Adaptive Conformal Inference under Distribution Shift using Adaptive Step Size

Theodora Ko

March 2023

1 Introduction

For our final project, we chose the theory option, focusing on understanding and expanding upon “Adaptive conformal inference under distribution shift” by Gibbs and Candes. First, we will provide a summary of the paper and how it modifies conformal prediction theory to accommodate for shifts in covariate distribution, all in a distribution-free way. Then, we take a closer look at the stepwise search the authors use in constructing adaptive conformal inference. The constant step size introduces a trade-off between adaptability and stability, so we experiment with an adaptive step size and compare coverage results with that of the constant step size.

We also explore the method’s applications to data in addition to its theory, with a particular interest in features typically seen in real-world scenarios but may challenge the ACI method. Specifically, we are interested in data sets that behave cyclically or exhibit sudden shocks. We simulate data sets with these features and run both the adaptive and constant step ACI methods.

2 Paper Summary

The paper develops a method named ”adaptive conformal inference” that constructs prediction sets that are robust to distribution shifts.

2.1 Conformal Prediction

The paper builds on the process of conformal prediction, so we provide a brief introduction to conformal prediction in the regression setting. In particular, we review split conformal prediction, which is less computationally intensive than full conformal prediction. Conformal prediction is a method for constructing prediction sets that does not depend on any specific distributions, sample size, or regression method.

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$, which are exchangeable from some distribution P, and we build a regression model for predicting Y based on X. We want to create a prediction set for Y_{n+1} using X_{n+1}, $\hat{C}_{n+1}$, such that $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \geq 1 - \alpha$.

First, we select a nonconformity score function, which indicates how well a data point conforms to the distribution. In the typical regression setting, a common score function is the absolute value of the residual:

$$S(X, Y) = |\hat{\mu}(X) - Y|$$

where $\hat{\mu}$ is some regression model. Other possible score functions include

$$S(X, Y) = \frac{|\hat{\mu}(X) - Y|}{\hat{\sigma}(X)}$$

which can be used for nonconstant and unknown variance, and

$$S(X, Y) = \max\{\hat{q}_{\alpha/2}(X) - Y, Y - \hat{q}_{1-\alpha/2}(X)\}$$

which is used for quantile regression.

Then we fit the score function on a portion of the data, $(X_1, Y_1), ..., (X_{n_0}, Y_{n_0})$ (e.g., taking $n_0 = n/2$). For the score function $S(X, Y) = |\hat{\mu}(X) - Y|$, for example, $\hat{\mu}$ would be some regression model fit on some fraction of the data. Then we compute the nonconformity scores on the holdout sample $S_i = S(X_i, Y_i)$ for $i = n_0 + 1, ..., n$ and define the quantile function of these scores as

$$\hat{Q}(p) = \inf\left\{ s : \left(\frac{1}{n - n_0} \sum_{(X_i, Y_i): i \in [n_0+1, n]} \mathbb{1}_{S_i \leq s} \right) \geq p \right\}$$

Less formally, $\hat{Q}(1 - \alpha)$ is equal to the $\lceil (1 - \alpha)(n_0 + 1) \rceil$-th smallest value of the set $\{S_{n_0+1}, ..., S_n\}$ where $\lceil x \rceil$ indicates the ceiling function which outputs the least integer greater than or equal to x. Then the prediction interval for Y_{n+1} at the α level is equal to

$$\hat{C}(X_{n+1}) = \{y \in \mathcal{Y} : S(X_{n+1}, y) \leq \hat{Q}(1 - \alpha)\}$$

In the example of $S(X, Y) = |\hat{\mu}(X) - Y|$, the prediction interval is equal to

$$\hat{C}(X_{n+1}) = \{y \in \mathbb{R} : |y - \hat{\mu}(X_{n+1})| \leq \hat{Q}(1 - \alpha)\} = \hat{\mu}(X_{n+1}) \pm \hat{Q}(1 - \alpha)$$

This method gives a marginal coverage guarantee $\mathbb{P}[Y_{n+1} \in \hat{C}(X_{n+1})] \geq 1 - \alpha$.

2.2 Conformal Prediction Under Covariate Shift

There may be cases where the distributions of the training and test data are not the same so the training and test data points are not exchangeable with each other. One paper that addresses the problem of covariate shift is "Conformal Prediction Under Covariate Shift" (Tibshirani 2019). This paper assumes that the covariate distribution is P_X for the training set, and $P_{\tilde{X}}$ for the test set, but the marginal distribution of Y given X remains the same between the training and test set. If we know the likelihood ratio between the two distributions P_X and $P_{\tilde{X}}$ (or if we are able to estimate accurately the ratio between the two distributions), we can still perform conformal prediction, "using a quantile of a suitably weighted empirical distribution of nonconformity scores" (Gibbs 2021).

2.3 Adaptive Conformal Inference Under Distribution Shift

Now consider the setting where we do not know how the distribution of the covariates shifts over time, and the marginal distribution of Y given X also may not be fixed. In this scenario, we would want to use adaptive conformal inference (ACI).

In ACI, we fit a score function and quantile function for each time t, S_t, and $\hat{Q}_t$. Using the score and quantile functions, we construct a prediction set

$$\hat{C}_t(\alpha) := \{y : S_t(X_t, y) \leq \hat{Q}_t(1-\alpha)\}$$

and define the miscoverage rate as

$$M_t(\alpha) := \mathbb{P}(S_t(X_t, Y_t) > \hat{Q}_t(1-\alpha))$$

which is the probability that the actual value of Y_t falls outside of the prediction interval $\hat{C}_t$. Since the distribution shift is unknown and the data points are not exchangeable, we don't expect the miscoverage rate $M_t(\alpha)$ to be equal α. However, we can calibrate α_t^* according to the local error rate such that our miscoverage rate follows $M_t(\alpha_t^*) \approx \alpha$.

Empirically, if the miscoverage frequency of the previous prediction sets were historically undercovering, then we want to increase our estimate of α_t^* such that our prediction set returns to approximately α coverage. Similarly, if our prediction sets were historically over-covering, we increase our estimate of α_t^*.

More precisely, if the prediction set at observation t did not contain Y_t, then we want to increase our prediction interval by incrementing α_t by a step γ.

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \mathrm{err}_t)$$

where

$$\mathrm{err}_t := \begin{cases} 1, & \text{if } Y_t \notin \hat{C}_t(\alpha_t) \\ 0, & \text{otherwise} \end{cases}$$

and err_t represents historical miscoverage rate. We initialize $\alpha_1 = \alpha$ so that our miscoverage rate tends toward α.

To explain the intuition behind the adaptive confidence interval, we can understand its logic the following way: given a significant distribution shift followed by a series of errors, it will return a constant negative value for $(\alpha - \mathrm{err}_t)$, decreasing the value of α_{t+1} gradually. Decreasing the value of alpha leads to widening the confidence interval until it captures the true value of Y_{t+1} and from there on, the value of $(\alpha - \mathrm{err}_t)$ will become positive and the value of α_{t+1} will increase again. Eventually, then the error rate will converge to α. The paper proves the above in section 4. By lemma 4.1, this is true with probability 1 for all $t \in \mathbb{N}$, $\alpha_t \in [-\gamma, 1+\gamma]$. Then, with the expansion of the recursive definition of α_{t+1}, $\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathrm{err}_t)$ using lemma 4.1, we get that with probability one for any $T \in \mathbb{N}$,

$$\lim_{T\to\infty} \sum_{t=1}^{T} err_t = \alpha.$$

Another method of updating α_t is by looking at all historical err_t and using weighting them to make a decision about adaptive α_t. One way of doing this is by weighting previous err_t such that we place more significance on recent err_t while also using err_1 to draw conclusions about how to adjust $alpha_t$. Therefore, an alternative update to α_{t+1} would be:

$$\alpha_{t+1} = \alpha_t + \gamma \left(\alpha - \sum_{s=1}^{t} \omega_s \mathrm{err}_s \right)$$

where ω_s must satisfy:

$$\sum_{s=1}^{t} \omega_s = 1$$

Finally, it should be noted that the ACI can't make any strong guarantees about the marginal coverage rate for each individual time t. However, the method does ensure that as the total time T tends towards infinity, the average miscoverage rate for the prediction intervals at each time converges almost surely to α. In other words, AIC only guarantees the asymptotic convergence of the average miscoverage level to α.

Additionally, the paper provides some additional theoretical results if the distribution shift is sufficiently small.

2.4 Real Data Examples

The paper includes two scenarios to demonstrate ACI's predictive powers. In each case, the authors find that the local coverage provided by the adaptive alpha is more complete than that of the fixed alpha. In the next section, we will draw upon these data and data methods to experiment with an adaptive step size mechanism.

The first is predicting market volatility using daily stock prices for four companies, where time intervals are measured by the next day the price for the company is reported. This analysis introduces the GARCH model for point predictions, as commonly used in financial analysis, that we will also employ in the next section. It also introduces a method of comparison that includes not only the fixed and adaptive alphas but also a Bernoulli sequence. The Bernoulli sequences intake a probability that is the targeted miscoverage rate of alpha, modeling the theoretical worst-case behavior for local coverage. We will also use this Bernoulli sequence as a benchmark when evaluating how an adaptive step size compares to a fixed step size, as well as when later evaluating how the original ACI performs on various data behaviors.

The second is predicting election outcomes for a set of counties based on a set of counties where the outcome is known. This scenario applies to political data where election results for all counties are reported in a sequence rather than simultaneously. As such, this data measures time by the next instance in which a county's voting outcome is reported, and the range of prediction is over a series of individual counties. This scenario differs from the market volatility one because the authors regress on the covariates when computing point predictions. When we later explore data with extreme shocks, we will also use covariate data in order to demonstrate covariate shift more explicitly, as well as employ a linear regression.

3 Step Size

The authors introduce the question of how to effectively choose the stepsize γ. Too high of a γ would make the trajectory of α_t unstable, but too low of a γ would mean that the method is not as adaptive to large distribution shifts. The authors formalize this intuition by finding that γ should be proportional to $\sqrt{|\alpha^*_{t+1} - \alpha^*_t|}$, meaning for larger distribution shifts, γ should be larger so that the method is more adaptable. For real-world data examples, the authors choose to use $\gamma = 0.005$ after testing multiple values of γ and choosing one that gave the right balance between stability and adaptability.

3.1 Randomized Step Size

We were interested in whether it was possible to ACI by using some type of non-constant γ. The authors write that it is possible to derive the online update

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \text{err}_t)$$

as "an online gradient descent algorithm with respect to the pinball loss" (Gibbs 2021). There is some research that suggests that a randomized step size can improve stochastic gradient descent (Musso 2020, Blier 2019), so we wanted to try various randomized stepsizes to see if they improved ACI.

In this paper, we will evaluate the effectiveness of the prediction intervals based on two metrics: (1) the local coverage rate, and (2) the size of the intervals. The addition of the second is necessary because the intervals must be useful as well as accurate. For example, an interval with a sufficiently large coverage rate is not necessarily informative if it is also inclusive of an extremely large range of the data.

Initially, we wanted to try the following stepsizes:

1. $\gamma_{t+1} = \gamma_t(1+a)$, $a \sim \text{Uniform}[-1,1]$
2. $\gamma_{t+1} = \gamma_t(1+a_t)$, $a_t \sim \text{Uniform}[-1,1]$
3. $\gamma_{t+1} = \gamma_t + a$, $a \sim \text{Uniform}[-\gamma_0, \gamma_0]$
4. $\gamma_{t+1} = \gamma_t + a_t$, $a_t \sim \text{Uniform}[-\gamma_0, \gamma_0]$

where values of $\gamma_t > 1$ would be set equal to 1 and values of $\gamma_t < 0$ would be set equal to 0. However, the problem with the first and third options was that if the $a \sim \text{Uniform}[-1,1]$ or $a \sim \text{Uniform}[-\gamma_0, \gamma_0]$ was negative, γ_t would reach 0, and if the a was positive, γ_t would eventually reach 1. Although the a_t was randomly generated for each time t for the second and fourth options, if the a_t was ever negative, γ_t would begin to decrease quickly towards 0. As a result, none of these randomized stepsize schemes were usable.

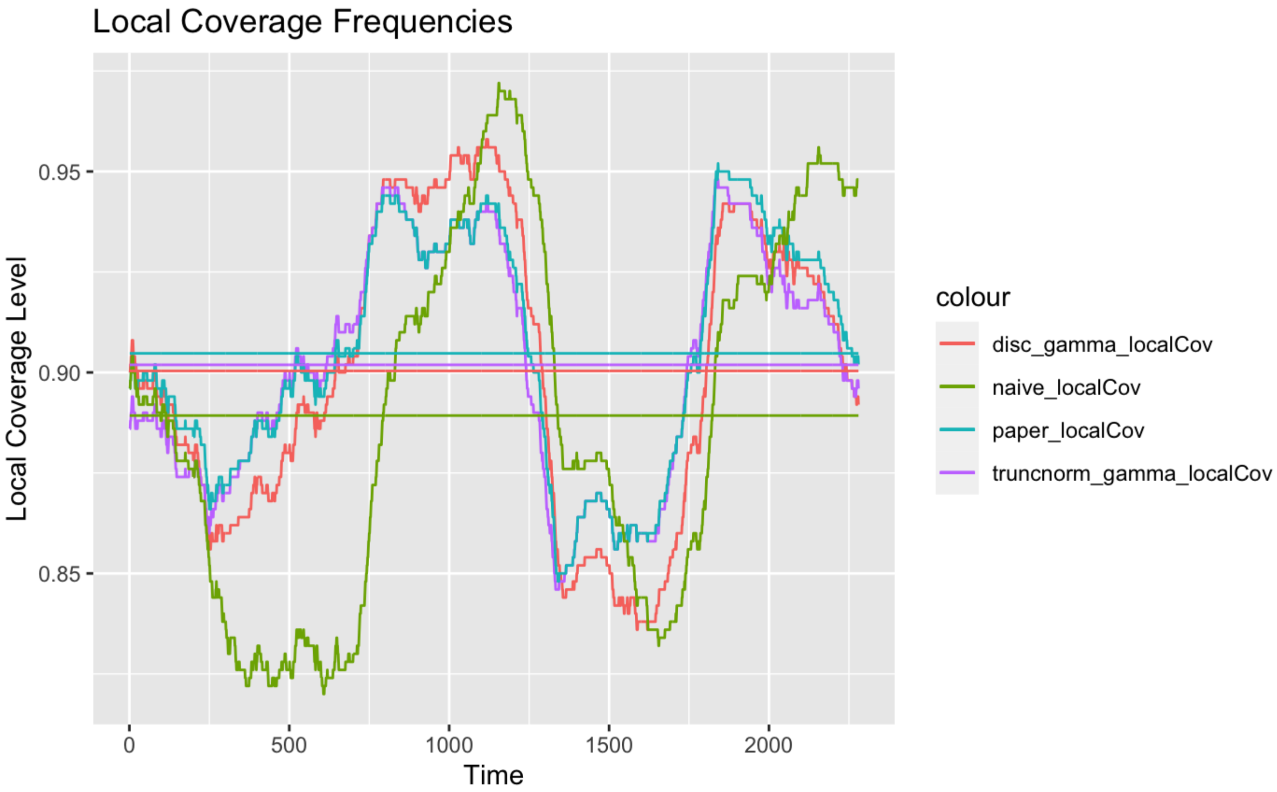
Next, we were interested in a stepsize that did not involve some kind of recursive formula. The two options we tried were:

1. $\gamma_t \sim \text{Uniform}\{e^{-10}, e^{-9}, ..., e^{-4}\}$ i.e. a discrete uniform distribution where each value has 1/6 probability of being chosen
2. $\gamma_t \sim TN(0, 1, 0.005, 0.005)$, a normal distribution with mean 0.005 and standard deviation 0.005 truncated at 0 and 1

We chose these two options because we wanted γ_t to be able to vary a wide range of possible step sizes, and we also didn't want γ_t to tend towards either 0 or 1 over time.

We tested each of these two randomized stepsizes with $\alpha = 0.1$ on the stock price data with a GARCH(1,1) model from section 2.2 in the paper and compared the results to the method in the paper using $\gamma = 0.005$. We focused only on the Blackberry open price from 2005 to 2020. The trajectory of α_t looked roughly similar between all three stepsizes.

	CI	ACI with $\gamma = 0.005$	ACI with option 1	ACI with option 2
CI size	0.1055986	0.1080479	0.10642	0.110085
average error	0.1047139	0.09751709	0.102195	0.1007557



ACI with the stepsize as option 1 gave the smallest confidence intervals on average, and ACI with $\gamma = 0.005$ gave the lowest average error. However, it did not look like any randomized stepsize made much of a difference from ACI with $\gamma = 0.005$, and all three methods performed similarly. The impact of a randomized stepsize could also potentially vary depending on the type of data being analyzed and the model being used.

3.2 Adaptive Step Size

In addition to the discussion of implementing gammas of different sizes, we were also interested in whether we can change the size of gamma along with the size of alpha, i.e. making the computation of α_{t+1} more responsive to the distribution shift and decreasing the rate of convergence of average miscoverage level to α.

More precisely, instead of using fixed γ value in the online update process, we wanted to decrease the volatility of α while guaranteeing a narrower confidence interval. To achieve so, we created an algorithm that does the following:

1. create error sequence that stores err_t for each time t and alpha sequence that stores α_t for each time t

2. check if the average miscoverage rate, or the mean of error sequence, is smaller than $\frac{\alpha}{2}$, our target error rate

3. - If the average miscoverage rate is lower $\frac{\alpha}{2}$, check if there is a constant decreasing trend of α_t. If there is a trend, stop decreasing the size of gamma. If not, decrease the size of γ by 0.999.
   - If the average miscoverage rate is higher than $\frac{\alpha}{2}$, increase the size of γ by 1.001.

4. store err_t to the error sequence and α_t to the alpha sequence

The intuition behind the logic is as follows:

- If the miscoverage rate on average is lower than $\frac{\alpha}{2} < \alpha$, we are already achieving the target coverage with the current α_t and also the distribution shift has either not been observed or is not influential enough that we need a significantly different $\alpha*_{t+1}$ from current α_t to adjust the miscoverage rate in the near future.

  The paper's suggested online update of α_{t+1} leads to constant increase of α_{t+1} value by $\gamma * \alpha$ in the case of no error at time t. If the average miscoverage rate is lower than $\frac{\alpha}{2}$ already and current α_t produces no error, we can decrease the size of increase from α_t to α_{t+1} by decreasing the size of γ. This would return us a narrower confidence interval for Y_{t+1} than the confidence interval for Y_{t+1} produced by a bigger γ value.

- Consider when the average miscoverage rate is bigger than $\frac{\alpha}{2}$. (Again note that we are being extremely conservative). Then, in the case of α_t that produces error, we want the increase of size of α_{t+1} to be bigger so that the average miscoverage rate can reach the convergence to α faster. Then, we increase the value of γ.

Algorithm 1 adaptive_gamma(data, $\alpha = 0.05, \gamma = 0.005$, lookback $= 1250$).

```
Input A time series data
Output sequence of adaptive error

initialise adaptive_error_sequence
adaptive_gamma ← $\gamma$
adaptive_gamma_sequence ← $\gamma$
adaptive_alphaSequence ← rep($\alpha$, T − lookback + 1)
T ← length(data)
adaptive_error_sequence ← rep(0, T − lookback + 1)

for $t \in$ lookback : $T$ do
    Define recentScores, scores as defined in the ACI paper
    Compute the error given adaptive_alphat and store it in adaptive_error_sequence
    Store current adaptive_alphat in adaptive_alphaSequence

    if $t -$ lookback $> 100$ then
        if mean(adaptive_error_sequence) $< \alpha/2$ then
            if is_decreasing(adaptive_alphaSequence) then
                pass
            else
                adaptive_gamma ← adaptive_gamma $\cdot$ 0.999
            end if
        else
            adaptive_gamma ← adaptive_gamma $\cdot$ 1.001
        end if
    end if
    Define weight w for each $\alpha_i$ for $1 \leq i \leq t$ as defined in the ACI paper
    Update adaptive_alphat using adaptive_gamma and w as defined in the paper
end for
return adaptive_gamma_sequence, adaptive_error_sequence
```

Note that

- We altered the code provided by the ACI paper and added our algorithm in simulating the adaptive gamma method.
- We chose a threshold of $\frac{\alpha}{2}$ for the average miscoverage rate so we can be extremely conservative when decreasing the step size. Note that simultaion of threshold α returns
- mean(adaptive_alphaSequence) is to prevent the size of γ constantly decreasing and reaching zero.
  - Consider a case where significant distribution shift has taken place and the value of α starts decreasing. The longer the alpha sequence decrease takes place, the smaller the size of gamma will be. Since adaptive_gamma $\cdot$ 0.999 is a geometric sequence, then adaptive gamma will converge to zero eventually.

  - Now, consider a case where the inference ahs been adjusted to the previous significant distribution shift. Then, since we want to achieve narrower confidence interval, we increase the size of gamma.

- The if statement - $t - \texttt{lookback} > 100$ in the above algorithm is to ensure that the training on the data has been done at least 100 times before we start adjusting the step size based on $\texttt{mean}(\texttt{adaptive\_alphaSequence})$

We tested for fixed γ from the ACI paper and adaptive γ based on the above algorithm on the Blackberry historical stock price data. We altered the code that was provided from the ACI paper in their GitHub and set initial $\gamma = 0.005$ and $\alpha = 0.05$.

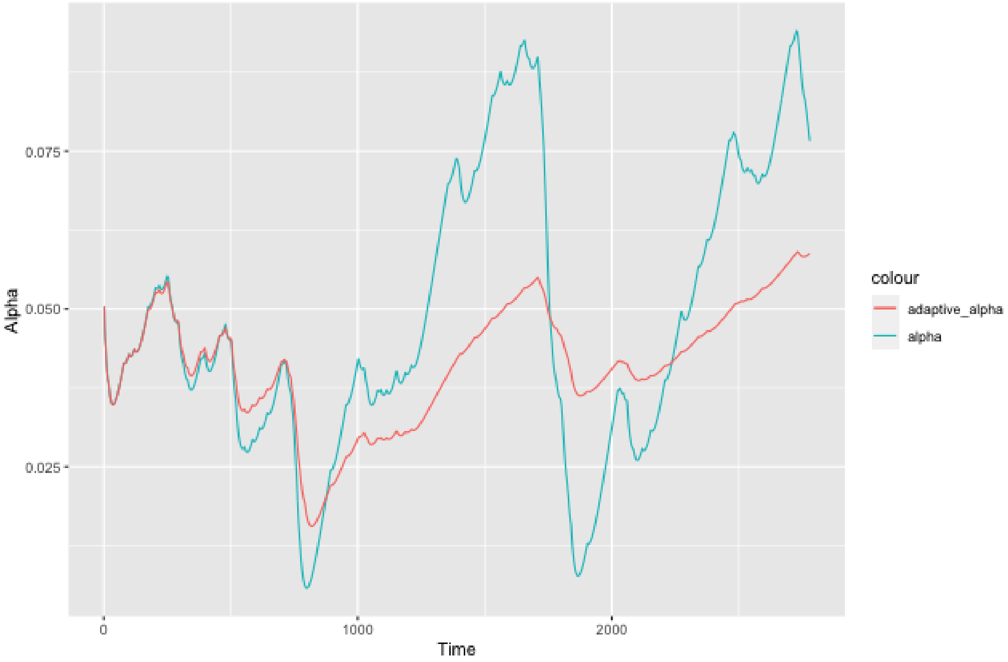


Fig 3.2.1. α values from adaptive gamma and fixed gamma method

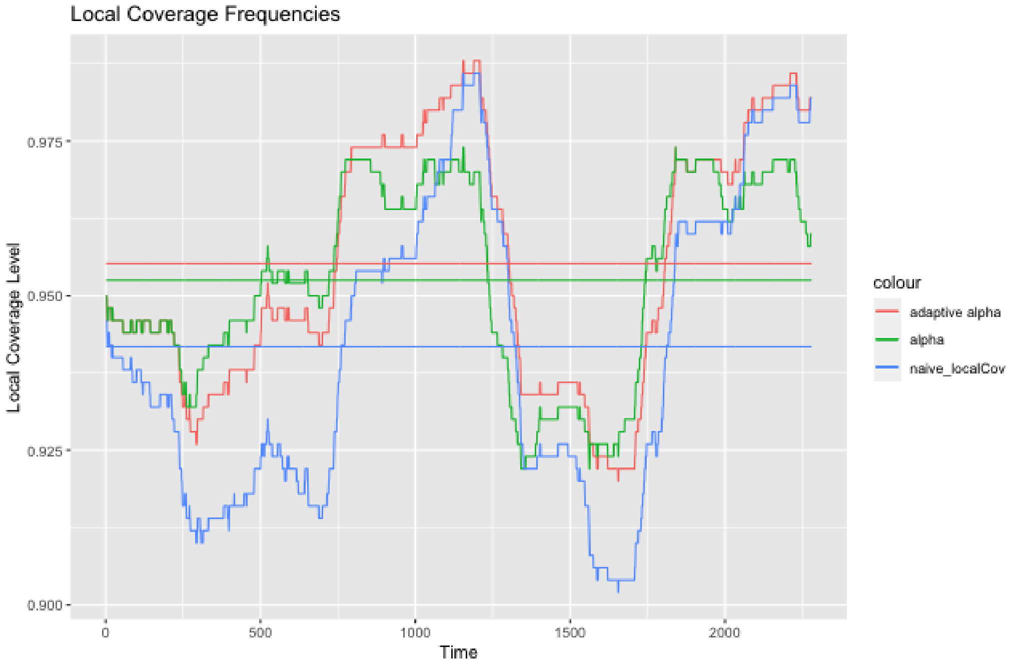


Fig 3.2.2. Local coverage frequencies from adaptive gamma and fixed gamma method

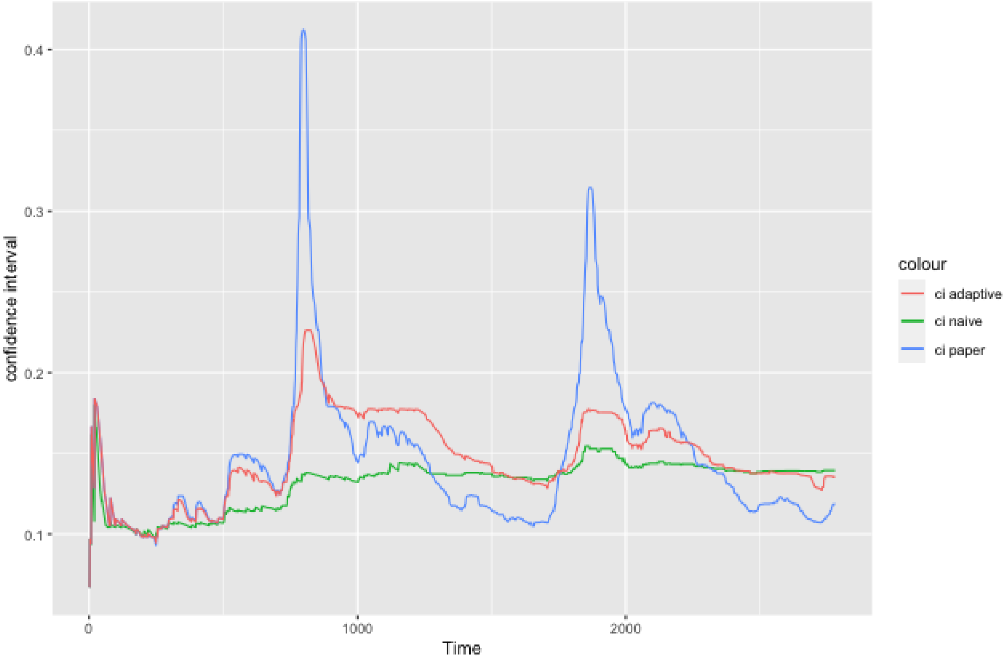


Fig 3.2.3. Confidence Intervals from adaptive gamma and fixed gamma method

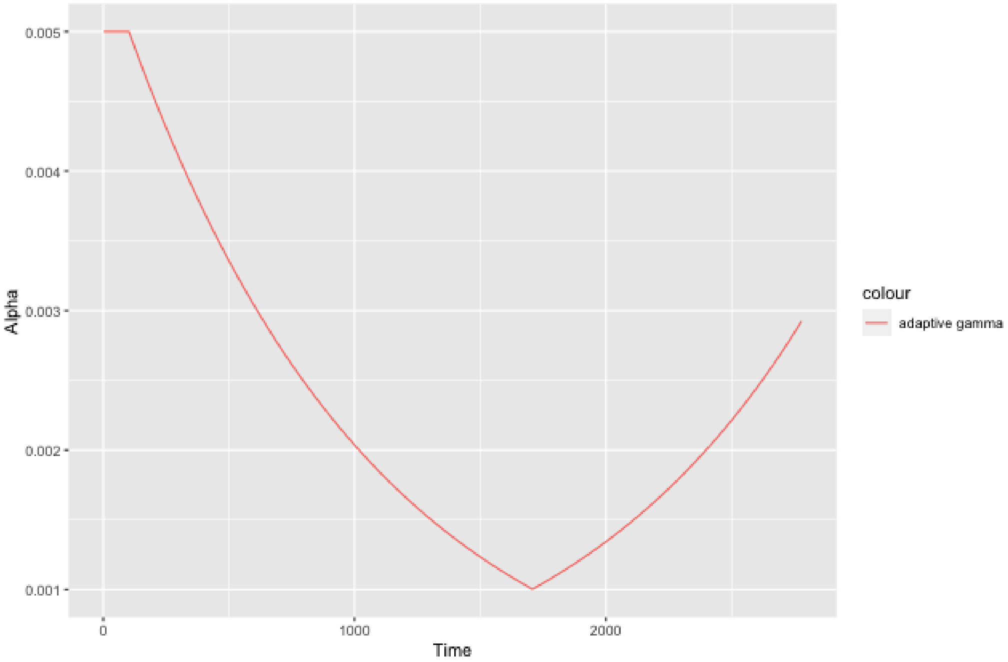


Fig 3.2.4. Gamma size change from gamma step size 0.001

From the simulation, we can observe that

1. Adaptive gamma works better than fixed gamma in terms of narrowing the confidence interval as shown by figure 3.2.3.

2. Adaptive gamma achieves better coverage level as shown by figure 3.2.2.

3. A smaller but nonzero γ works better than naive method of constant alpha (i.e. $\gamma = 0$) in terms of coverage level. This is because while there is no adaptation of confidence interval for naive method, smaller gamma still does lead to some adaptation of confidence interval to distribution shift.

4. The adaptive gamma method decreases the volatility of $alpha_t$ compared to that of fixed gamma, as shown by figure 3.2.1

5. In the extreme cases where confidence interval significantly widens (reference figure 3.2.3), the adaptive gamma achieved similar, or higher, coverage level with a narrower confidence interval.

We also tested if updating γ by different step size (0.001 and 0.002) would make difference. In the following figures, note that `adaptive_alpha1` is the α_t computed using gamma step size 0.001 and `adaptive_alpha2` is the α_t computed using gamma step size 0.002. From the simulation, we observed the followings:

- Bigger gamma step size leads less volatile alpha trend as shown by figure 3.2.4.
- Bigger gamma step size leads to higher coverage level and narrower confidence interval for extreme distribution shift as shown in figure 3.2.6. and 3.2.5.
- Bigger gamma step size leads to not only faster rate of change for gamma size but faster reaction to extreme events shown by the earlier trough of trend of gamma for step size 0.002 than that of step size 0.001 as shown in figure 3.2.7.

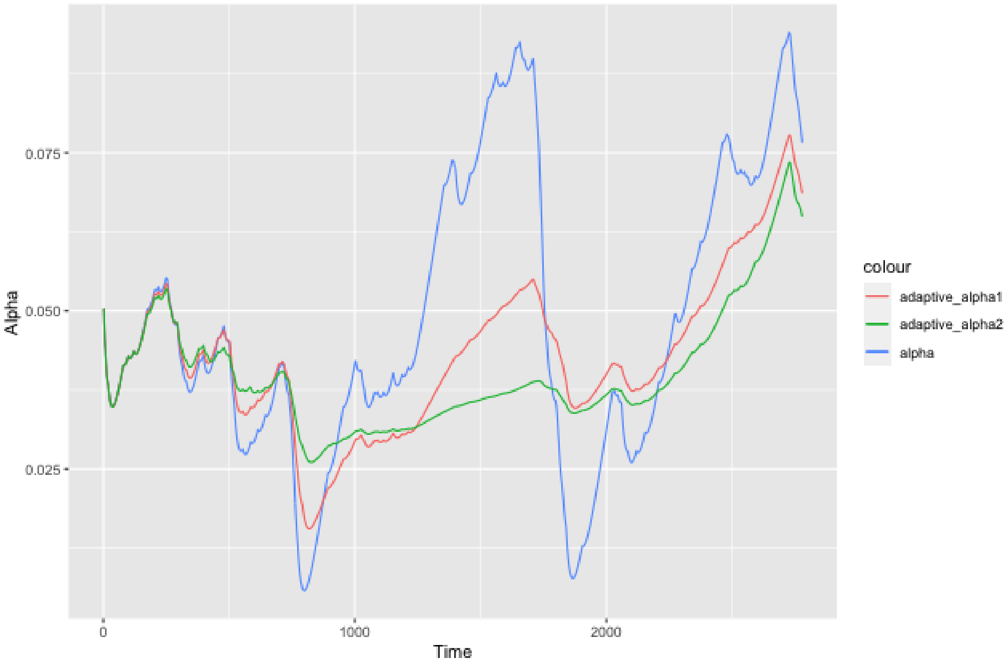


Fig 3.2.5. α values from adaptive gamma and fixed gamma method

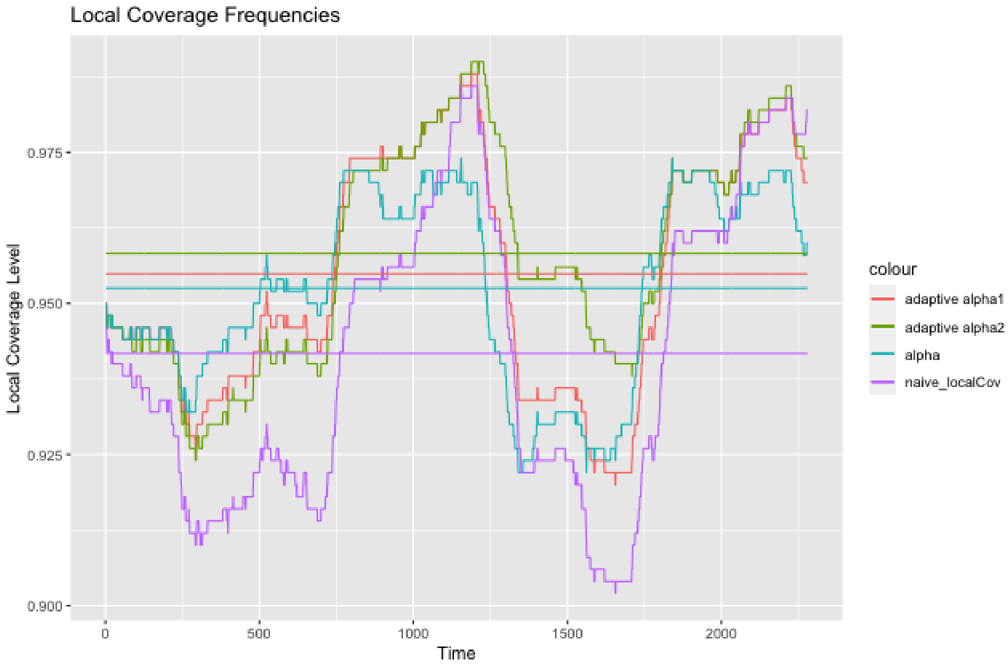


Fig 3.2.6. Local coverage frequencies from adaptive gamma and fixed gamma method

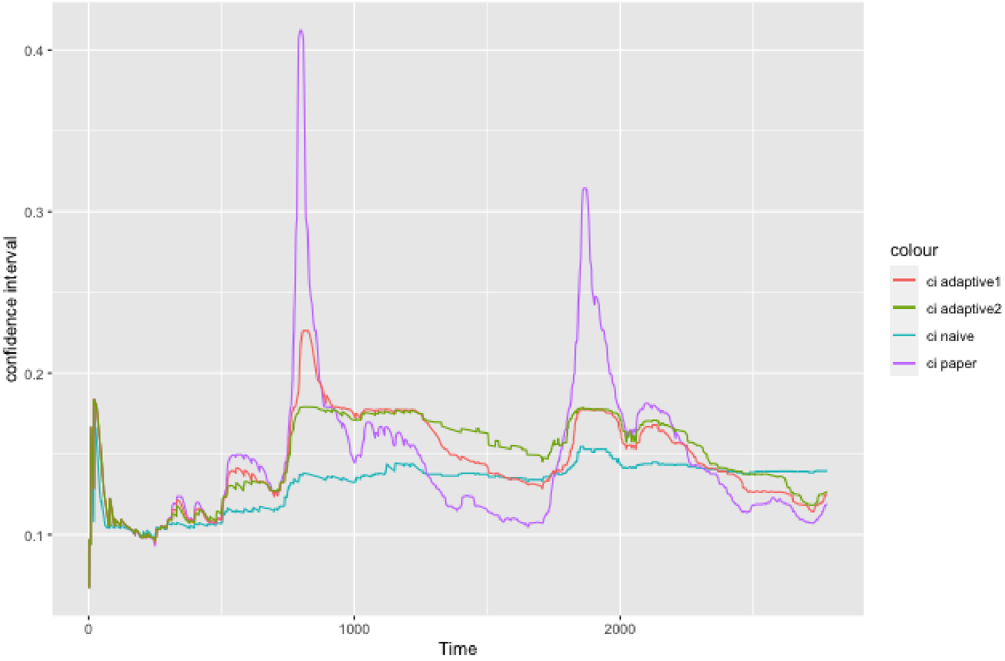


Fig 3.2.7. Confidence Intervals from adaptive gamma and fixed gamma method

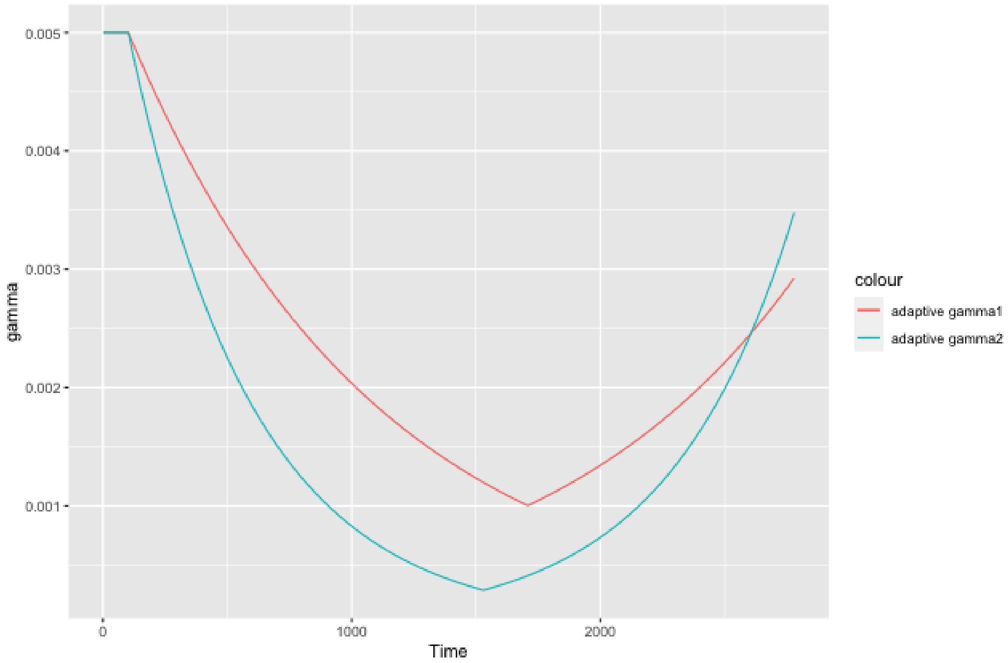


Fig 3.2.8. Gamma size change from gamma step size 0.001 and 0.002

4 Unconventional Data

In addition to developing an adaptive gamma to complement the adaptive alpha, we are also interested in how ACI may perform in other everyday settings. In considering common datasets describing the economy, politics, or geography, we noticed that they may have three unique features: cyclical patterns, sudden events, and white noise. For example, airport traffic ebbs and flows depending on the time of day, gas prices can suddenly skyrocket due to geopolitics, and telescope imaging can include static from image processing.

These scenarios are common in real life but may be considered more statistically unconventional. While the paper's two examples of stock prices and electorate counties may address these data types, it is unclear to what extent they exist. As such, we will simulate data that explicitly replicates these three features in order to evaluate how ACI performs in their most extreme cases.

For each condition, we will provide a hypothesis about how we expect ACI to perform, and then compare and analyze the actual results.

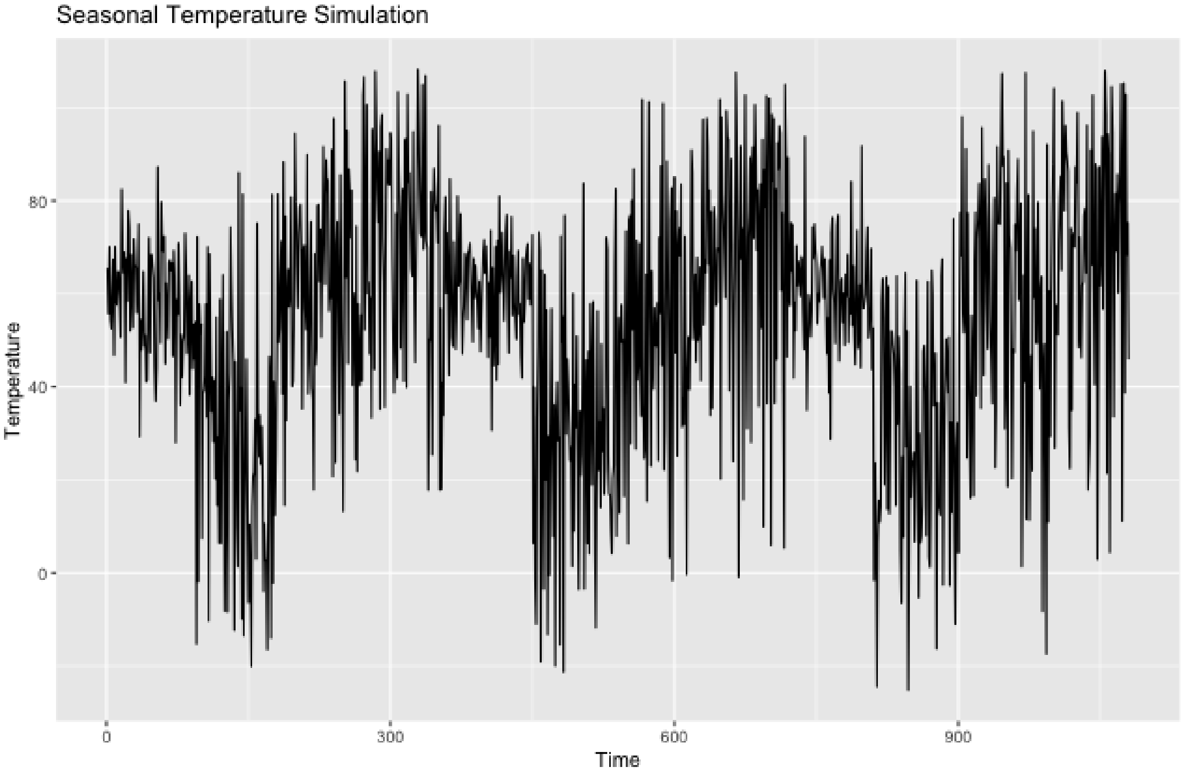
4.1 Cyclical Distribution Shifts

In this paper, we define a cyclical distribution shift as a sequence of distributions that repeat over time. Cyclic data is of interest to us because it presents two potential challenges: consistency and efficiency. The issue of consistency arises because sections from similar points in the cycle could yield very different predictions. While they may be statistically valid, those who rely on ACI output may find it to be less practically useful if they deviate from expected behavior. Furthermore, the issue of efficiency arises because intervals are always calculated despite the same behavior repeating. As such, there is likely an opportunity to reduce computational costs with pattern recognition. While this is not a strategy we will develop in this paper, it provides another reason to experiment with cyclical data.

In this paper, cyclical data allows us to examine how ACI performance depends on the regression model used. This is because we can fit a seasonal ARIMA model that theoretically can closely match the data's patterns, and then compare it to a random (in this case GARCH) model that only accounts for volatility. For this simulation, we will rotate through a collection of normal distributions, each of which we will draw from for a set number of times before the distribution changes. As such, we will fit a seasonal ARIMA model that is designed for a seasonal random walk, or ARIMA(0, 0, 0)(0, 1, 0).

The real-world example of cyclical patterns we model the data is upon is temperatures in Chicago during the four seasons of fall, winter, spring, and summer. Each season comes with its own unique range of temperatures and how they might vary, and each year, these four seasons repeat. We use normal distributions with different means and variances to model each season's temperatures, switching from one season to the next every 90 days. The total range of data is T=1080 days, or 3 years. The minimum and maximum allowed temperatures are Chicago's record low of -27 degrees and record high of 109 degrees Fahrenheit. It follows then that in this simulation, distribution shift only occurs when the season changes. Otherwise, temperatures from the same season will be drawn from the same given distribution for that season. The simulated data is graphed below in 4.1.0.

Fig 4.1.0. Seasonal temperature simulation



4.1.1 Hypothesis

For the cyclical data simulation, we will focus on how the regression model used affects ACI results. As previously mentioned, we will use ACI with both SARIMA and GARCH and compare the results between the two. Note that for this section, we have used α=0.1 and γ=0.1, with a lookback period of 50 and starting point of 100.

We expect the ACI using SARIMA to have higher local coverage specifically when seasons change. This is because we tell the SARIMA how many seasons there are so it can anticipate distribution shifts, whereas ACI with GARCH must "learn" this through trial and error. However, we expect ACI with SARIMA and GARCH to perform similarly right before seasonal changes, since both models have had an adequate range of time to accommodate to the new distribution.

4.1.2 Results

Below is both local coverage frequencies and interval sizes for the temperature data. First, we will group by adaptive versus fixed alpha to test our original hypothesis that ACI would perform better specifically when seasons change. Figures 4.1.1 and 4.1.2 show the differences in local coverage rates for each regression method. Contrary to what we expected, the mean coverage rate for both inference methods is higher when GARCH is used. Furthermore, if we examine time periods right after seasonal changes (T > 90, 180, 270, etc), we find that the GARCH and SARIMA methods alternate in which is performing better. The difference between coverage is also vast, reaching almost 0.1 at times.

Fig 4.1.1. Adaptive local coverage frequencies

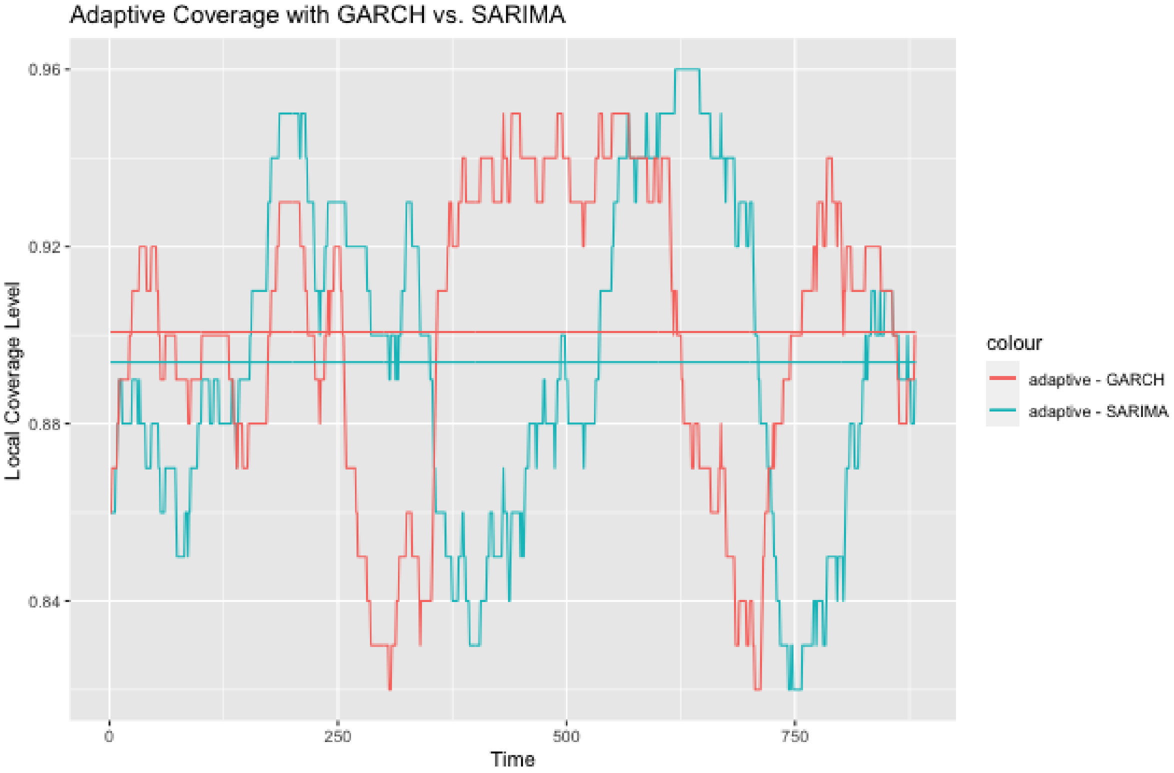
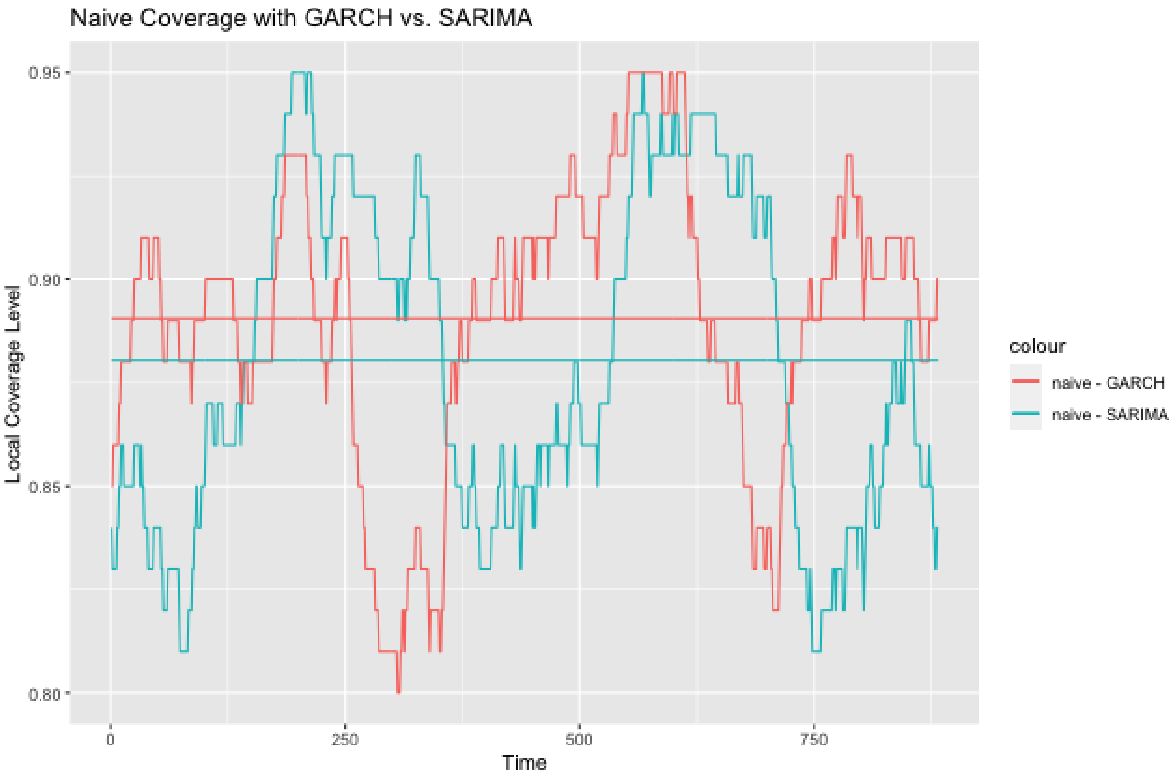


Fig 4.1.2. Naive local coverage frequencies 0



To further examine the difference in inference, we've outputted graphs of the actual interval sizes for each regression model in Figures 4.1.3 and 4.1.4. Similarly, both regression models yield fluctuating interval sizes. However, the SARIMA model in both alpha cases yields much larger interval sizes than GARCH, with the mean sizes being similar across both alpha cases as well.

Fig 4.1.3. Interval sizes using GARCH

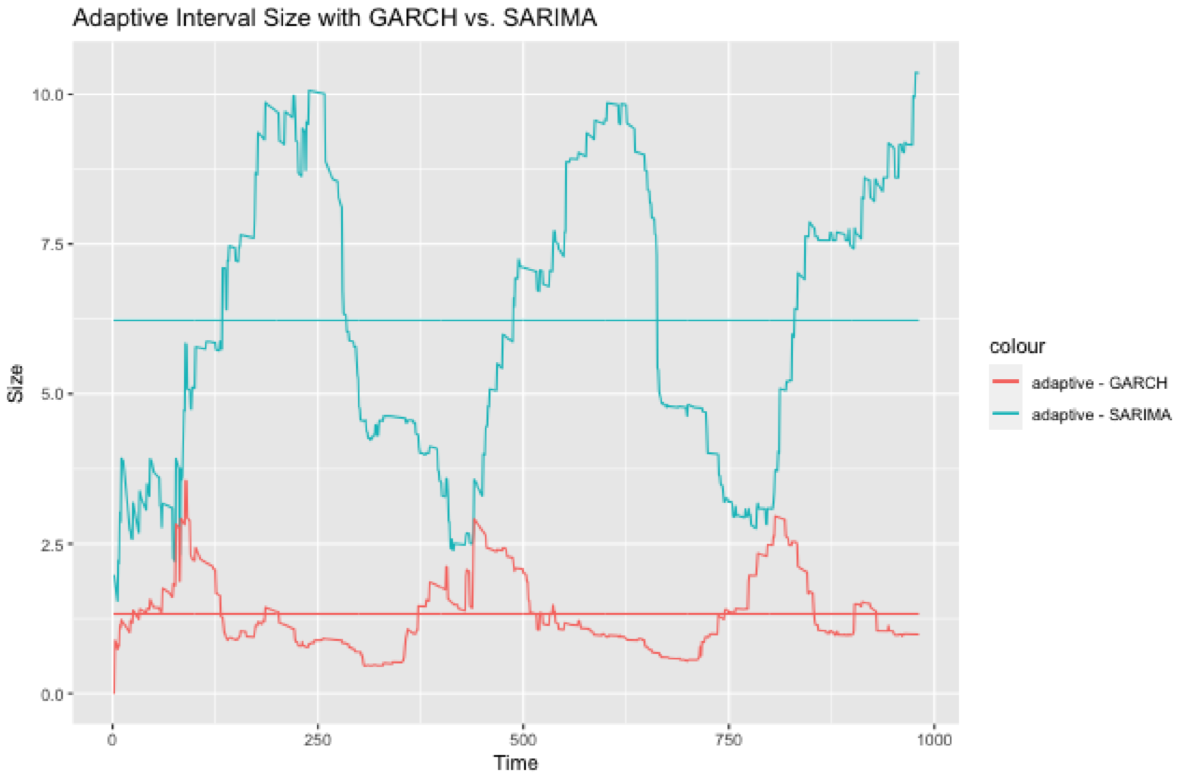
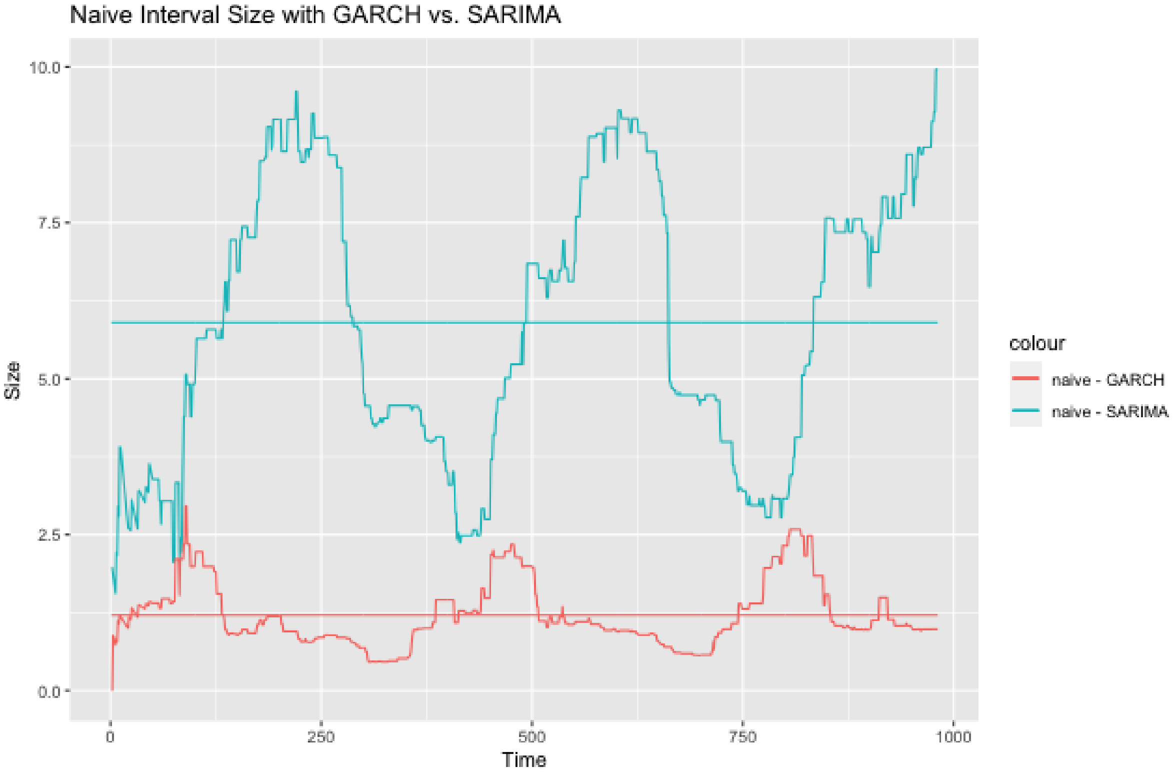


Fig 4.1.4. Interval sizes using SARIMA



We can also group by regression model rather than inference model to ask the separate question of which type of alpha fares better with a specialized regression. Since we found SARIMA to be worse performing than GARCH, this also doubles as asking the question of whether adaptive or fixed alpha will adapt better to a regression model that underperforms. As demonstrated in the paper, we have also included the Bernoulli sequence and the static alpha trends for comparison. Figures 4.1.1 and 4.1.2 demonstrate both adaptive and non-adaptive CI performance using the SARIMA regression. We see that in Figure 4.1.1 (GARCH), coverage dips at T=90, 180, and 360, for example, and this occurs for both adaptive and fixed alpha methods. In Figure 4.1.2 (SARIMA), we see that local coverage is higher for ACI compared to the fixed alpha inference at seasonal change time points such as T=90, 360, 540, 630, etc. The difference in local coverage is notable; for example, at T=90, ACI has a ¿ 0.05 increase in coverage over the fixed-alpha method.

Fig 4.1.3. Local coverage frequencies using GARCH

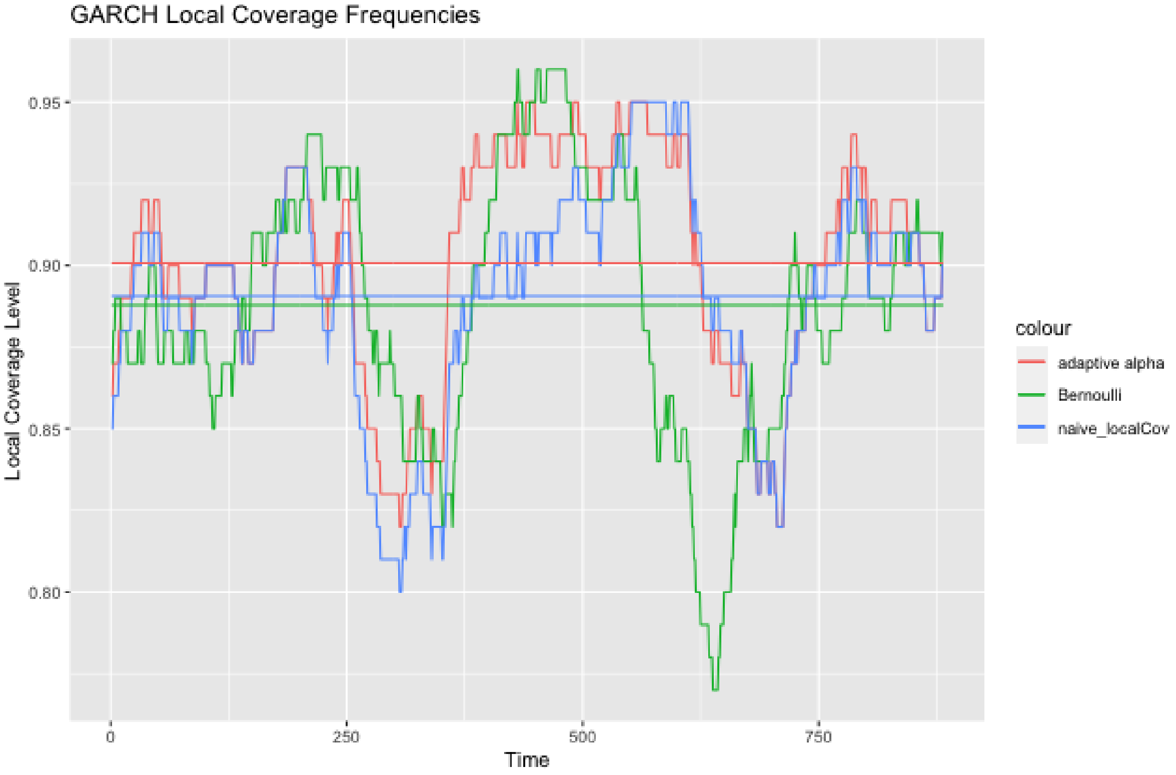
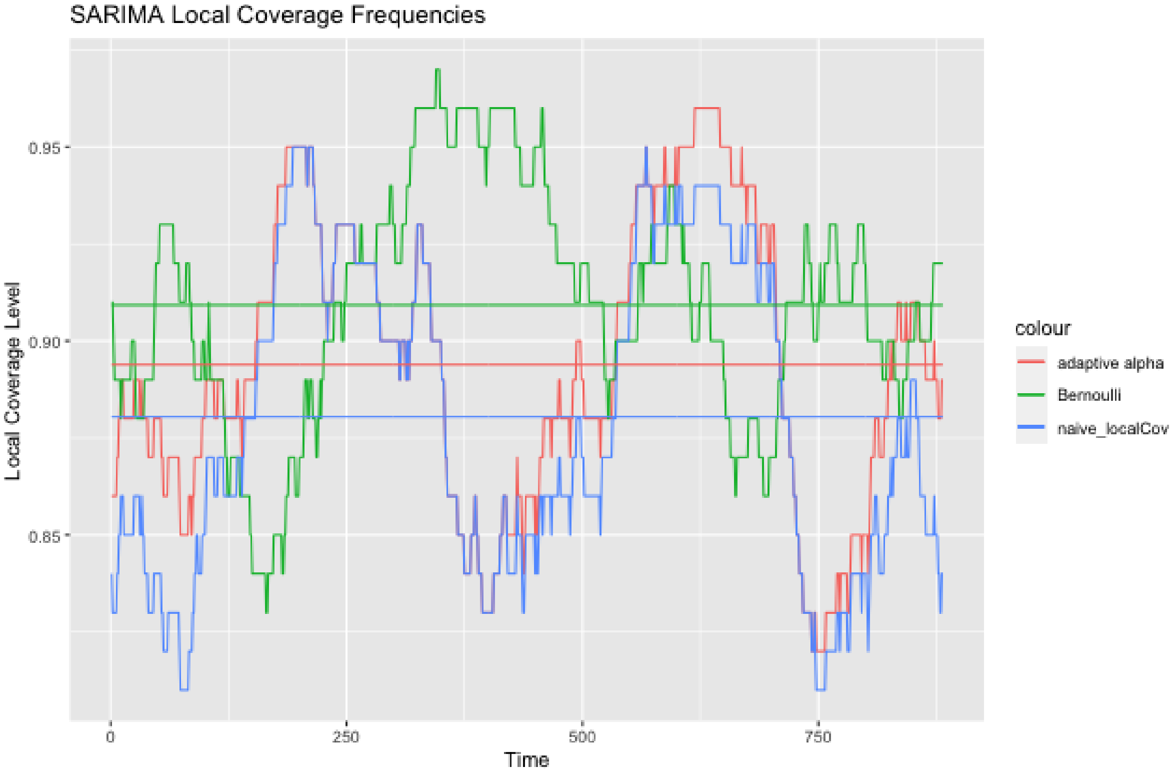


Fig 4.1.4. Local coverage frequencies using SARIMA



We will again also compare using the interval sizes to see whether there is a tradeoff between size and accuracy. In this case, we see that the results using adaptive alpha have marginally higher prediction interval sizes. Meaning, the higher local coverage was achieved by expanding the interval rather than a more accurate placement.

Fig 4.1.5. Interval sizes using GARCH

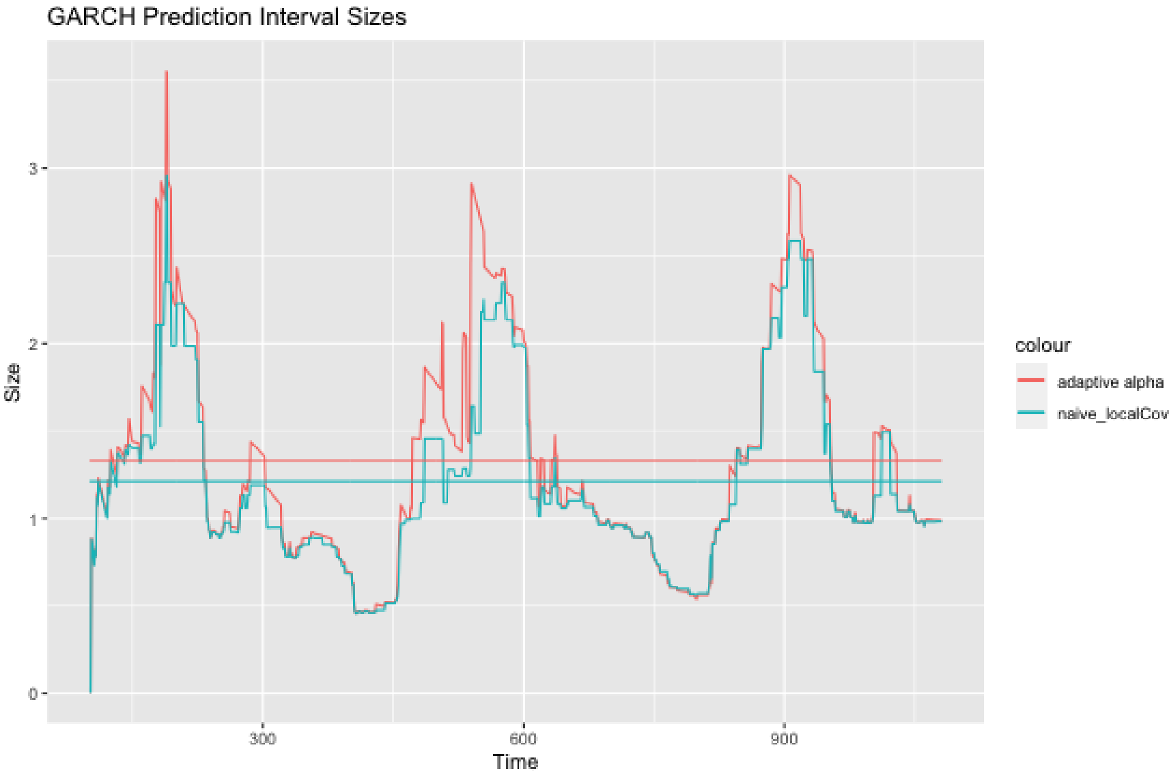
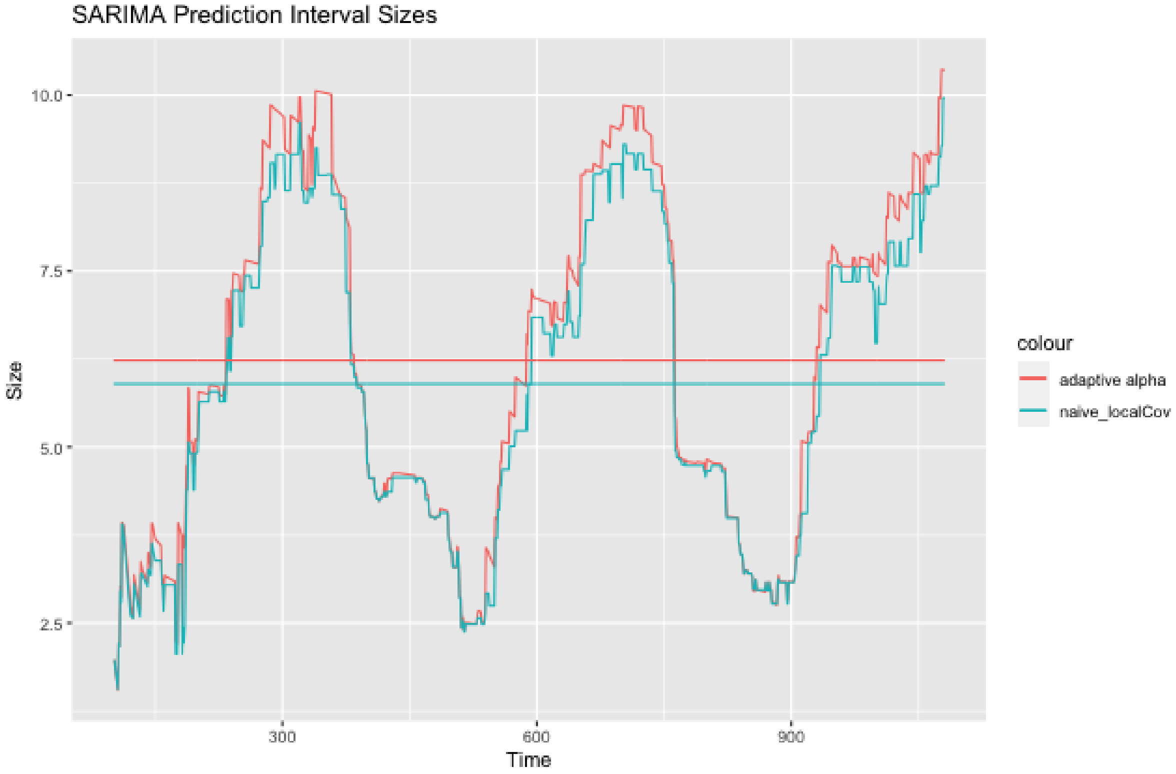


Fig 4.1.6. Interval sizes using SARIMA



Ultimately, in experimenting with cyclical data, we find that for either adaptive or fixed alpha, a seasonal regression model performs worse than a generic volatility model such as GARCH. This goes against our initial intuition that if a regression model was devised for a particular behavior, then the resulting inference would have better coverage with more accurate point predictions. One possible reason for this is there may be computational conflict when both the regression model and ACI attempt to account for the same shifts in distribution. However, it is also possible that the results only apply to cyclical distribution shifts. We would invite further discussion into the theoretical basis for these results, as well as any room for improvement in how the seasonal-aware regression model was implemented.

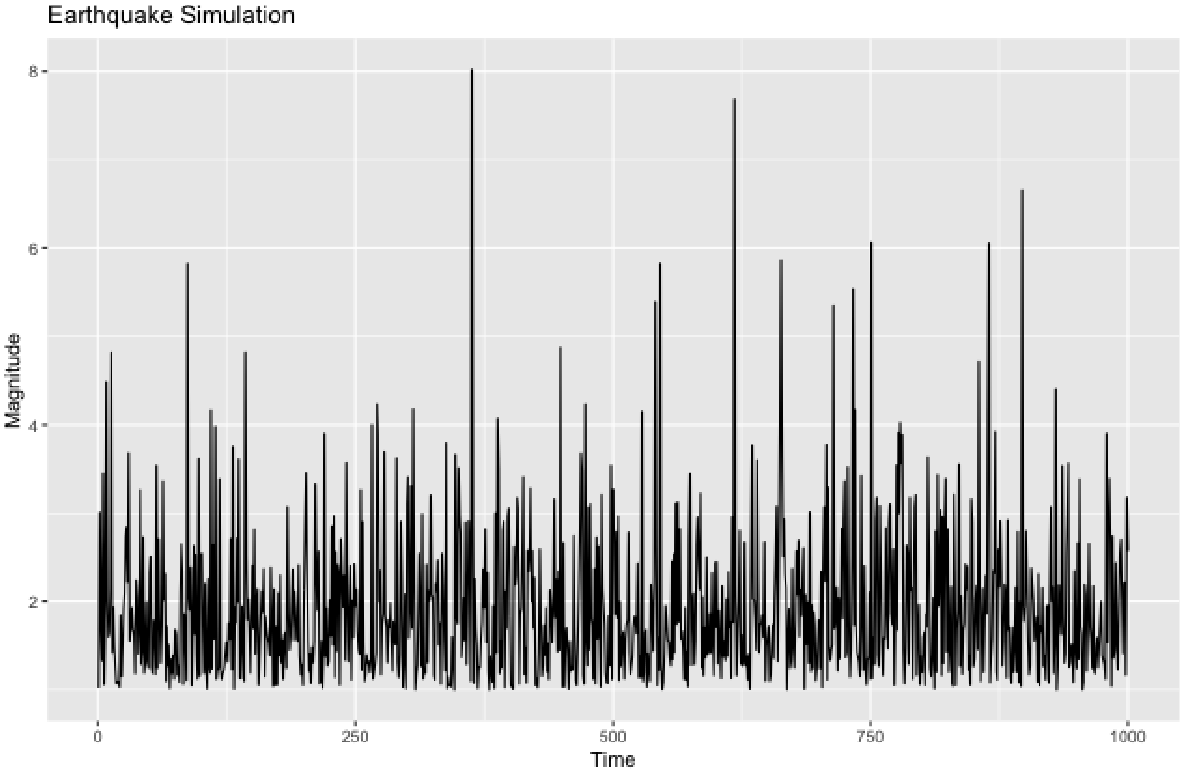
4.2 Data with Extreme Events

The next irregular time-series data will simulate the occurrence of extreme events. We define an extreme event as a Y_t that is much higher than its previously recorded points. We are interested in extreme shocks to evaluate both ACI's statistical robustness and its performance in real-world scenarios. Extreme shocks may challenge ACI for two reasons. First, a constant gamma would not be able to adjust to the step size required to accommodate for an extreme shock quickly enough. Meaning, it is unlikely that the predictions may ever estimate such an extreme metric at the time it occurs. Secondly, it follows that when gamma does adjust, ACI may overshoot for a time after the shock has occurred.

The real-world example of extreme shocks we model this data set upon is earthquake magnitudes. For the most part, the region near an active fault line will experience minor magnitudes, since major earthquakes occur very sparsely across the globe. We aim to simulate this behavior for a hypothetical region, measuring earthquake magnitude based on the Richter magnitude scale, since the moment magnitude scale is lesser known. The Richter scale begins at a minimum of 1.0 and theoretically has no maximum, although we cap our simulated magnitudes at the world record of 8.6. Hence, the data will typically be drawn from a distribution that falls within a ”micro” to ”minor” range, with a mean that has a slight random deviation from 2.0 on the scale. With probability $\frac{1}{250}$ and $\frac{1}{100}$, we will determine whether our fault region experiences a moderate (mean of 4.0) or major (mean of 7.0) earthquake a given time point. The denominators were chosen as the number of years it might take for another earthquake of that given scale to occur (i.e. a major earthquake once every 250 years).

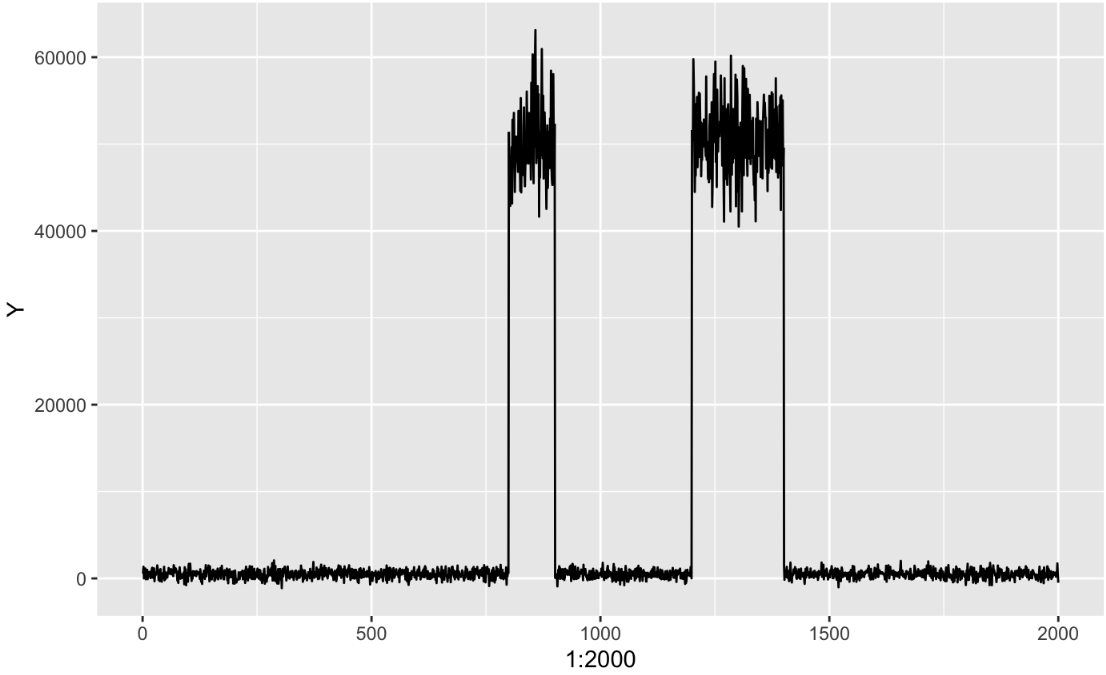
As such, in this simulation, distribution shifts occur in two ways. The first is from occasional shocks due to moderate and large earthquakes. The second is that when minor magnitudes are drawn, we include a random change in mean to the normal distribution for minor magnitude earthquakes. The results from this data simulation are plotted below in Figure 4.2.0.

Fig 4.2.0. Earthquake data magnitudes



We also considered another simulated data set with two covariates, X_1 and X_2. In this data set, the extreme events took place from $t = 800, ..., 900$ and $t = 1200, ..., 1400$. During these times we had $X_1 \sim \mathcal{N}(100, 1000)$ and $X_2 \sim \mathcal{N}(200, 2000)$ with $Y = X_1 + 2X_2 + \sigma\epsilon$. During the extreme events, we had $\sigma = 100$ and $\epsilon \sim \mathcal{N}(500, 10)$. Outside of the extreme events, we had $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathcal{N}(1, 1)$ with $Y = 10X_1 + 5X_2 + \sigma\epsilon$. Here, we had $\sigma = 100$ and $\epsilon \sim \mathcal{N}(5, 5)$. In this simulation, we have extreme distribution shift not only in the marginal distributions of X_1 and X_2, but also in the conditional distribution of Y given X_1 and X_2. The plot for Y can be seen below in figure 4.2.1.

Fig 4.2.1 Extreme Event Simulation



We used a linear regression of Y on X_1 and X_2 as the model in the ACI with $\alpha = 0.1$. For our conformal score function, we used the square of the residuals of our linear model. We then followed the same steps for a simple adaptive conformal inference.

4.2.1 Hypothesis

We believe ACI could both fail to capture the extreme event and allow the event to skew its predictions afterward. We are interested in testing whether ACI is better in comparison to the naive confidence intervals at adjusting back to predicting the usual distributions after extreme shock events. We will use a naive and an adaptive confidence interval to determine both the coverage rate and the confidence interval size of our predictions as we observe extreme shocks.

4.2.2 Results

The results from the earthquake simulation are represented in Figures 4.2.2 and 4.2.3.

Fig 4.2.2. Local Coverage Frequencies for Simulated Earthquakes

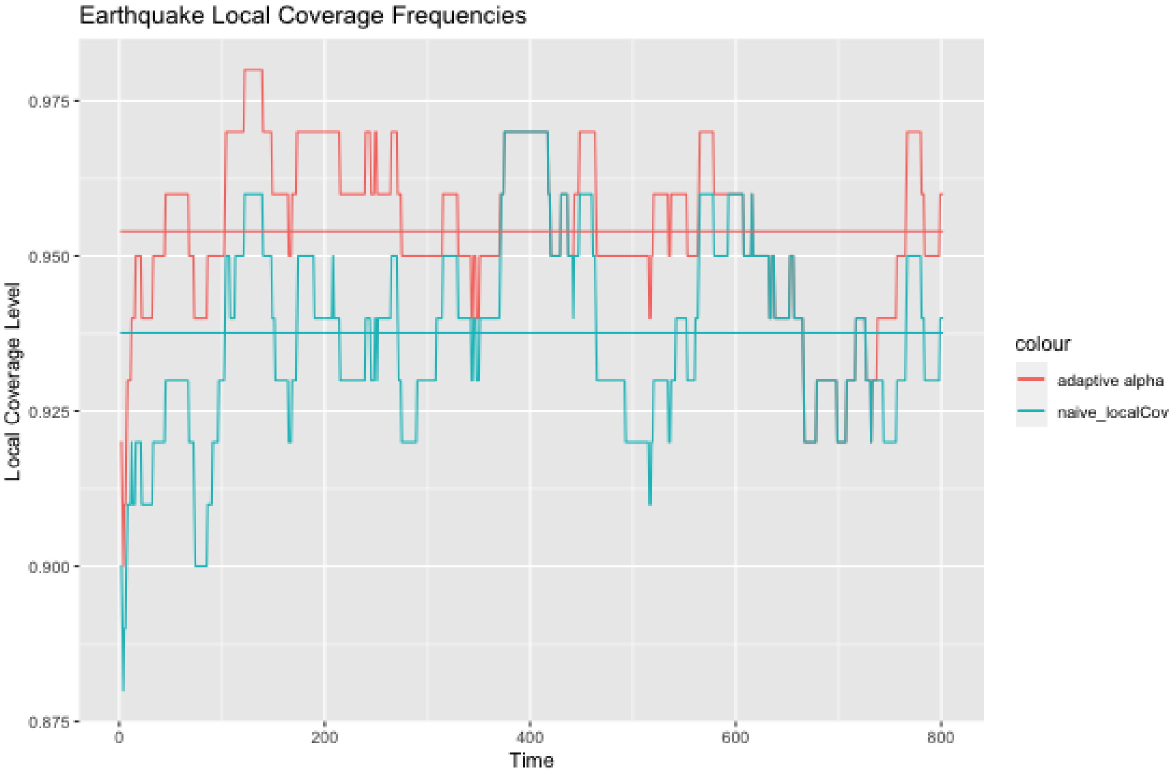
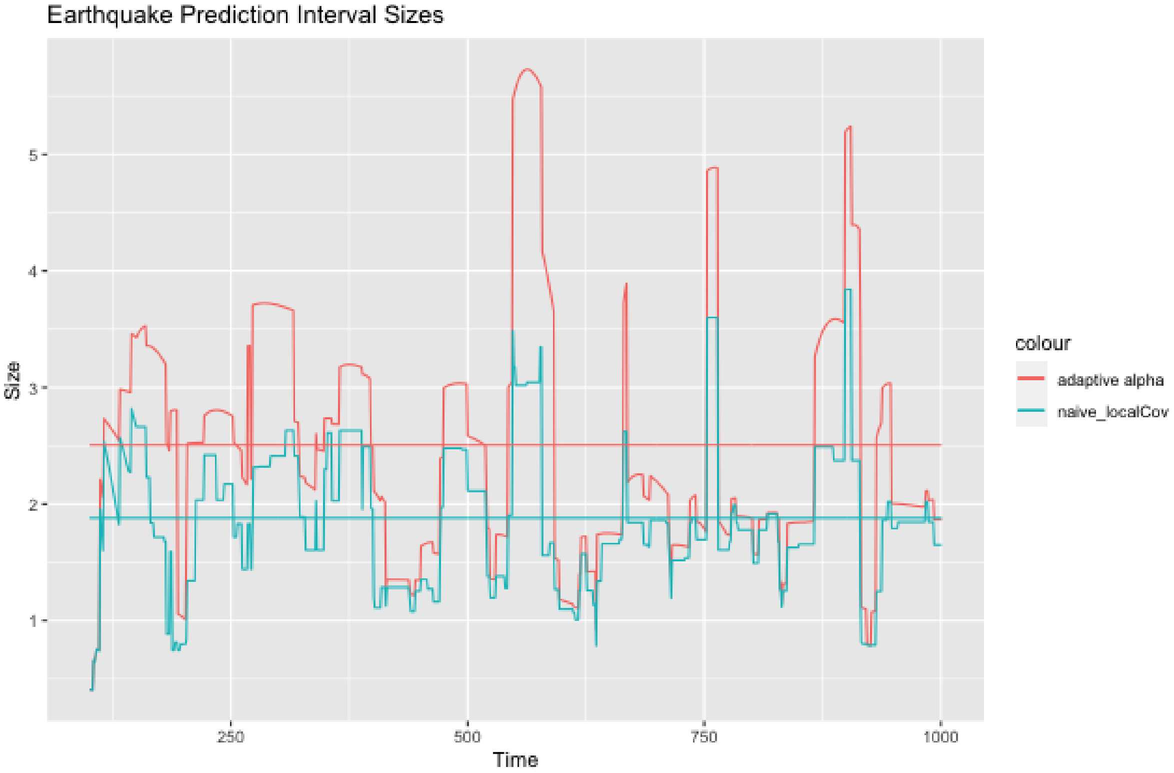


Fig 4.2.3. Confidence Intervals for Simulated Earthquakes



For the simulation of extreme events through the example of earthquakes, we used $\alpha = 0.05$ to compute a confidence interval using both the adaptive and naive methods. The adaptive method consistently showed a local coverage level close to $0.95 = 1-\alpha$ while the naive local coverage rate was lower. To achieve consistent coverage rates, the adaptive α method increased α_t during earthquake shocks, making the prediction interval size increase. The increased size of the adaptive α intervals on average contributes to an increase in Earthquake coverage frequency for the adaptive method.

As expected, we also see that post-earthquake estimations are skewed because of their larger magnitudes. In Figure 4.2.3., confidence interval sizes rise occasionally above the average size, sometimes moderately and other times more dramatically. The times the interval sizes spike at roughly correspond to the peaks in Figure 4.2.0. Interestingly, however, some earthquake occurrences did not affect the size. For example, the largest simulated earthquake occurred between T=250 and T=500, but there is no corresponding large spike in interval size during that time.

Now, we will analyze the results from the simulation with covariates.

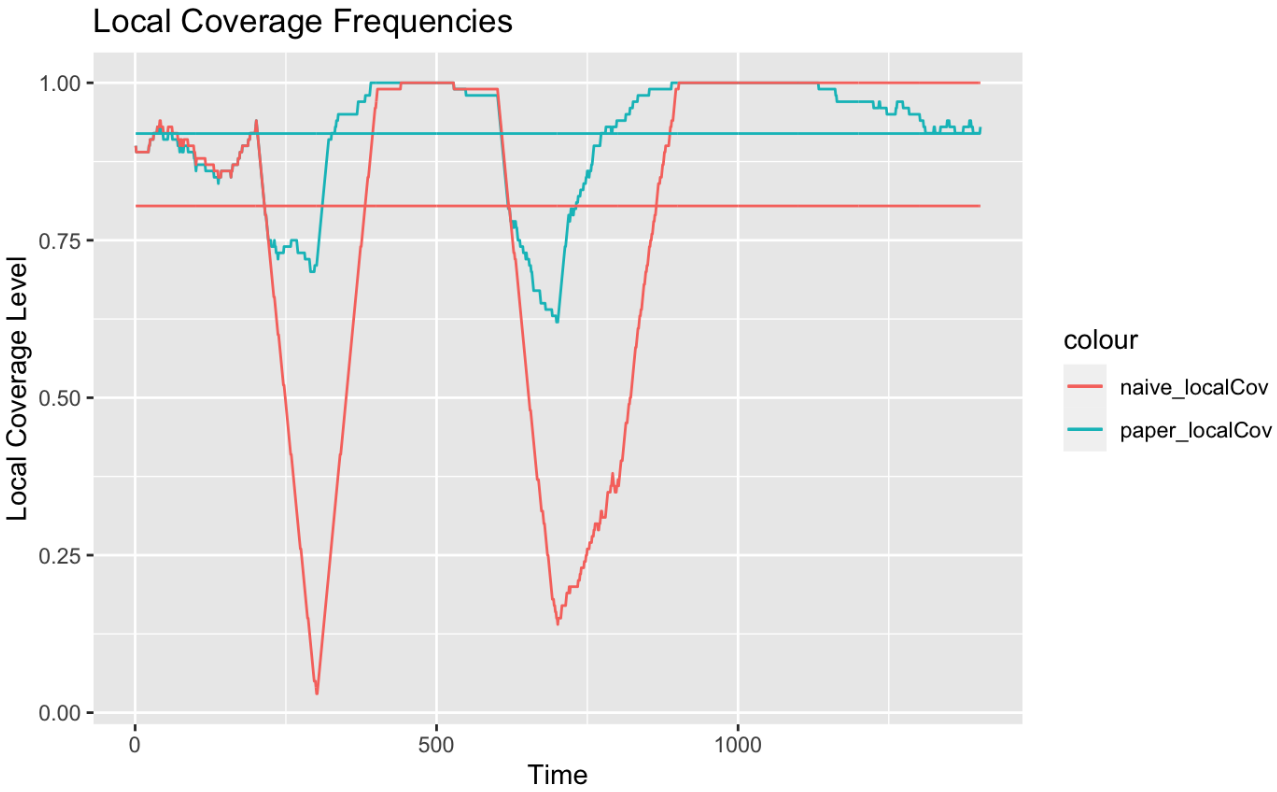


Fig 4.2.4. Local Coverage Frequencies for Extreme Events

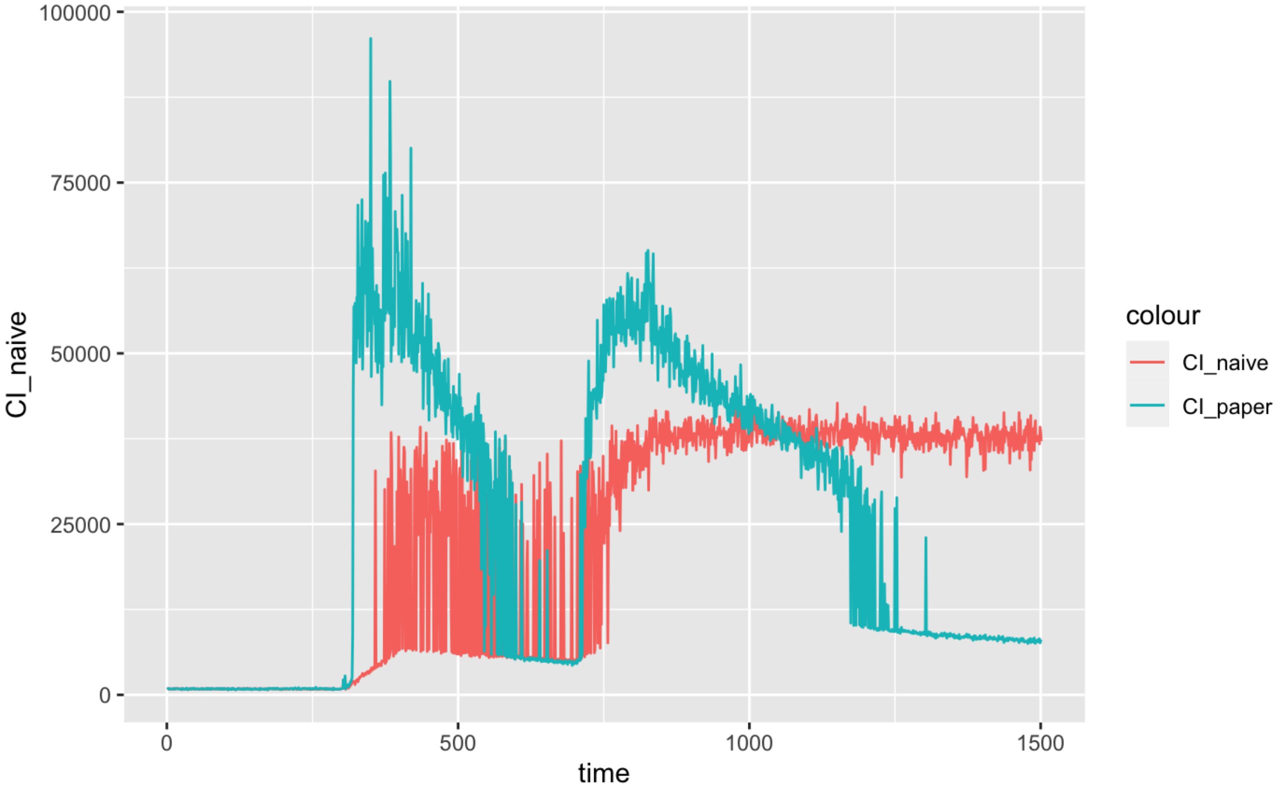


Fig 4.2.5. Confidence Intervals for Extreme Events

From the second simulation for extreme events, the local coverage rate remains close to 0.9 with $\alpha = 0.1$. This is because, during extreme events, ACI reacts to distribution shifts by increasing α_t and thus increasing the confidence interval size. This effect can be seen in Figure 4.2.5. However, after the extreme event has passed (around time 450), the ACI confidence interval size decreases in size as it recognizes and adapts to the distribution shift back to "normal". Since the simulated data draws from very simple normal distribution shifts, it is not difficult for the method to "learn" from past points to anticipate future changes.

On the other hand, the naive confidence interval, which does not recognize local coverage rate as a factor in computing new confidence intervals, relies on consistently large confidence interval sizes in order to achieve a significant coverage rate. Even so, due to the extreme events and the inability for the naive confidence intervals to recognize the change in distribution shifts, the coverage level for the naive confidence intervals is significantly lower than ACI during the extreme events. The confidence interval size of the naive method, while not as large as ACI confidence intervals during the extreme events, stays at a large size even after the extreme event has passed. On the other hand, the ACI method recognizes the change in distribution after the extreme event and, through the adaptive α_t, adjusts the confidence interval back to a more reasonable size while making the same, and often better, local coverage rates.

Unlike the earthquake simulation where extreme shocks were more frequent, the second extreme event simulation had extreme events for an extended period of time and "breaks" between extreme events. This allowed the adaptive α method to adjust to the distribution shifts and return to confidence intervals of a reasonable size. However, we can see that the naive method is unable to recognize the change back to the distribution after the interruption of an extreme shock, even after extreme events have passed.

From our exploration of ACI in extreme events, we have noticed that the adaptive power of ACI allows the confidence interval size to adjust according to the distribution at the moment instead of using a catch-all larger confidence interval as the naive method. In future explorations, we hope to determine how momentum-adaptive ACI or adaptive γ ACI may impact how quickly the ACI responds to distribution shifts and how the confidence interval size changes before and after extreme shocks. We believe that a more precise γ through the adaptive γ method may improve the ability of ACI to respond to extreme shocks.

5 Conclusion

Adaptive conformal inference improves upon fixed-alpha conformal inference by making it agnostic of both a data set's covariate distribution and regression model. Building upon these ideas, we explored both changes in its step-wise mechanism and new applications in data, with a particular emphasis on more unconventional randomness like a cyclical or extreme event component.

For our study in stepsize, we tried both a randomized and adaptive stepsize. The randomized stepsize was based on previous research that suggested it can improve stochastic gradient descent. However, we found that randomized stepsize performed similarly to the paper's fixed gamma ACI. On the other hand, the adaptive stepsize achieved better coverage levels and narrower prediction intervals. For the specific data set this was tested on, we found that a slightly larger gamma size performed better on these metrics as well.

For our study in unconventional data, we found that a seasonal autoregressive model does not lead to better inference for either adaptive or fixed-alpha methods, which we expected to happen. However, our hypothesis regarding data with extreme shocks was affirmed, since we found that after an extreme event, inference became too extreme for subsequent data points.

The process of writing this paper also led to a few more ideas for exploration. Although we have evidence that support how adapting gamma size can improve the ACI method, there are some observations that also require further exploration. Firstly, smaller adaptive gamma leads to higher coverage level and narrower confidence interval. Also, although adaptive gamma was able to achieve similar coverage level with a narrower confidence interval, for predictions following the extreme distribution shift, adaptive gamma method returned wider confidence interval than that of fixed gamma method. Additionally, we could also incorporate more erratic distribution shifts. For example, the mean of a distribution could change drastically for every change in time.

6 References

Blier, Léonard, et al. “Learning with Random Learning Rates.” ArXiv.org, 29 Jan. 2019, https://arxiv.org/abs/1810.01322.

Musso, Daniele. “Stochastic Gradient Descent with Random Learning Rate.” ArXiv.org, 11 Oct. 2020, https://arxiv.org/abs/2003.06926.

Gibbs, Isaac, and Emmanuel Candès. “Adaptive Conformal Inference under Distribution Shift.” ArXiv.org, 28 Oct. 2021, https://arxiv.org/abs/2106.00170.