Univariate Extreme Value Analysis

Practice problems using the **extRemes** ($\geq 2.0 - 5$) package.

1 Block Maxima

1. **Pearson Type III distribution**

   (a) Simulate 100 maxima from samples of size 1000 from the gamma (aka Pearson type III) distribution (**Hint:** see `rgamma`).

   (b) Fit the GEV to these simulated data using `fevd`.

   (c) Check the plot diagnostics. Do the assumptions for fitting a GEV to these data appear to be reasonable?

   (d) What is the value of the shape parameter?

   (e) Use `ci` to find the 90% confidence interval for the shape parameter (**Hint**: you may find it helpful to see the help file for `ci.fevd`).

   (f) Does zero fall inside the interval?

   (g) Use `ci` to find the 25- and 100-year return levels along with their normal approximation 90% confidence intervals.

2. **Max stable property**: Repeat exercise 1.1 but draw 100 maxima from samples of size 1000 from the GEV distribution with a shape parameter equal to 0.25 (**Hint:** see `?revd`).

   (a) Given that the GEV is max stable, do the results make sense?

   (b) Is 0.25 within the 95% confidence interval for the shape parameter?

   (c) Use `profliker` to plot the profile likelihood of the shape parameter.

   (d) Does it appear to be reasonable to use the normal approximation for the confidence intervals for the shape parameter?

   (e) Find the profile likelihood 95% confidence interval for the shape parameter. **Hint**: you may need to change the `xrange` argument in order to get an adequate plot (use `verbose = TRUE` in order to get the plot); the profile likelihood should cross the blue horizontal line in two places on either side of the MLE with dashed vertical lines also crossing at this intersection. Does the resulting interval agree with the normal approximation one?

(f) Look at the profile likelihood plot for the 100-year return level. Does it seem reasonable to use the normal approximation for obtaining confidence intervals for the 100-year return level?

(g) Try to find a profile likelihood confidence interval for the 100-year return level. Even if unsuccessful, what would you say about the normal approximation interval based on these results?[1]

3. **Sample size and estimation strategies**: Maximum likelihood estimation is probably the most common strategy for fitting an EVD to data, but it may not always be the best.

   (a) Simulate samples of size 10, 20, 50 and 100 from the GEV distribution with location parameter 2, scale parameter 1.5 and shape parameter -0.5 (**Hint**: see `?revd`).

   (b) Fit the GEV df to each sample from 9 above. Check the diagnostic plots for each, and estimate 95% CI's for the parameters in each case. Are the fits reasonable?

   (c) Fit the GEV to the sample of size 10 again, but this time using the GMLE method instead of the default MLE method, and make the diagnostic plots for the fit (**Hint:** use the `method` argument in the call to `fevd`). How do they compare with the fit using MLE? Calculate 95% normal approximation confidence intervals for the parameter estimates. Do the "true" parameter values fall inside the intervals?

4. **Denver minimum temperature data set**

   (a) Load the Denver minimum temperature example data set, called `Denmint`.

   (b) What is the temporal frequency of the data?

   (c) What are the units of the temperature data?

   (d) Create a new data set that has only the annual maxima of the minimum temperature along with the corresponding values of the other variables that occur at the time of the annual maxima. **Hint:** use the `blockmaxxer` function, and note that the `blocks` argument must be a numeric vector.

   (e) Plot the annual maxima of the minimum temperature data against year. Are there are any obvious trends over time?

[1] Something is amiss with the current version of **extRemes** and the profile likelihood plotting, which may also affect the resulting intervals. We hope to resolve this issue sooner than later. Stay tuned.

(f) Fit the GEV distribution to the annual maxima of the Denver minimum temperature data.

(g) Plot the diagnostics. Do the assumptions for fitting the GEV to these data appear to be reasonable?

(h) What is the value of the shape parameter?

(i) Find the normal approximation 95% confidence intervals for the parameter estimates and the 100-year return level.

(j) What can be said from these results concerning annual maximum of Denver's minimum temperature data? **Note:** the fitted object will be used again later, so you may want to keep it handy.

(k) According to this model fit, what is the highest possible value (i.e., with non-zero probability of occurring) for Denver minimum temperature? Do you believe it? **Hint**: Check the help file for `fevd` to find the formula for finding the upper bound of the Weibull distribution.

(l) Use `pextRemes` to find the probabilities of exceeding the values 66, 67, ... 90 degrees centigrade. Do the results concur with your answer to 1.4 (k)?

(m) Use `rextRemes` to simulate a data set from the same fitted model as the Denver minimum temperature data. Plot the simulated data against the year along with the actual data using different colors and/or symbols.

5. The normal distribution is in the domain of attraction of the Gumbel distribution.

(a) Simulate 100 samples of size 50 from the normal distribution (e.g., using `rnorm`), find the maximum of each, and store these maxima in an object called `Z`.

(b) Fit the GEV distribution to the sample of maxima, and plot the fit diagnostics. Do the assumptions for fitting the GEV to these simulated data appear reasonable?

(c) Find the normal approximation confidence intervals for the shape parameter. Is the Gumbel hypothesis (i.e., that the shape parameter is zero) met?

2 Frequency of Extreme Events

1. Load the example data set called `Rsum` and look at its help file. What do these data represent? Make appropriate plots of the "count" variable to determine if there are any trends that should be taken into account.

2. Use `fpois` to fit the homogeneous Poisson to the count data and to test for equality of mean and variance.

3. Can the null hypothesis of equality of mean and variance be rejected?

4. According to the results, about how many hurricanes can be expected each year?

5. Analyze how the ENSO state may affect the hurricane count per year. Is its inclusion in a non-homogeneous Poisson fit significant at the 5% level? Does this result concur with your subjective assessment of the plots you made in 1?

3 POT Models

1. Load the data set called `Denversp` from the package `extRemes`.

2. See the help file for this data set to learn what it contains.

3. Make a scatter plot of precipitation against hour. What do you notice?

4. Make a scatter plot of precipitation against day and then year. Any patterns or trends?

5. Choose a threshold for fitting a PP model to these precipitation data (**Hint**: use 0.1 and 0.8 as the lower and upper limits). Does 0.395 mm appear to be a reasonable choice for a threshold?

6. Make auto-tail dependence plots for the data using a probability threshold of 0.8 (note that this 0.8 refers to a transformation of the data, and is not related to the 0.8 from above). Do the data appear to be tail dependent?

7. Estimate the extremal index using the intervals method and a threshold of 0.395 mm. What would you say about whether or not the data are independent over the threshold?

8. Fit a point process (PP) model to the precipitation data with a threshold of 0.395, and plot the diagnostics. Do the assumptions for the PP model fit to these data appear to be reasonable?

9. Find the estimated Poisson rate parameter. **Hint**: use the relation $\hat{\lambda} = \left[1 + \frac{\hat{\xi}}{\hat{\sigma}}(u - \hat{\mu})\right]^{-1/\hat{\xi}}$).

10. Create a variable, called `z`, that indicates whether or not each value of precipitation exceeds 0.395 mm and multiply it by 365.25 days per year (i.e., `z <- 365.25 * (Denversp$Prec > 0.395)`). Now, use `fpois` to estimate the rate parameter. How does the estimate compare with that from above? Can the null hypothesis of equality of mean and variance be rejected?

4 Parameter covariates

1. Continuing with the `Denmint` data set of the `extRemes` package, plot the negative of the annual maximum of the minimum temperature against year.

2. Does there appear to be any temporal trend in these data?

3. Fit a linear regression of year against negative minimum temperature (**Hint**: See the help file for `lm`). Is there a significant linear trend in these data (**Hint**: use the `summary` function on the `lm` fitted object)?

4. Fit a GEV distribution to the negative of the annual maximum of the minimum temperature data with a linear trend according to `Year` in the location parameter and plot the diagnostics. Do the model assumptions appear to be reasonable?

5. Use the likelihood-ratio test to determine if inclusion of the annual trend is significant.

6. Continuing with the Sept-Iles minimum winter temperature data example from the lecture slides, find the "best" GEV model for analyzing these data.

7. Continuing with the `Denversp` data.

   (a) Fit a model $Y(t) = \beta_0 + \beta_1\delta$ where $\delta = 0$ if `Hour` ≤ 12 and 1 otherwise, and Y represents the precipitation values. Is there any significant diurnal effect for this model?

   (b) Fit a point process model to the Denver precipitation data with no parameter covariates, and a varying threshold of 0.395 mm for hours larger than 12 and 0.2 otherwise. Be sure to check the model diagnostics.

   (c) Fit a point process model to the Denver precipitation data using the same varying threshold, and with a diurnal indicator in the location parameter as $\mu(\delta) = \mu_0 + \mu_1\delta$ with δ as above. Do the assumptions for the fit appear to be reasonable?

(d) Perform the likelihood-ratio test on these two fits, and also compare their AIC and BIC values. Which model would you choose for these data?

(e) Use `make.qcov` to set up a matrix that can be used with `ci` in order to estimate the 20-year return levels based on hour of the day (before or after noon) along with 95% confidence intervals, and assign the results to an R object. Plot the Denver precipitation data against hour of the day and use `lines` to add the estimated return levels and their confidence intervals. Do the intervals make physical sense?

(f) Fit a model with diurnal variation in both the location and scale parameters. Is the additional trend significant?

(g) Fit a model with diurnal variation in all of the parameters. Is the additional trend significant? Are the model assumptions reasonable based on the diagnostic plots?

5 More Practice

1. See the help file for **extRemes** to see, among other things, a list of the data sets included with the package.

2. Analyze the `Peak` data set. Is a block maxima or threshold excess model more appropriate here? Do there appear to be any trends in the data?

3. **Fitting an EVD at multiple sites**: Often it becomes necessary to fit an individual EVD to multiple data vectors that, e.g. occur at multiple spatial locations, and in such cases, it can be too cumbersome to store the entire fitted object returned by `fevd` for each and every location. While **extRemes** does not currently have functionality for directly performing such a task, it does contain a `distill` method function (cf. package **distillery**, which is installed/loaded with **extRemes**) to make the task easier.

   (a) Create a 100×500 matrix whose columns each represent a random draw from a GEV distribution (you may or may not decide to vary the parameter estimates for each column).

   (b) Write a function that takes a numeric vector argument `x`. The function should fit a GEV (using `fevd`) to the data `x`, and return the distilled version of this fit.

   (c) Using the `apply` function in conjunction with the function you wrote in 5.3 (b), fit the GEV to each column of the data simulated in (a).

(d) (optional) Make image plots of the parameter estimates.

(e) Perhaps there is other information that should be returned in (b) besides just the parameter estimates (and everything else returned by `distill`), for example, return level estimates. Modify the function in (b) to additionally return the 100-year return level and the AIC and BIC values (**Hint**: the `summary` method function will return the AIC and BIC values if you assign its output to an object), and repeat the exercise.

4. **Super-heavy tails**: The law of small numbers that gives credence to using the EVD's for maxima and threshold excesses is not as strong as the law of large numbers in that convergence to a non-degenerate distribution is not guaranteed. One example is a random variable whose distribution is super heavy-tailed.

   (a) Simulate a sample of size 1000 from a GP distribution with shape parameter value of 0.25, and assign to an object calle `y`.

   (b) Make a histogram of `y`.

   (c) Fit the GP to these data using a threshold of zero, and check the diagnostic plots.

   (d) Create a new variable, `x`, that is simply `exp(y)`.

   (e) Make a histogram of `x`.

   (f) Fit the GP to `x`.

   (g) What do the diagnostic plots indicate for this fit?

5. **Changing default optimization arguments**

   (a) Load the `rain` data from package **ismev**.

   (b) Use `pp.fit` from package **ismev** to fit the PP model to these data with a threshold of 30.

   (c) Use `pp.diag` from package **ismev** to plot the fit diagnostics. How does it look?

   (d) Use `fevd` from package **extRemes** to perform the same fit, and plot the diagnostics. how do they look now?

   (e) Which fit do you believe?

   (f) Re-do the fit with `pp.fit` using the arguments: `muinit = 40, siginit = 10`, and `shinit = 0.1`. How do the diagnostic plots appear now?

(g) Plot the `fevd` fitted object using the argument `type = "trace"`. Do you believe that you at least achieved a local minimum?

(h) Re-do the fit using `fevd` with the argument `initial = list( location = 50, scale = 20, shape = 0.5 )`, which will force `fevd` to use these values as the initial parameters instead of the default, which computes the L-moments, MOM, and, in the case of the PP model, GP estimates, and uses the one that gives the lowest negative log-likelihood value. Anything different?

(i) Re-do the fit again using these same initial values, but this time also use the argument `optim.args = list(method = "Nelder-Mead")`. Most optional arguments to `optim` can be passed in this way via `fevd`. The default optimization routine for **ismev** functions is to use Nelder-Mead with the MOM estimates as the initial parameter estimates. `fevd` uses the BFGS algorithm as the default along with the actual likelihood gradient values rather than finite differences. While this approach may sound better, with EV likelihoods, it does not always give good results; where finite differences may often give better ones. Changing the optimization routine to Nelder-Mead is one relatively easy way to force the optimization routine to not use the exact gradients.

6. **Bayesian estimation** can be performed with `fevd`, but it generally requires a savvy user to make it work; at least so far. The important pieces are the proposal and prior functions. To see how you can write your own proposal and/or prior functions, look at the default ones, and be sure to have appropriate arguments and output value. The default functions are called: `fevdProposalDefault` and `fevdPriorDefault`, respectively. Try fitting the GEV to data simulated from the GEV using the Bayesian method with its default values. Use the `type = "trace"` argument in the call to plot to check the MCMC chains for mixing and convergence, as well as the posterior densities. If they look ok, try plotting the fit diagnostics. Note that the GMLE method, a penalized likelihood approach, also uses a prior function via the same arguments (i.e., `priorFun` and `priorParams`), but has a different default (namely, `shapePriorBeta`).