A clustering framework for conditional extremes models

Patrick O'Toole, Christian Rohrbeck, Jordan Richards

March 21, 2025

Same (present), tense throughout if possible

Abstract

Conditional extreme value models have proven useful for analysing the joint tail behaviour of random vectors. Conditional extreme value models describe the distribution of components of a random vector conditional on at least one exceeding a suitably high threshold, and they can flexibly capture a variety of structures in the distribution tails. One drawback of these methods is that model estimates tend to be highly uncertain due to the natural scarcity of extreme data. This motivates the development of clustering methods for this class of models; pooling similar within-cluster data drastically reduces parameter estimation uncertainty.

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. As a motivating example, we explore tail dependence of meteorological variables across multiple spatial locations and seek to identify sites which exhibit similar multivariate tail behaviour. To this end, we introduce a clustering framework for conditional extremes models which provides a novel and principled, parametric methodology for summarising multivariate extremal dependence.

In a first step, we define a dissimilarity measure for conditional extremes models based on the Jensen-Shannon divergence and common working assumptions made when fitting these models. One key advantage of our measure is that it can be applied in arbitrary dimension and, as opposed to existing methods for clustering extremal dependence, is not restricted to a bivariate setting. Clustering is then performed by applying the k-medoids algorithm to our novel dissimilarity matrix, which collects the dissimilarity between all pairs of spatial sites.

A detailed simulation study shows our technique to be superior to the leading competitor in the bivariate case across a range of possible dependence structures and uniquely provides a tool for clustering in the multivariate extremal dependence setting. We also outline a methodology for selecting the number of clusters to use in a given application. Finally, we apply our clustering framework to meteorological data from Ireland and air pollution data in cities across the US (United States).

Contents

1 Introduction

2 Motivating examples

2.1 Irish meteorological data

-

2.2 US urban air pollution

-

3 Methods

3.1 Peaks-over-threshold

3.2 Extremal dependence

3.3 Conditional extremes

3.4 Jensen-Shannon divergence

3.5 Clustering

4 Simulation study

- Throughout this section, we evaluate the effectiveness of our clustering method using a simulation study.
- In **??**, we describe the mixture of normal and t-copulas used to generate the data for our simulated experiments.
- **??** compares our method to the leading competitor in the bivariate case, Vignotto et al. [2021].
- **??** shows how our method can be extended to three dimensions, and uniquely cluster in the multivariate extremal dependence setting.
- In **??**, we apply our method to a more realistic example, with the simulation study designed to somewhat mimic the structure of the Irish dataset introduced in **??**.
- In **??**, we evaluate the uncertainty in parameter estimates pre- and post-clustering using the bootstrapping scheme from Heffernan and Tawn [2004].
- Finally, **??** outlines a methodology for selecting the number of clusters to use in a given application, and show it's effectiveness using our simulated data.

4.1 Simulation design

Our simulation design for a single dataset, which we refer to as a "location" to match the spatial nature of our applications, is as follows:

- We generate multivariate data from a mixture of Gaussian and t-copulas, where we control the dependence structure through their respective correlation parameters.

Where to describe Vignotto method? in introduction?

Will have to mention that other paper that does bivaraite clustering as well in background, right?

Copy mostly from TFR, will have to include NI data

See Huser paper for inspiration

Or call subsections Marginal and dependence modelling? See other papers on CE

May not be subsections, but will form structure of this section

Do I want to add any visualisations of LRI here? Any ideas there?

May want tables instead of figures for some/all simulation results

How do I get the ρ in ggplot to

- Each variable in this multivariate dataset may represent, for example, a meteorological or air pollution variable at a given location.
- For simplicity, we sample from each of these copulas and transform to GPD margins with shape $\xi = -0.05$, scale $\sigma = 1$, and take the mixture of these as our variables.
- Throughout these simulations, we take an equal mixture from the margins of each copula.
- In this way, the Gaussian copula generates observations exhibiting extremal independence, whereas the t-copula induces extremal dependence, the strength of which is determined by their respective correlation parameters.
- We can use the CE model to estimate the multivariate extremal dependence structure of this data.
- One downside of this design is that while simulation is easy, it is not possible to ascertain what the "true" CE parameters are.
- We can easily extend this design by generating multiple locations with the same and different correlation parameters in the Gaussian and t-copulas.
- The knowledge of these parameters provides our "known" clustering solution.
- For example, we could generate data for four locations, where two have a Gaussian copula correlation of 0.5 and a t-copula correlation of 0.9, and the other two have a Gaussian copula correlation of 0.5 and a t-copula correlation of 0.1.
- We can use the CE model to estimate the multivariate extremal dependence structure at each location, and then apply the JS divergence to cluster these locations based on the similarity of the estimated CE parameters.
- Using this method, we can evaluate the ARI against the known clustering solution to see how well our clustering algorithm performs under various dependence scenarios.

4.2 Comparison to competing methods

- Compared to method from Vignotto et al. [2021] for two dimensions.
- Compared using Adjusted Rand Index (ARI) (), which compares a clustering solution
- Shown to be an improvement over this method for the overwhelming majority of simulations.
-

Justify use of these parameters, negative shape gives more predictable behaviour(?)

Include here? Maybe discussion point..

Plot such an example? Also is there a need?

Replace Gaussian copula correlation with symbol throughout, like in plots?

Say that we used the max risk function in Vignotto

cite!

Make interesting comments about how clustering performs better when Gaussian copula correlation is higher!

Rewrite caption

Mention what points and lines represent!

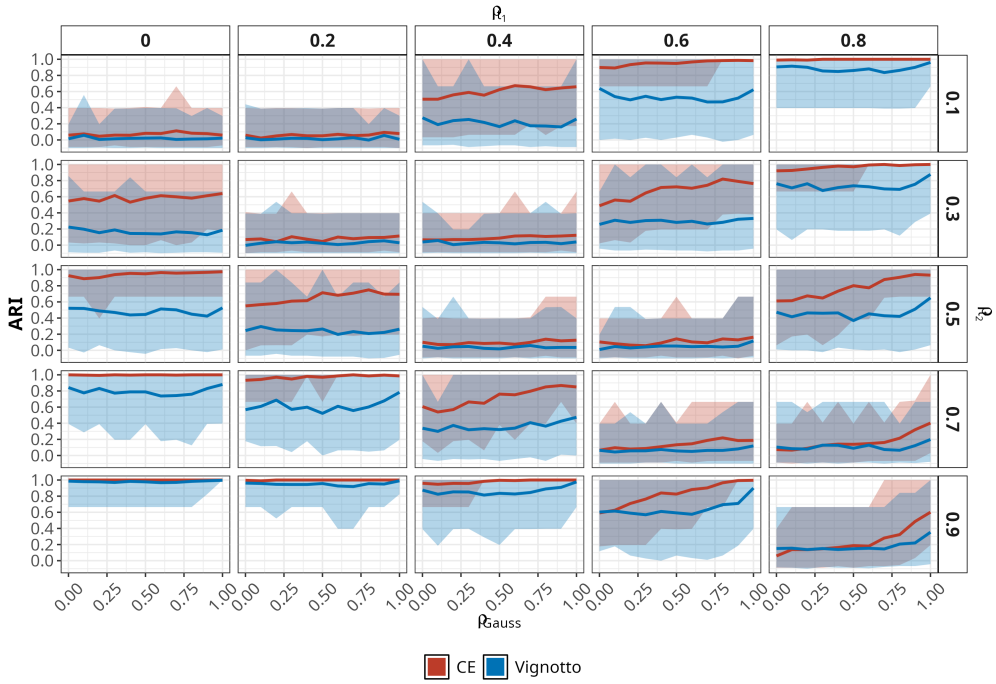

Figure 1: *Comparison of clustering methods for two dimensions and two known clusters. The x-axis represents the Gaussian correlation parameter for both clusters, while the facet labels present the t-copula correlation parameters for the two known clusters the data is simulated with. The Adjusted Rand Index (ARI) is used to compare the clustering solutions. The proposed method is shown to be superior to the leading competitor in the bivariate case, Vignotto et al. [2021]. Both models perform better when the difference in the t-copula correlation parameters between the two clusters is larger, and when Gaussian correlation is higher.*

4.3 Extension to > 2 dimensions

- Also shown to work well for three dimensions ...

write caption

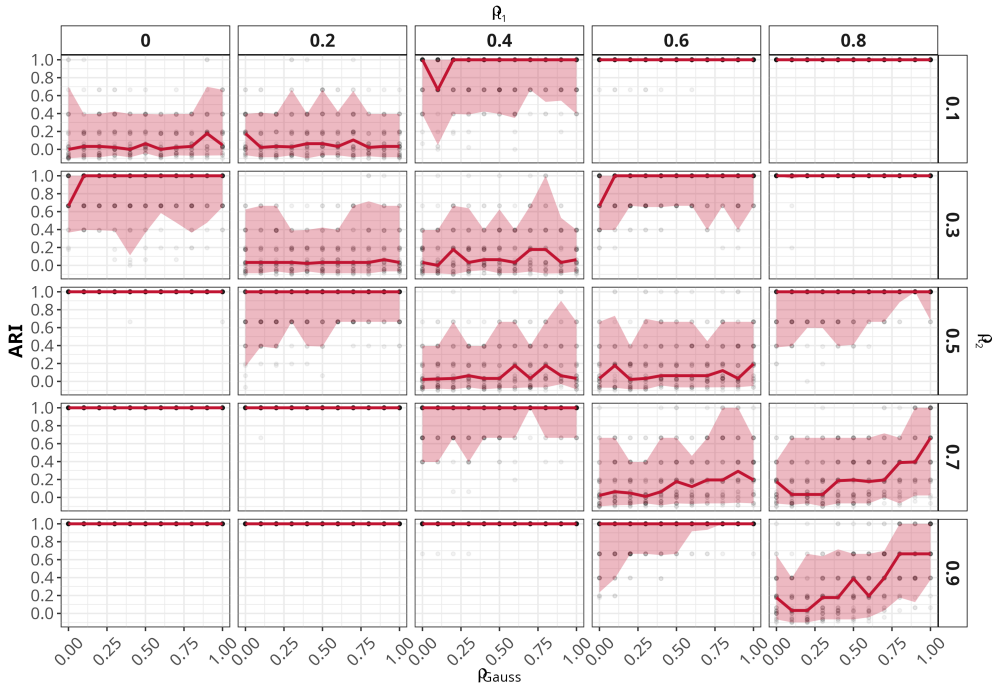

Figure 2: *caption*

4.4 More realistic example

- Generated more realistic example to somewhat match the structure of the Irish dataset.
- ? locations, ...

fill in

Write caption

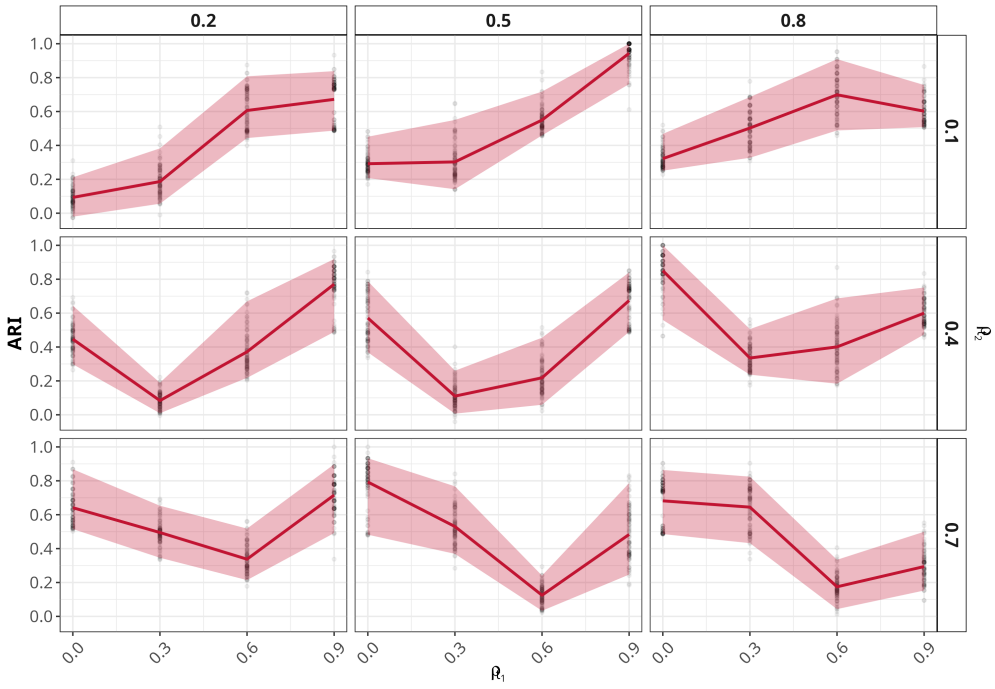

Figure 3: *caption*

4.5 Parameter estimation pre- and post-clustering

- Desire to ascertain whether dependence parameters are less uncertain after clustering.
- Can bootstrap using scheme in Heffernan and Tawn [2004] to determine uncertainty in parameter estimates.
- Post clustering, can see that uncertainty is vastly reduced for both α and β parameters in this simulation study.

4.6 Choosing the number of clusters

-
-

Describe TWGSS

Describe AIC

5 Applications

5.1 Irish meteorological data

-

5.2 US urban air pollution data

-

6 Discussion

Code availability

References

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.

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.