A Simulation Study of Extreme Temperatures

Kevin O'Connor

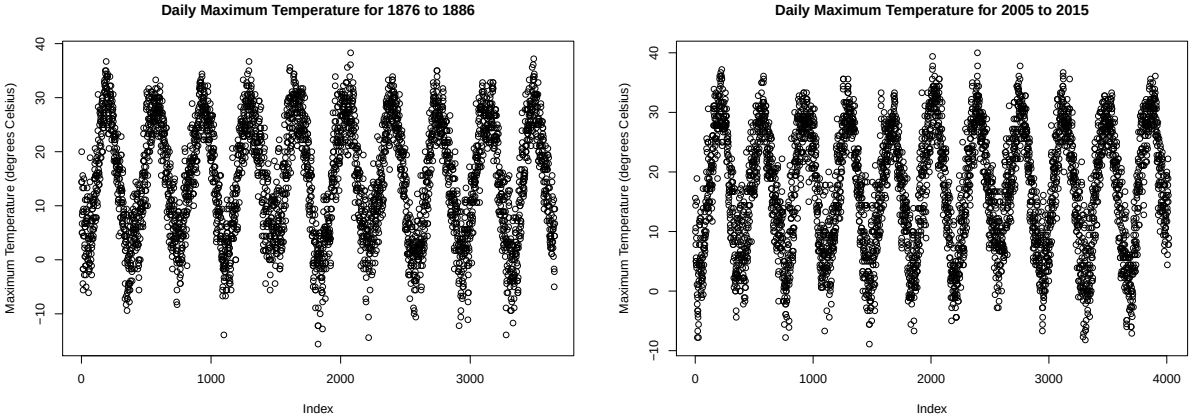
May 31, 2016

Introduction

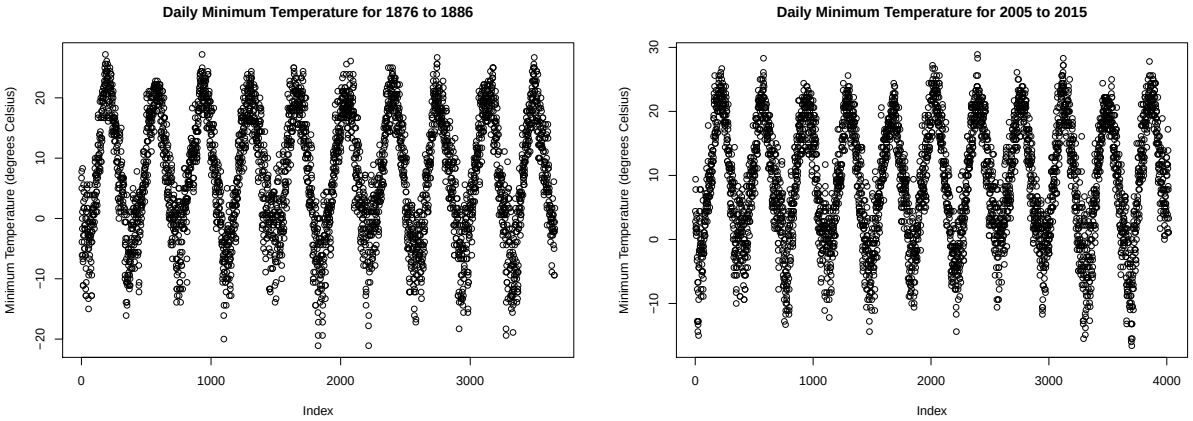
In this paper, I introduce and investigate the results of a simulation study on extreme temperature models. This work is motivated by a recent paper by Stein [1], in which the limiting GEV distribution for annual maximum temperatures are cast into doubt. Such a result is achieved by imposing an upper bound on the maximum temperature on any given day which varies throughout the year. In the work that follows, I present such a model for daily maximum temperature and try to emulate the Central Park data used in Stein's paper. I then investigate the properties of the model via simulation, including the distribution of its maximum to assess the validity of the GEV distribution in this case. The code for the simulations can be found on GitHub at the link provided at the end of this document.

Data Exploration

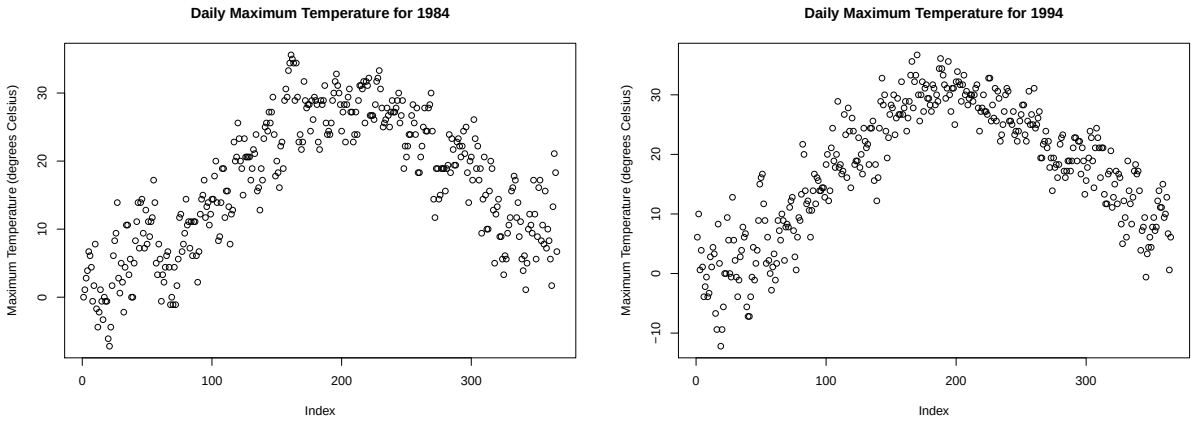
Before developing the model, it will be helpful to look at the Central Park data for defining characteristics. Below, plots of the daily maximum from 1876 to 1886 and 2005 to 2015 are given.



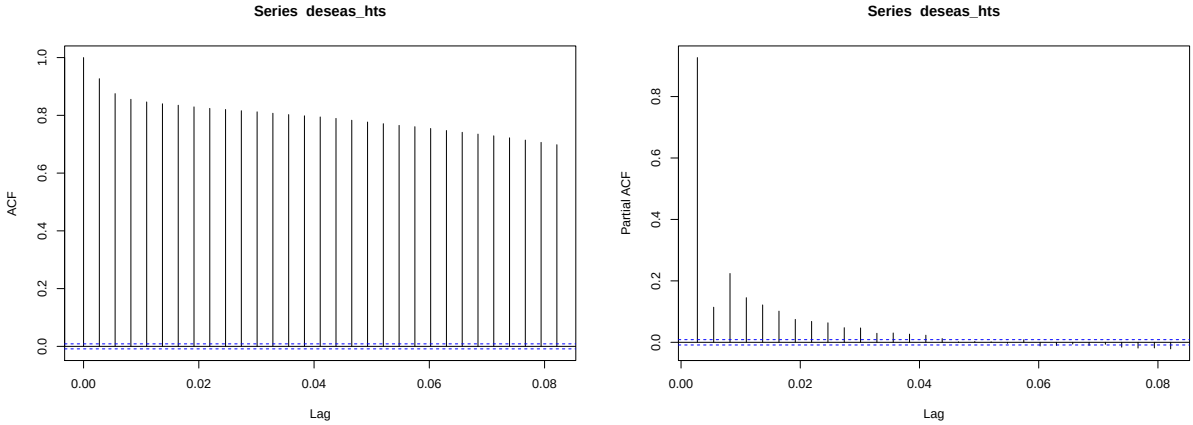
Similarly, plots of the daily minima are given below as well.



Besides the obvious seasonality, we note that the daily extreme temperatures tend to vary as a triangle wave. We can observe this by looking at individual years. Plots of the daily maximum temperatures for 1984 and 1994 can be found below.



These give further support for the triangle wave model of daily extreme temperature. However, such a choice of bounds on the daily maximum temperature gives rise to issues of undifferentiability which will be addressed later. Furthermore, we can look at the dependence for the data by first removing the seasonality via a Loess smooth and calculating the ACF and PACF.



We can see from the PACF that significant dependence exists up to about lag 10 based on the Loess smoothing method. Furthermore, the covariance seems to decay exponentially with increasing lag.

So with this data in mind, we can move on to develop our model.

Model Development

As suggested by Stein, we impose an upper bound $u(i/n)$ and a lower bound $l(i/n)$ on the daily maximum temperature such that $l(i/n) < u(i/n)$ for any i and $l((i+an)/n) = l(i/n)$ and $u((i+an)/n) = u(i/n)\ \forall a \in \mathbb{Z}$. Thus, we have set n as the period of the bounds but have left out a phase shift parameter. This is done for simplicity as the specific day which we label as the first day is arbitrary. So, we can select it such that the phase shift disappears. As suggested earlier, the Central Park data indicates that a triangle wave might be a good fit for both $u(i/n)$ and $l(i/n)$. Despite differentiability issues, we let

$$u(i/n) = u_1 \frac{2}{a}\left(t - a\left\lfloor \frac{t}{a} + \frac{1}{2}\right\rfloor\right)(-1)^{\left\lfloor \frac{t}{a}+\frac{1}{2}\right\rfloor} + u_0$$

$$l(i/n) = l_1 \frac{2}{a}\left(t - a\left\lfloor \frac{t}{a} + \frac{1}{2}\right\rfloor\right)(-1)^{\left\lfloor \frac{t}{a}+\frac{1}{2}\right\rfloor} + l_0$$

We will fix this choice of bound for the model. We could have also selected a superposition of sines and cosines that approximated a triangle wave and more closely fit our data but this would introduce too many parameters than would be practical.

Now, we must consider how the daily maximum temperature should be distributed within the bounds. The

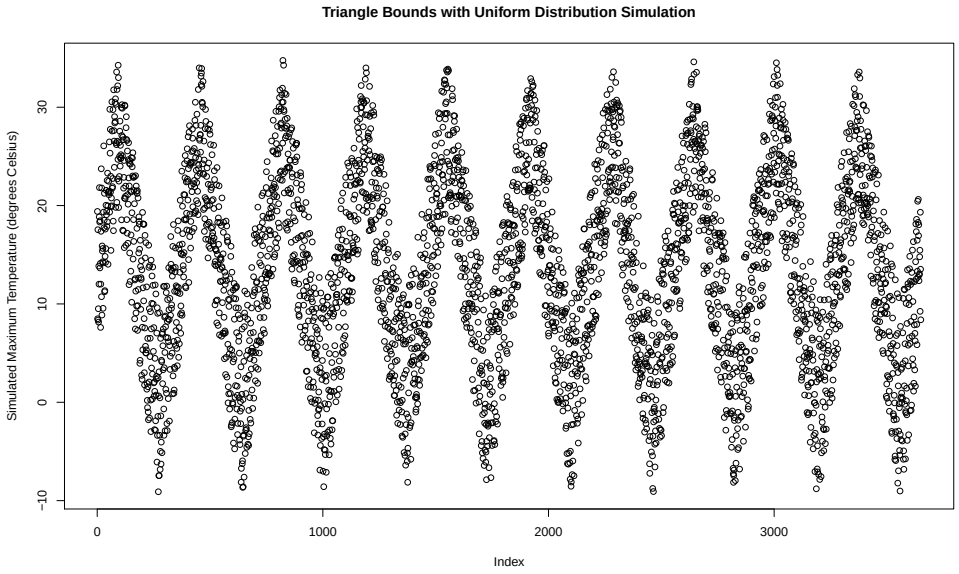
simplest case would be that of the uniform distribution. Specifically, supposing independence from day to day, we could let the maximum temperature on a day i, X_{in}, be distributed as

$$X_{in} \sim \text{Uniform}[l(i/n), u(i/n)]$$

such that

$$F_{X_{in}}(x) = \frac{x - l(i/n)}{u(i/n) - l(i/n)}$$

A simulation was carried out using this model with empirically determined parameters over a simulated time period of 10 years and can be seen below.

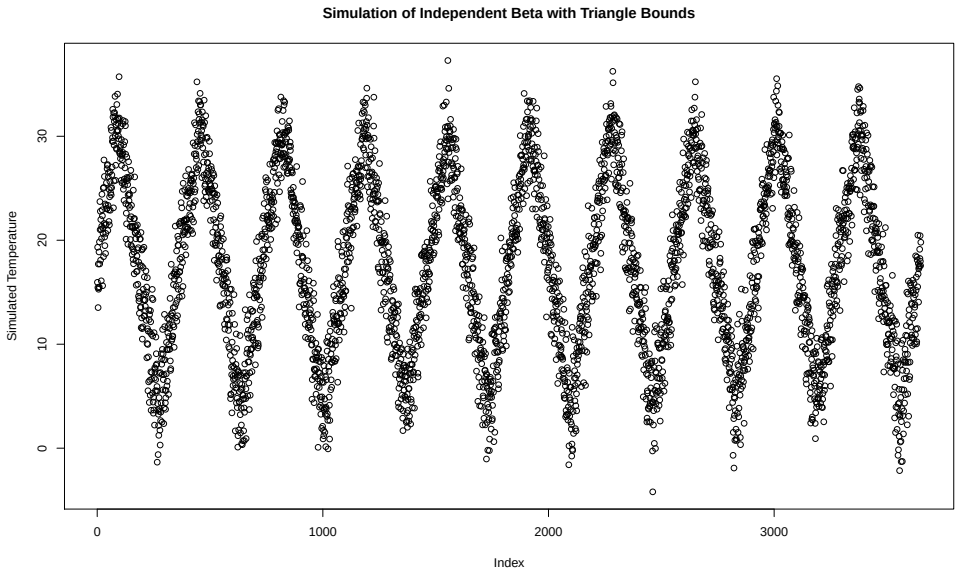


The plot indicates the presence of two issues with this model. For one, the data follow the bounds too closely. As such we see many values occurring near the bounds on any given day. This is not observed in the Central Park data. And second, we find that during the extreme times of the year, consecutive extreme observations occur infrequently. This relates to the need for some underlying dependence structure that makes consecutive similar values more likely, and therefore making it more likely for day-specific record temperatures to occur in the same year as outlined by Stein.

To address the first issue, we can use the shifted and scaled beta distribution instead of the uniform distribution. Specifically, still assuming independence, we suppose

$$X_{in} \sim \text{SSBeta}(\alpha, \beta, l(i/n), u(i/n))$$

A short treatment of some basic properties of the shifted and scaled beta distribution can be found in the appendix. A plot of 10 years of daily maximum temperatures simulated from this distribution can be found below.



Ignoring scaling, we see that this somewhat captures the roughness at the edges of the Central Park data.

Now, we may ask how the second discrepancy between our model and the data, that regarding the dependence structure, might be addressed. Most simply, we may write our model as an autoregressive process of order p. In former work, this model was written as

$$X_{tn} = \sum_{i=1}^{p} \phi_i X_{(t-i)n} + \phi_0 \left[(u(t/n) - l(t/n)) b_{tn} + l(t/n) \right]$$

where $b_{tn} \sim \text{Beta}(\alpha, \beta)$ and $\sum_{i=0}^{p} \phi_i = 1$. However, for large p, this acts to over-smooth X_{tn} especially during times of the year when $u(t/n)$ and $l(t/n)$ are changing rapidly. So, in order to introduce dependence without smoothing, we can write our model like so,

$$X_{tn} = (u(t/n) - l(t/n)) \left[\sum_{i=1}^{p} \phi_i \frac{X_{(t-i)n} - l((t-i)/n)}{u((t-i)/n) - l((t-i)/n)} + \phi_0 b_{tn} \right] + l(t/n)$$

So in effect, X_{tn} depends on the past p observations with their seasonality removed.

Choice of Coefficients

It is worth investigating whether imposing a specific functional form on the coefficients might simplify parameter estimation. As evidenced by the autocorrelation and partial-autocorrelation functions for the Central Park data, it would make sense first of all to impose the condition that $\phi_i \leq \phi_j$ for all $i > j$. In past work we assumed that $\phi_i = 1/(1+p)$ for all i. However, better results may be achieved via a Triangular or Epanechnikov kernel function. In the case of the triangular kernel, for $|u| \leq 1$, we have the function

$$K(u) = 1 - |u|$$

Whereas for the Epanechnikov kernel, again for $|u| \leq 1$, we have

$$K(u) = \frac{3}{4}\left(1 - u^2\right)$$

Of course, since we are considering discrete values of u for $|u|$ not necessarily ≤ 1 and considering only past observations in computing X_{tn}, we must modify these functions. For the triangle kernel, we can write it as

$$K(i) = \frac{2p - 2i + 1}{2(p+1)^2}$$

for $i \in \{0, ..., p\}$, while for the Epanechnikov kernel we write,

$$K(i) = \frac{3(p+1)^2 - 3i^2 - 3i - 1}{2(p+1)^3}$$

for $i \in \{0, ..., p\}$. The details of these transformations can be found in the appendix. These two choices of weighting function, in addition to the constant weighting, will be explored in later simulations.

Expectation and Dependence

We now solve for the expectation of X_{tn} under this model.

$$\begin{aligned}\mathbb{E}X_{tn} &= \mathbb{E}\left[(u(t/n) - l(t/n))\left[\sum_{i=1}^{p}\phi_i \tfrac{X_{(t-i)n} - l((t-i)/n)}{u((t-i)/n) - l((t-i)/n)} + \phi_0 b_{tn}\right] + l(t/n)\right] \\ &= (u(t/n) - l(t/n))\left[\sum_{i=1}^{p}\phi_i \mathbb{E}\left[\tfrac{X_{(t-i)n} - l((t-i)/n)}{u((t-i)/n) - l((t-i)/n)}\right] + \phi_0 \mathbb{E}b_{tn}\right] + l(t/n)\end{aligned}$$

Now, supposing that the mean of the standardized temperature on any given day is some constant μ, we write

$$= (u(t/n) - l(t/n))\left[\sum_{i=1}^{p}\phi_i\mu + \phi_0 \tfrac{\alpha}{\alpha+\beta}\right] + l(t/n)$$

and find that

$$\mathbb{E}X_{tn} = (u(t/n) - l(t/n))\frac{\alpha}{\alpha+\beta} + l(t/n)$$

Next, we wish to consider the conditional distribution of X_{tn}. Suppose we condition X_{tn} on $\mathcal{F}_{t-p,t-1}$, which we use to denote the smallest σ-algebra such that $\{X_{(t-i)n}\}_{i=1}^{p}$ is measurable. We can see that

$$X_{tn}|\mathcal{F}_{t-p,t-1} \sim \text{SSBeta}\left(\alpha, \beta, (u(t/n) - l(t/n))\sum_{i=1}^{p}\phi_i\tilde{b}_{(t-i)n} + l(t/n), (u(t/n) - l(t/n))\left[\sum_{i=1}^{p}\phi_i\tilde{b}_{(t-i)n} + \phi_0\right] + l(t/n)\right)$$

where

$$\tilde{b}_{(t-i)n} = \frac{X_{(t-i)n} - l((t-i)/n)}{u((t-i)/n) - l((t-i)/n)}$$

and so,

$$\mathbb{E}\left[X_{tn}|\mathcal{F}_{t-p,t-1}\right] = (u(t/n) - l(t/n))\left[\phi_0\frac{\alpha}{\alpha+\beta} + \sum_{i=1}^{p}\phi_i\tilde{b}_{(t-i)n}\right] + l(t/n)$$

The problem of finding X_{tn} conditioned on $\mathcal{F}_{t-j,t-1}$ where $1 \leq j < p$ (and therefore $\mathcal{F}_{t-j,t-1} \subset \mathcal{F}_{t-p,t-1}$) is a bit more complicated. We will consider the extreme case of $j = 1$ as an example. Suppose that according to $\mathcal{F}_{t-1,t-1}$, $X_{(t-1)n} = x_{t-1}$. Then we have

$$(u((t-1)/n) - l((t-1)/n))\left[\sum_{i=1}^{p}\phi_i\tilde{b}_{(t-1-i)n} + \phi_0 b_{(t-1)n}\right] + l((t-1)/n) = x_{t-1}$$

Furthermore, suppose the process is causal, giving the representations,

$$X_{(t-1)n} = \left(1 - \sum_{i=1}^{p}\phi_i B^i\right)^{-1}\phi_0 b_{(t-1)n}$$

and

$$X_{tn} = \left(1 - \sum_{i=1}^{p}\phi_i B^i\right)^{-1}\phi_0 b_{tn}$$

where B is the backshift operator. Here, the causality condition is satisfied as long as $\phi(z)^{-1}$ has a power series expansion, i.e. the roots of $\phi(z)$ lie outside of the unit circle. Then, X_{tn}, is distributed as

$$\begin{aligned}X_{tn}|\mathcal{F}_{t-1,t-1} &\overset{d}{\sim} X_{tn} - X_{(t-1)n} + x_{t-1} \\ &= \left(1 - \sum_{i=1}^{p}\phi_i B^i\right)^{-1}\left(\phi_0 b_{tn} + \phi_0 b_{(t-1)n}\right) + x_{t-1} \\ &= \sum_{i=0}^{t} a_i b_{(t-i)n} + x_{t-1}\end{aligned}$$

Thus, X_{tn} conditioned on $\mathcal{F}_{t-1,t-1}$ is distributed as the linear combination of t independent beta-distributed random variables. While this distribution is not analytically tractable, a saddle-point approximation is given

by Nadarajah, Jiang, and Chu (2014) [2]. Now, lifting the conditioning on $\mathcal{F}_{t-1,t-1}$ and supposing causality once again, we can write X_{tn} as a linear process,

$$X_{tn} = (u(t/n) - l(t/n)) \sum_{i=0}^{t} a_i b_{(t-i)n} + l(t/n)$$

And from this, we can write the variance of X_{tn},

$$\mathrm{Var}(X_{tn}) = \frac{(u(t/n) - l(t/n))^2 \alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \sum_{i=0}^{t} a_i^2$$

Parameter Estimation

Now, for purposes of performing a realistic simulation, a method of estimating the parameters of the model introduced above based on the data. As we have seen how the joint distribution of the data is not analytically tractable, likelihood maximization over the entire parameter space would require further research. Instead, we can split the estimation problem into two steps: first, estimating the bounds, and second, estimating the order, shape, and weightings. Below I present a simplified version of this fitting procedure in which no further iterations are performed after both estimation steps. However, it seems that there is potential here for the modification of this procedure into an iterative optimization problem.

Step 1: Bound Estimation

As stated above, we will first estimate the bounds. Looking at the data, one would first decide on a functional form of the bounds. We have chosen a triangle wave, however, one might opt for a superposition of sinusoids for purposes of smoothness. Then, one would choose parameters, θ_u and θ_l, such that $u(i/n;\theta_u) \geq X_{in}$ and $l(i/n;\theta_l) \leq X_{in}$ for all $i \in \mathbb{N}$ and $u(i/n;\theta_u) - l(i/n;\theta_l)$ is minimized. In effect, we are finding the bounds which "hug" the data most closely. While not optimal in all cases, for noisy data, this should provide a close enough approximation.

Step 2: Unseasonalized AR(p) Estimation

Having estimated the bounds, the data can then be standardized to lie on the interval $[0, 1]$ by the transformation,

$$T : X_{in} \to \frac{X_{in} - \hat{l}(i/n)}{\hat{u}(i/n) - \hat{l}(i/n)}$$

Label $\tilde{X}_{in} = T(X_{in})$. What remains after this transformation is an AR(p) process with mean $\alpha/(\alpha+\beta)$. This gives us one equation from which to estimate α and β. If we impose a functional form on the weight functions, such as the triangle kernel as introduced before, we can remove the dependence of $\tilde{X}_{tn}$ on $\{\tilde{X}_{(t-i)p}\}_{i=1}^{p}$ for any t and use the variance of b_{tn} as our second equation with which to estimate α and β. Supposing we have done this, we have a second equation, this time involving the variance, which can be used to estimate α and β. As discussed by Owen (2008) [3], we have estimates of α and β based on the method of moments,

$$\hat{\alpha} = \hat{\mu}\left(\frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}^2} - 1\right)$$

and

$$\hat{\beta} = (1-\hat{\mu})\left(\frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}^2} - 1\right)$$

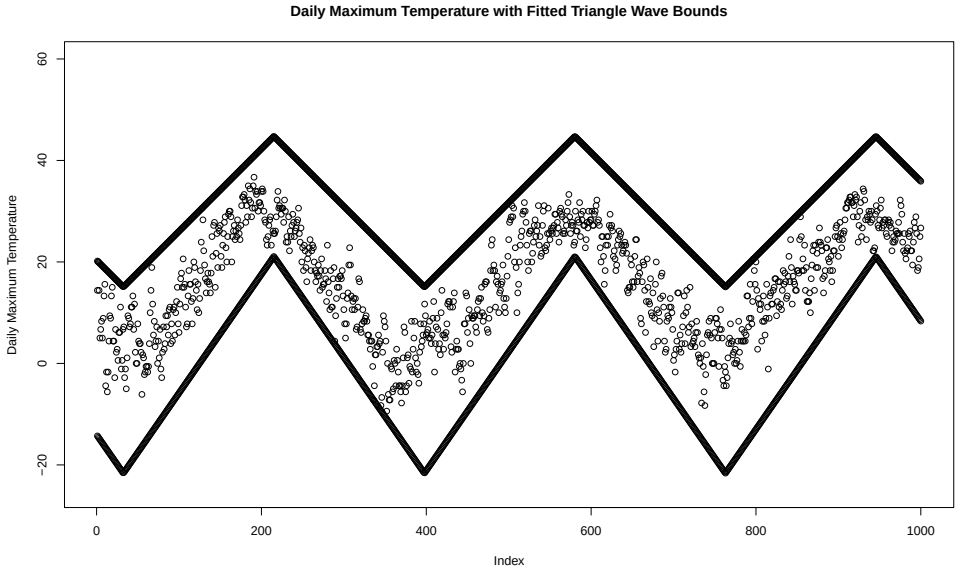
Obviously, restricting the weighting coefficients to have some specific functional form is in most cases not a valid assumption. In its general form, we have linear model with coefficients that can be estimated using least squares, written as

$$\tilde{X}_{tn} = \sum_{i=1}^{p} \tilde{\phi}_i \tilde{X}_{(t-i)n} + \phi_0 b_{tn}$$

However, further study into regression models with beta-distributed error terms would be necessary before such a method could be implemented.

Simulations

Now, we present results on the simulation of daily maximum temperature generated from the model that has been developed. Parameters for the model have been estimated based on a crude implementation of the procedure outlined above, which can be found in the *R* file on GitHub. First, we estimate the bounds based on a random subset of the data. Below, the resulting bounds are plotted with the first 1000 observations of the daily maximum,



The fitted parameters are given by

$$a = 365.25/2$$

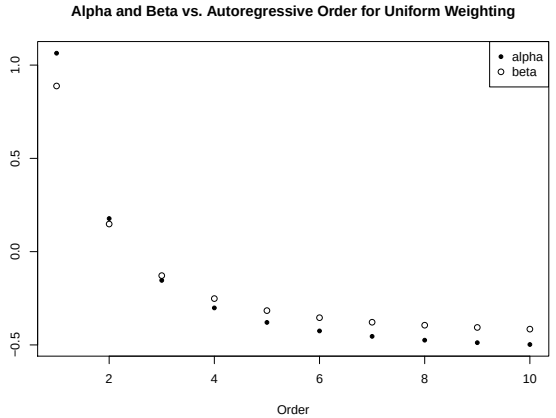
$$u_0 = 29.91$$

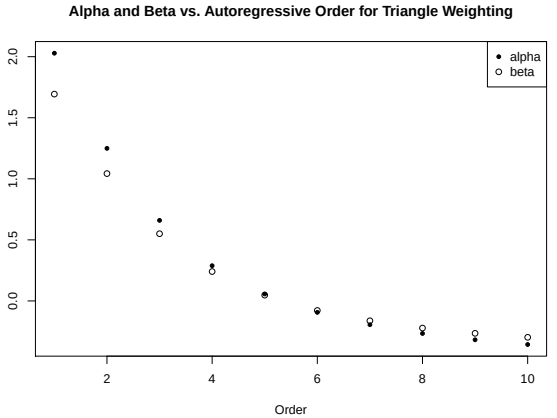
$$u_1 = 14.84$$

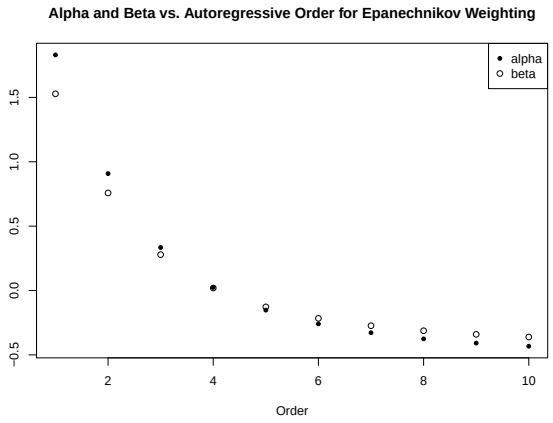
$$l_0 = -0.28$$

$$l_1 = 21.36$$

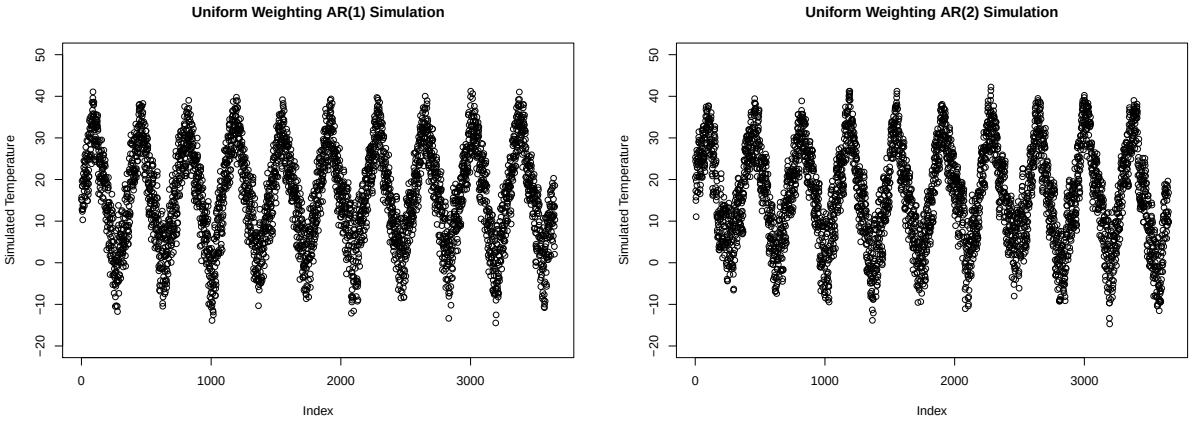
Next, without a robust method of estimating the weighting coefficients, $\{\phi_i\}_{i=0}^{p}$, we will impose a different functional forms on the weighting. Specifically, we will use the uniform, triangle, and Epanechnikov weighting functions as discussed previously. Then with the weight coefficients, we will subtract out the dependence of $\tilde{X}_{tn}$ on $\{\tilde{X}_{(t-i)n}\}$ for $i \in \{1, ..., p\}$. From this we use the method of moments to find estimates of α and β. These estimates are plotted over a range of autoregressive orders for each weighting function below.

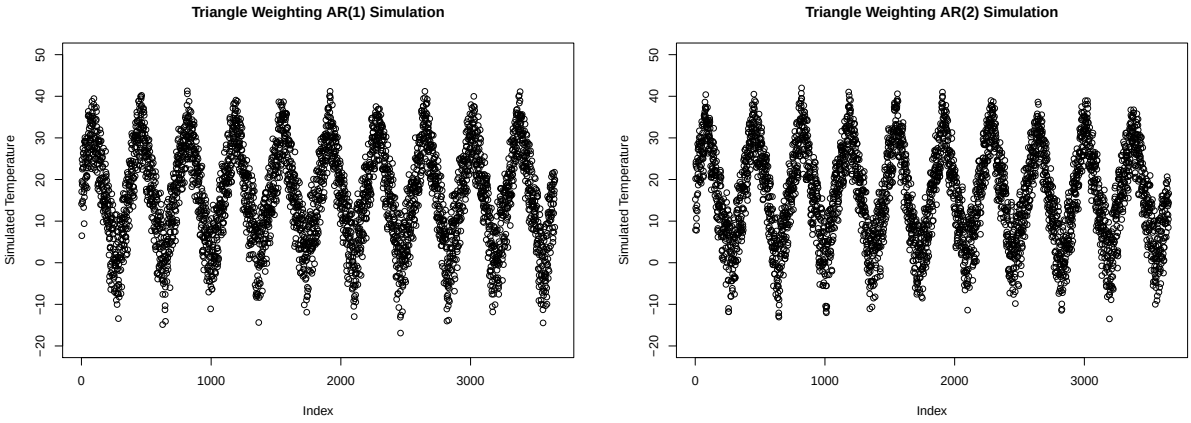


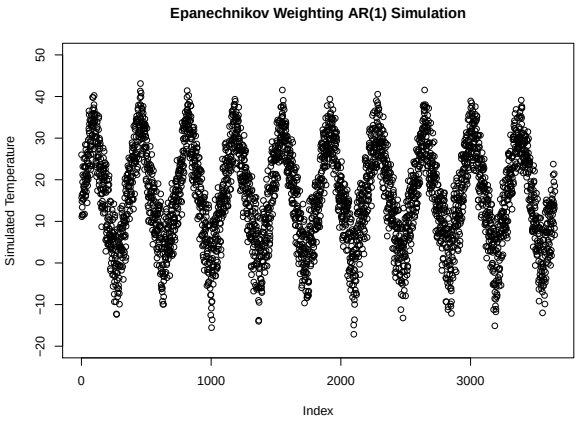


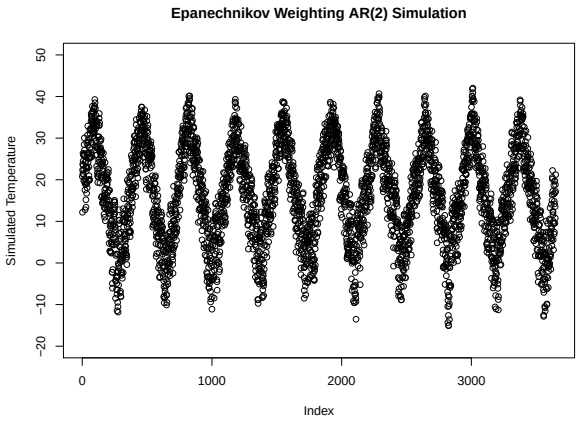


We see that for large values of p, we tend to get impossible (negative) values for α and β via the method of moments. This comes about likely from an error in the removal of dependence of $\tilde{X}_{tn}$ on $\{\tilde{X}_{(t-i)n}\}$ for $i \in \{1, ..., p\}$. For large values of p, it is likely that all of these weighting functions over estimate the dependence of daily maximum temperature at high lags. So, we will ignore orders where α or $\beta < 0$ in our simulation moving forward. Below, we give example simulations of 10 years of AR(1) and AR(2) processes for each of the three weighting functions with the estimated parameters.

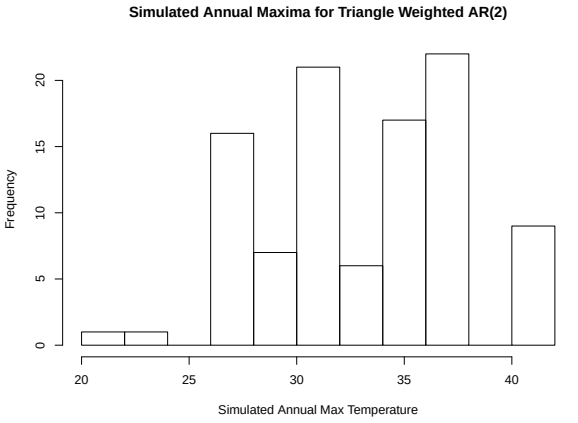


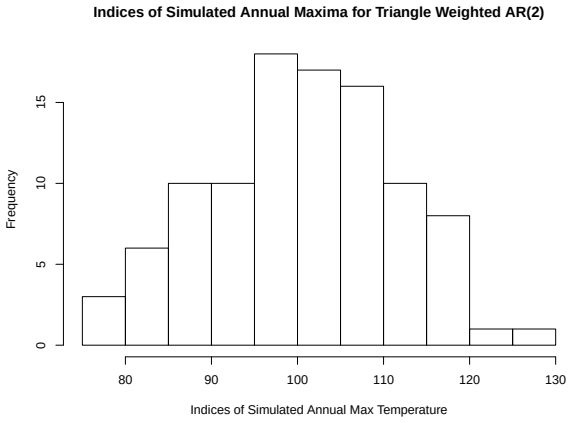


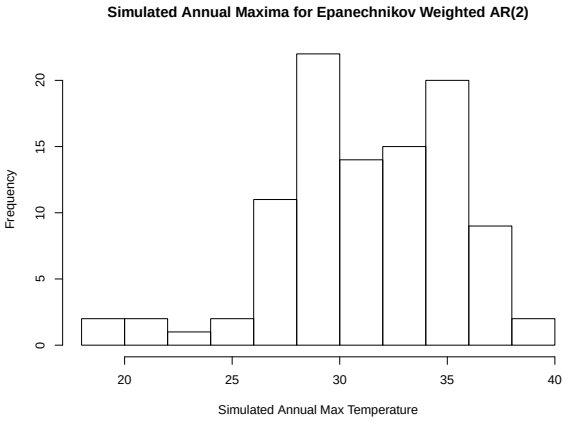


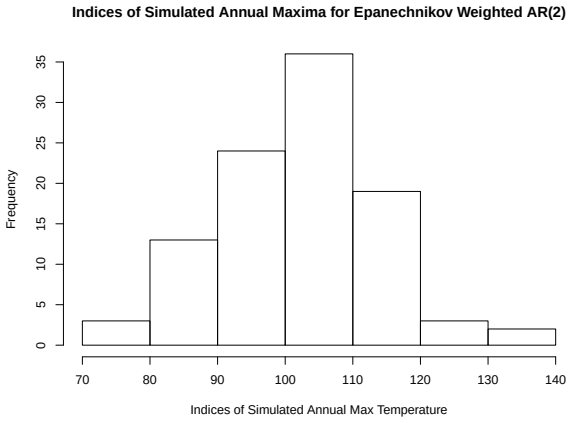


Now, we may finally ask how the annual maximum, generated from these distributions, is distributed. The simulations carried out above were extended to cover a 1000 year time period and the annual maximum was found for each year. Histograms of these values for the AR(2) processes with triangle and Epanechnikov weightings as well as the indices of the day on which the maximum occurred are given below.

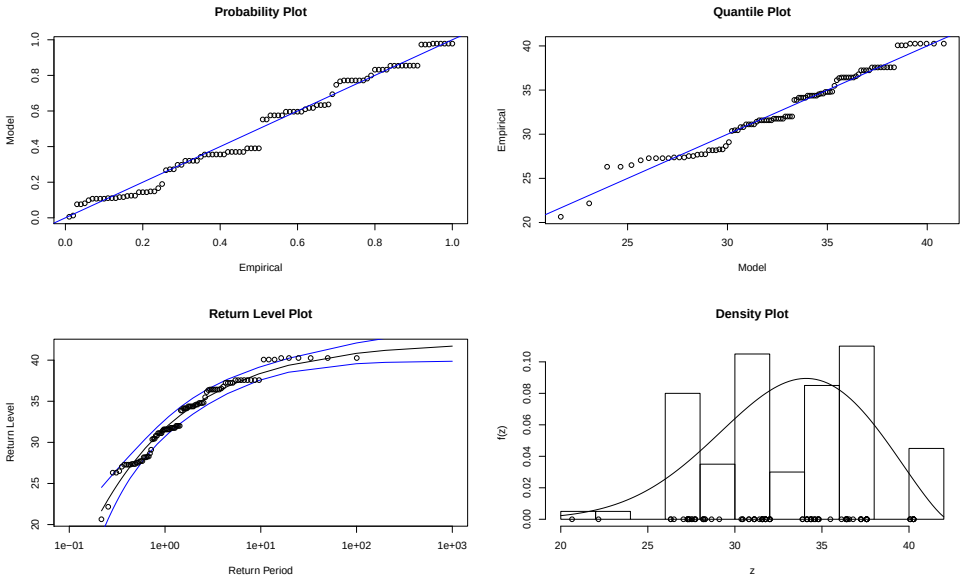




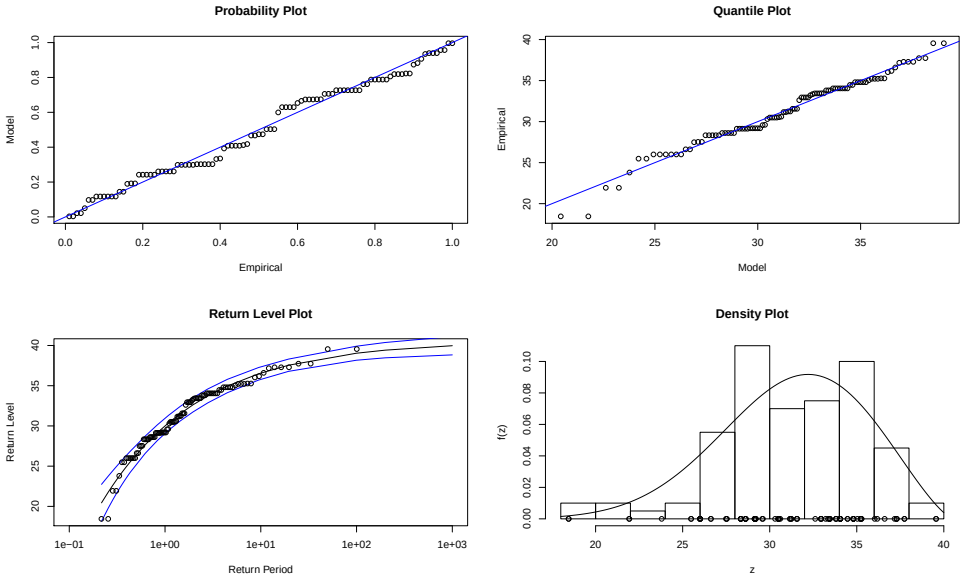




We notice that, as expected, the annual maximum for this data only occurs during a certain subset of the year. This is in agreement with one of the conditions specified by Stein on the annual extreme temperatures. Now, we will fit these two data sets to a GEV distribution using the 'ismev' package in *R*. The fit parameters for the triangle weighting coefficients are given by location: 31.7 (0.5), scale: 4.6 (0.4), and shape: -0.44 (0.08) with a negative log-likelihood of 286.3556. A diagnostic plot is given below.



Similarly, for the Epanechnikov weighting, the fit parameters were location: 30.0 (0.5), scale: 4.5 (0.3), and shape: -0.42 (0.5) with negative log-likelihood 282.3147. A diagnostic plot for this fit is provided below as well.



Conclusion

Looking at the diagnostic plots, there is some evidence that the annual maximum under this model does not follow a GEV distribution. In the quantile plots for both data sets, we see substantial deviation from linearity, while in the histograms overlaid with the density function, we see substantial disagreement, especially in the triangle weighting case. This evidence seems to support the claim of Stein, that by bounding the value of a random variable, the claim that its maxima approach a GEV distribution could be invalid. However, further study into better better ways of parametrizing and fitting this model, as well as alternative models altogether should be investigated to gain further insight into this phenomenon. This includes experimentation with smooth boundary functions, as the choice of triangle bounds in this simulation violated the smoothness condition imposed by Stein in his work.

Appendix

Beta Distribution Preliminaries

A random variable, X, is said to be Beta(α,β) distributed if it has the pdf,

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

This r.v. is supported on the interval $[0,1]$. In modeling temperature with a Beta distribution, we thus have to shift and rescale X. Specifically, if we wish to model temperature as having some finite lower bound, l, and some finite upper bound, u, we use the transformation,

$$T : X \to l + (u-l)X$$

Then $T(X)$ is supported on $[l,u]$ and temperature is bounded by $[l,u]$ in this model. Solving for the pdf of our transformed random variable,

$$\begin{aligned} f_{T(X)}(y) &= f_X(x)/T'(x)\Big|_{x:y=T(x)} \\ &= \tfrac{1}{u-l} f_X\left(\tfrac{y-l}{u-l}\right) \\ &= \tfrac{1}{(u-l)B(\alpha,\beta)} \left(\tfrac{y-l}{u-l}\right)^{\alpha-1} \left(1-\tfrac{y-l}{u-l}\right)^{\beta-1} \end{aligned}$$

The cdf of X is given by the regularized incomplete Beta function, $I_x(\alpha,\beta)$. The cdf for $T(X)$ can be found as follows,

$$\begin{aligned} F_{T(X)}(T(x)) &= \int f_{T(X)}(T(x))dT \\ &= \tfrac{1}{(u-l)B(\alpha,\beta)} \int \left(\frac{y-l}{u-l}\right)^{\alpha-1} \left(1-\frac{y-l}{u-l}\right)^{\beta-1} dy \\ &= \tfrac{1}{(u-l)B(\alpha,\beta)} \int x^{\alpha-1}(1-x)^{\beta-1}(u-l)dx \\ &= \tfrac{B(x:\alpha,\beta)}{B(\alpha,\beta)} \\ &= I_{\frac{y-l}{u-l}}(\alpha,\beta) \\ &= F_X\left(\tfrac{y-l}{u-l}\right) \end{aligned}$$

where $y = T(x)$. Next consider the maximum of n independent Beta distributed random variables, X_1, ..., X_n, labeled M_n. Then the cdf of the maximum is given by

$$\begin{aligned} F_{M_n}(m) &= F(X_1 \le m, ..., X_n \le m) \\ &= F(X_1 \le m)...F(X_n \le m) \\ &= F_{X_1}\left(\tfrac{m-l}{u-l}\right)...F_{X_n}\left(\tfrac{m-l}{u-l}\right) \\ &= \left[F_X\left(\tfrac{m-l}{u-l}\right)\right]^n \\ &= \left[I_{\frac{m-l}{u-l}}(\alpha,\beta)\right]^n \end{aligned}$$

Now, suppose we we lift our restriction that the finite bounds on daily temperature, l and u, be constant. Thus, define

$$l(i/n) : (-\infty,\infty) \to \mathbb{R}, \quad u(i/n) : (-\infty,\infty) \to \mathbb{R}$$

where $l(i/n)$ and $u(i/n)$ are the lower and upper bounds respectively on temperature in the distribution of X_i, where i varies from $[1,n]$. Thus, for some X_i, the pdf is given by

$$f_{X_i}(x) = \frac{1}{\big(u(i/n)-l(i/n)\big)B(\alpha,\beta)} \left(\frac{x-l(i/n)}{u(i/n)-l(i/n)}\right)^{\alpha-1} \left(1-\frac{x-l(i/n)}{u(i/n)-l(i/n)}\right)^{\beta-1}$$

with cdf,

$$F_{X_i}(x) = I_{\frac{x-l(i/n)}{u(i/n)-l(i/n)}}(\alpha, \beta)$$

It is clear that, from the same reasoning as in the case of constant bounds, the cdf of the maximum will be given by

$$F_{M_n}(m) = \prod_{i=1}^{n} I_{\frac{m-l(i/n)}{u(i/n)-l(i/n)}}(\alpha, \beta)$$

Thus, for a random variable, X_{in}, distributed as a shifted and scaled beta with shape parameters α and β, lower bound $l(i/n)$ and $u(i/n)$, we write $X_{in} \sim \text{SSBeta}(\alpha, \beta, l(i/n), u(i/n))$.

Transformation of Kernel Functions

As discussed previously, we wish to modify the triangle and Epanechnikov kernel functions to use as weight functions for the autoregressive coefficients in our model. To do this we must do three things: "fold" and extend the distribution to be non-zero on the interval $[0, p+1]$, find a constant such that this modified distribution integrates to 1, and discretize the function. Below, we do this for both kernel functions.

Triangle Kernel

The triangle kernel has the continuous form,

$$K(u) = 1 - |u|$$

So, we can modify it to be non-zero on the interval $[0, p+1]$ by writing

$$K(u) = 1 - \frac{u}{p+1}$$

Now solving for the constant c that makes this integrate to 1,

$$\begin{aligned}
\int_0^{p+1} cK(u)du &= \int_0^{p+1} c\left(1 - \frac{u}{p+1}\right) du \\
&= c\left(u - \tfrac{u^2}{2(p+1)}\right)\Big|_0^{p+1} \\
&= c\tfrac{p+1}{2}
\end{aligned}$$

Thus we find $c = 2/(p+1)$. Now discretizing,

$$\begin{aligned}
\int_i^{i+1} K(u)du &= \int_i^{i+1} \frac{1}{p+1}\left(1 - \frac{u}{p+1}\right) du \\
&= \tfrac{2}{p+1}\left(u - \tfrac{u^2}{2(p+1)}\right)\Big|_i^{i+1} \\
&= \tfrac{2p-2i+1}{(p+1)^2}
\end{aligned}$$

So, we find that the triangle weighted coefficients are given by

$$\phi_i = \frac{2p - 2i + 1}{(p+1)^2}$$

Epanechnikov Kernel

The Epanechnikov kernel has the continuous form,

$$K(u) = \frac{3}{4}(1 - u^2)$$

Modifying its support,

$$K(u) = \frac{3}{4}\left(1 - \frac{u^2}{(p+1)^2}\right)$$

Normalizing,

$$\begin{aligned}\int_0^{p+1} cK(u)du &= c\int_0^{p+1} \frac{3}{4}\left(1-\frac{u^2}{(p+1)^2}\right)du \\ &= c\tfrac{3}{4}\left(u-\tfrac{u^3}{3(p+1)^2}\right)\Big|_0^{p+1} \\ &= c\tfrac{p+1}{2}\end{aligned}$$

Thus $c = 2/(p+1)$. Now discretizing,

$$\begin{aligned}\int_i^{i+1} K(u)du &= \int_i^{i+1} \frac{1}{p+1}\frac{3}{2}\left(1-\frac{u^2}{(p+1)^2}\right)du \\ &= \tfrac{3}{2(p+1)}\left(u-\tfrac{u^3}{3(p+1)^2}\right)\Big|_i^{i+1} \\ &= \tfrac{3p^2+6p-3i^2-3i+2}{2(p+1)^3}\end{aligned}$$

So the Epanechnikov weighted coefficients are given by

$$\phi_i = \frac{3p^2+6p-3i^2-3i+2}{2(p+1)^3}$$

Links

- GitHub: https://github.com/oconnor-kevin

References

[1] Stein, Michael. (2016). *Should Annual Maximum Temperatures Follow a GEV Distribution?*.

[2] Nadarajah, Jiang, and Chu. (2014). *A saddlepoint approximation to the distribution of the sum of independent non-identically beta random variables.*

[3] Owen, Claire. (2008). *Parameter Estimation for the Beta Distribution.*