}} \gamma_{i,j,k} logP(Y_i|X_i, G_i = g_j, Z = z_k, \theta)\Bigg] \\
&= \sum_{j\in\{0,1\}}\sum_{i=1}^{N} \gamma_{i,j,k}\frac{\partial}{\partial \sigma_k}\Bigg[logP(Y_i|X_i, G_i = g_j, Z = z_k, \theta)\Bigg] \\
&= \sum_{j\in\{0,1\}}\sum_{i=1}^{N_j} \gamma_{i,j,k}\frac{\partial}{\partial \sigma_k}\Bigg[log\mathcal{N}(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i, \sigma_k^2)\Bigg] \\
&= \sum_{j\in\{0,1\}}\sum_{i=1}^{N_j} \gamma_{i,j,k}\frac{\partial}{\partial \sigma_k}\Bigg[-0.5log(2\pi\sigma_k^2) - \frac{(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i)^2}{2\sigma_k^2}\Bigg] \\
&= \sum_{j\in\{0,1\}}\sum_{i=1}^{N_j} \gamma_{i,j,k}\Bigg[\frac{-1}{\sigma_k} + \frac{(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i)^2}{\sigma_k^3}\Bigg]
\end{aligned}$$

...where N_j is subset of N s.t. $G_i = g_j$

Setting to 0, we get

$$\sigma_k^2 = \frac{\sum_{j\in\{0,1\}}\sum_{i=1}^{N_j}\gamma_{i,j,k}(Y_i - \beta_{0_{j,k}} - \beta_{1_{j,k}}X_i)^2}{\sum_{j\in\{0,1\}}\sum_{i=1}^{N_j}\gamma_{i,j,k}}$$