Thesis Formulation Report

Developing clustering algorithms for conditional extremes models

Patrick O'Toole
Supervised by: Christian Rohrbeck and Jordan Richards
August 27, 2024

Abstract

Conditional extreme value models have proven useful for analysing the joint tail behaviour of random vectors. While an extensive amount of work to estimate conditional extremes models exists in multivariate and spatial applications, the prospect of clustering for models of this type has not yet been explored. This project will review existing methods in the area of conditional extremes models, and develop and research ideas on how some of these models can be embedded into a clustering framework. It will also involve the review of existing state-of-the-art clustering methods within extreme value analysis.

Taken from contract, edit to form proper abstract

Responsible Research and Innovation

Fill in!

Make linebreak the same as abstract

Add Responsible Research and Innovation statement

Contents

1 Introduction 4

2 Peaks-over-threshold method for univariate extremes 4

3 Conditional extremes model 5

4 Clustering for extremes 6

5 Motivating example 7

6 Discussion and future work 8

7 Code availability 9

Prospective ordering of sections:

1. Introduction
2. Peaks-over-threshold method for univariate extremes
3. Conditional extremes model
4. Clustering for extremes
5. Motivating example
6. Discussion and future work

Order important, hard to think of what might be best

1 Introduction

- What are extremes (modelling tails of distributions where underlying stochastic process is assumed), why they are used (anywhere where we are more interested in tails of distributions).
- Some simple, examples of uses in marginal estimation (and more complicated examples from reading course).
- Modelling of dependence between marginals required to improve return level estimation (see Winter et al. [2016]), with one such model being the conditional extremes model of Heffernan and Tawn [2004].
- Extremal models are notoriously difficult to fit, largely due to shape parameter, and this propagates into CE model parameters,
- There is a wealth of literature on clustering in extremes. However, clustering on the CE model has never been done before.
- In particular, hierarchical clustering would be useful for improved parameter estimation, with a Bayesian framework further introducing priors for our parameters which are useful in extremes where there is a lack of available data.
- Summarise what will be talked about in the rest of the report.

Need to motivate need for dependence modelling as in H&T 2004 and Winter 2016 papers

Do this section!

Want to do as future research, keep vague

2 Peaks-over-threshold method for univariate extremes

Follow other papers using GPD, like Carreau et al. [2017].

- Peaks over threshold (reference Halkema and de Haan 1974, Pickands 1975, see Carreau 2017 for example) definition for extremes (including reference to limiting extreme value distribution G from Coles 2001)
- Conditional CDF of GPD distribution
- Different tails of GPD for different signs of shape parameter
- Talk about different threshold selection methods, e.g. stability plots, including quantile regression method used in Youngman [2019].
- IID assumption in GPD as defined above, can model parameters as covariates, numerous methods to do this, such as using `evgam` as first introduced in Youngman [2019], and also Gaussian processes and INLA (with references).

- Return levels (m-observations, n-year, from Coles [2001]) can be estimated from GPD parameters (derived from conditional probability and exceedance probability, as in Cooley et al. [2007] and Coles [2001]).
- Reference difficulty in estimating GPD and particularly ξ due to lack of data and strange likelihood (following Coles [2001]).
- Also mention inadequacy of return level estimates with univariate methods when dependence is seen, as seen in Pacific Northwest in Winter et al. [2016], motivating use of dependence models.

3 Conditional extremes model

Below narrative is a bit jumbled, can order better by looking at Heffernan and Tawn [2004], Keef et al. [2013] and applied conditional extremes papers again.

Anything else to add?

- Introduction to dependence modelling for extremes, following Heffernan and Tawn [2004] for modelling $\mathbb{P}(\boldsymbol{X} \in \boldsymbol{C})$, i.e. an observation which is extreme in more than one dimension, be that for multiple variables at a single location (e.g. rain and wind speed) or across multiple locations and/or times.
- Following Tawn et al. [2018] and Heffernan and Tawn [2004], introduce other methods for modelling dependence in extremes, such as max-stable, Pareto and Gaussian processes, and copulas, and their limitations, both in the restrictiveness of the type of dependence these methods can model, and their computational feasibility.
- Something on marginal model as piecewise ECDF and GPD, which can be fitted using any method mentioned in 2.
- Starting from asymptotic motivation, define conditional extremes model as in Heffernan and Tawn [2004] and in applied paper as the non-parametric regression equation $Y_{-i} = a_{|i}(Y_{|i}) + \rfloor i(Y_{|i})Z_{|i}$
- Requires Gumbel transformation of margins to have exponential upper tail, but this gives complex form for a for negative dependence, so instead following Keef et al. [2013] Laplace margins are used which have doubly exponential tails and thus have the conceptually simple $a_{|i} = \alpha_{|i} Y_{|i}, b = Y_{|i}^{\beta_{|i}}$, which can be interpreted as the slope and spread parameters for our semi-parametric regression line (is there a reference I can use which makes this interpretation?).
- Extrapolation through MC algorithm as in Heffernan and Tawn [2004] (alternative in Keef et al. [2013].
- Talk about diagnostics through independence of residuals and tail exceedances in the limit.
- Uncertainty through bootstrap scheme,
- Return levels ... (improved over univariate, as shown in Winter et al. [2016])

Can also talk about (if time/space):

- Additional constraints from Keef et al. [2013] on possible values of α and β to ensure stochastic ordering of $Y_j \mid Y_i = y$ for large y
- Can also talk about how Tawn et al. [2018] shows CE is an improvement over other methods like max-stable/Pareto processes (how much detail? Look at what other papers have done)
- Spatial extension in Wadsworth and Tawn [2018], which has α which slowly decays as distance between site and some reference site increases, several choices of β to fit different tails, and the

representation of the residuals as a Gaussian process with some constraints to ensure it's value is 0 when the site in question is the same as the reference site.

Also have section on applications of CE model!

4 Clustering for extremes

Mainly lit review on clustering.

- Clustering on extremes done for two reasons: improved explainability and improved parameter estimation.
- First kind mainly focuses on deriving distance matrices for some metric (such as marginal GPD parameters, F-madogram, etc) and applying some classical clustering algorithm such as k-mediods. Can talk about different papers which have done this.
- Second kind uses hierarchical modelling and generally produces latent, "data-driven" group structure, methods for estimating likelihood can be split between Frequentist methods which use some flavour of EM algorithm, and Bayesian which uses MCMC, in particular Bottolo et al. [2003] and Rohrbeck and Tawn [2021], which use RJMCMC (other methods (possibly from outside extrems) include use of "stick-breaking prior" and latent Dirichlet allocation, reference).
- This kind particularly useful for extremes because of lack of data causing "naive" marginal ξ values to have large variance (which in turn feeds into estimates of α and β for CE model).
- Earlier paper in this vein is Cooley et al. [2007], but this only used domain-knowledge to fit two seperate ξ values for different regions (mountainous and plains).
- Talk about Frequentist EM papers for GEV (Dupuis et al. [2023]) and GPD (Carreau et al. [2017])
- Talk about Rohrbeck and Tawn [2021], how it changes RJMCMC algorithm from Bottolo et al. [2003] and uses GPD rather than PP (review both papers), how it might be an improvement on Frequentist EM methods for various reasons:
  1. Use of priors desirable where data is naturally lacking for extremal context (but may also be shunned for being opinionated?), mention use of Penalised complexity prior for ξ in INLA implementation of GPD.
  2. Inference for the Carreau et al. [2017] is quite conceptually difficult, making use of U-statistics for probability weighted moment estimators and in the GEV paper required a consistency analysis of the QML methods used
  3. In contrast, inference for Bayesian problem could be seen as somewhat simpler, can be simply expressed through DAG with hyperpriors to enable partial pooling, can more easily limit e.g. ξ to more reasonable values, spatial interpolation is easy (but can be made more complex), number of regions/clusters to use can be estimated with within MCMC scheme rather than e.g. cross-validation as in Carreau et al. [2017].
  4. Any improvements in computational speed? Can talk about INLA implementation of CE and how it allows for modelling of many regions, opens up use in geostatistical context,
  5. Probabilistically defines uncertainty around parameter estimates, which is nice, compared to having to define complicated bootstrapping schemes or other methods to quantify uncertainty in Frequentist setting.

6. Also reference context of conditional extremes model, rather than GPD or GEV explicitely (although GPD must be estimated to then estimate CE parameters), probably more easily formulated in hierarchical Bayesian model (how?)

- Therefore, seems like a Bayesian clustering scheme for Conditional Extremes, something which has never been done before, would be desirable. Would likely largely follow Rohrbeck and Tawn [2021], but wouldn't need two likelihood components as marginal estimates required to then estimate alpha parameters (or would it? Scheme could be more complicated then as you couldn't partition likelihood into decoupled components)

Should/Do I need to go into much detail about what this would look like?

5 Motivating example

1. **Introduction**: Why we look at motivating example (to show how CE model works, how parameter estimates have large variance, and a simple, likely incorrect clustering procedure can be performed on the outputted parameters/regression lines) Fill in more
2. **Data**: Weekly Winter precipitation sum and daily wind speed maxima for Ireland from 1990 to 2020 inclusive, as in Vignotto et al. [2021]. Include single exploratory plot to showcase data (left plot could have locations of weather sites, right could have rainfall plotted against wind speed for a given site).
    - Precipitation data from Met Eireann, wind speed data from ERA5 reanalysis, reference accordingly (also map from
3. **Model**:
    (a) **Marginal model**: uses `evgam` to estimate σ, ξ at each location smoothing both over space (can take from ITT2 report) for GPDs fitted to precipitation and wind speed, respectively.
        - Also uses method from Youngman [2019] whereby location-specific thresholds are defined as fixed quantiles and estimated by quantile regression.
    (b) **Dependence model** uses `texmex` to fit conditional extremes model on top of marginal models to estimate α, β for rain | wind speed and vice versa for each location. Vanilla CE used for simplicity of implementation for this motivating example.
    (c) **Clustering**: (likely incorrect/significantly flawed or overly simplistic, which is fine as it motivates further work) simple clustering algorithms:
        - Following similar method to Vignotto et al. [2021], at each location have Laplace distributed fitted regression lines $Y_{-i} = \alpha_{|i}(Y_{|i}) + Y_{|i}^{\beta_{|i}} Z_{|i}$, take excesses over high quantile (same as in dependence modelling) (no need for their risk function), partition into subsets of points which are extreme for one of or both rainfall and wind speeds, calculate KL divergence between locations as in paper to get distance matrix, perform k-mediods on this (gives centroid to each cluster which corresponds to an actual data point). Choose optimal number of clusters using silhouette method.
    (d) **Refitting**: Refit model using data from all cluster members centred at cluster centroid, see if this reduces variance in estimates for ξ (could leave σ estimated for each individual site) Still to do!
4. **Results**:
    (a) **Marginals**: Show maps of σ and ξ for both rain and wind speed, ... (what else?)
    (b) Possibly show uncertainty in ξ estimates

(c) **Dependence**:

i. Show diagnostic and quantile plots for CE for some location(s),

ii. Show bootstrapped α and β values for rain | wind speed and wind speed | rain for different thresholds, motivating choice of CE threshold at 70th quantile, and fixing β at 0.1 so that all variability is in α (also makes interpretation easier, and need for this further highlights need for better parameter estimation through grouping and/or some hierarchical model).

iii. Bootstrapped values for ξ (under vanilla `texmex` marginal estimates, rather than `evgam`) and α show that uncertainty high in both, even when fixing β.

iv. Maps of α values conditioning on rain and windspeed, possibly cross-hatch where bootstrapped α values have 95% CI which intersects 0.

v. Plot of α values versus longitude and latitude (possibly coloured by distance to coast), showing how space is main driver in difference (unsurprising as used as only covariate in marginal `evgam` model).

(d) **Clustering**

i. Plot Laplace regression lines with quantiles and separation of bivariate space into regions where one or both variables are extreme, as in Vignotto et al. [2021].

ii. Show map of cluster membership, possibly under multiple K values.

(e) **Refitting**

i. New bootstrapped values for ξ and α, hopefully see reduction in variance, but likely not much as our clustering hasn't been very principled and is merely used as a motivating example.

ii. New maps of α values, with points coloured by cluster membership and cluster centroids denoted with different shape.

5. **Discussion**:

- Marginal and dependence parameter estimates show how ξ and *sigma* vary over space for Ireland, with the north-west having the most extreme weather conditions.
- Uncertainty in ξ values shown in bootstrapping motivates need for clustering to reduce variance in estimates.
- Clustering not very principled and done more for explainability purposes than improving parameter estimation, but can comment on how it does at this task, and if this is likely done poorly than this motivates further work.

6 Discussion and future work

Summary of report:

1. Introduced extremes, why they are important, why dependence modelling is important,
2. Showed how univariate extremes can be modelled using GPD,
3. Introduced dependence modelling and the conditional extremes model, why it's an improvement over previous models, where it may be susceptible to uncertainty in ξ estimates and some previous applications of the model.

4. Reviewed clustering methods for extremes, shown examples of both types of clustering, explanatory clustering likely not applicable in this context, but hierarchical Bayesian clustering methods likely to be useful,
5. Presented a motivating example which shows how the conditional extremes model can be fit, how parameter estimates can have very large variance, how clustering could be done on these parameters, and how simple methods are inadequate, leading to the need for more involved research into this topic.

Future work:

- Try Bayesian clustering approach similar to Rohrbeck and Tawn [2021], will involve deriving likelihood for CE model and priors for parameter values.
- Later, can extend the model to spatial CE from Tawn et al. [2018].

7 Code availability

- Code for analysis available at https://github.com/potoole7/TFR.
- Fork of `texmex` package available at http://github.com/potoole7/texmex, adds functionality to fix β values and only estimate α.

anything else added?

References

Hugo C. Winter, Jonathan A. Tawn, and Simon J. Brown. Modelling the effect of the El Niño-Southern Oscillation on extreme spatial temperature events over Australia. *The Annals of Applied Statistics*, 10(4):2075 – 2101, 2016. doi: 10.1214/16-AOAS965. URL https://doi.org/10.1214/16-AOAS965.

Janet E. Heffernan and Jonathan A. Tawn. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(3): 497–546, July 2004. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2004.02050.x. URL http://dx.doi.org/10.1111/j.1467-9868.2004.02050.x.

Julie Carreau, P Naveau, and Luc Neppel. Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resources Research*, 53(5):4407–4426, 2017.

Benjamin D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for u.s. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019. doi: 10.1080/01621459.2018.1529596. URL https://doi.org/10.1080/01621459.2018.1529596.

Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer series in statistics. Springer, Guildford, England, 2001 edition, November 2001.

Daniel Cooley, Douglas Nychka, and Philippe Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007.

Caroline Keef, Ioannis Papastathopoulos, and Jonathan A. Tawn. Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the heffernan and tawn model. *Journal of Multivariate Analysis*, 115:396–404, March 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2012.10.012. URL http://dx.doi.org/10.1016/j.jmva.2012.10.012.

Jonathan Tawn, Rob Shooter, Ross Towe, and Rob Lamb. Modelling spatial extreme events with environmental applications. *Spatial Statistics*, 28:39–58, 2018. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2018.04.007. URL https://www.sciencedirect.com/science/article/pii/S2211675317302786. One world, one health.

JL Wadsworth and JA Tawn. Spatial conditional extremes. *Manuscript submitted for publication*, 2018.

Leonardo Bottolo, Guido Consonni, Petros Dellaportas, and Antonio Lijoi. Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6:25–47, 2003.

Christian Rohrbeck and Jonathan A Tawn. Bayesian spatial clustering of extremal behavior for hydrological variables. *Journal of Computational and Graphical Statistics*, 30(1):91–105, 2021.

Debbie J Dupuis, Sebastian Engelke, and Luca Trapin. Modeling panels of extremes. *The Annals of Applied Statistics*, 17(1):498–517, 2023.

Edoardo Vignotto, Sebastian Engelke, and Jakob Zscheischler. Clustering bivariate dependencies of compound precipitation and wind extremes over great britain and ireland. *Weather and Climate Extremes*, 32:100318, 2021.