A clustering framework for conditional extremes models

Patrick O'Toole, Christian Rohrbeck, Jordan Richards

April 1, 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

Where to describe Vignotto method? in introduction?

2 Motivating examples

2.1 Irish meteorological data

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

-

2.2 US urban air pollution

Copy mostly from TFR, will have to include NI data

-

3 Methods

See Huser paper for inspiration

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

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 4.1, we describe a simple simulation design using a Gaussian copula to generate data. This data has the advantage of having theoretical asymptotic values for the CE model parameters, to which we can directly compare our estimates.
- In 4.2, we extend this study to a more complex simulation design using a mixture of Gaussian and t-copulas, for a variety of simple and more complex dependence structures and scenarios. We evaluate our clustering solution using the ARI, and ascertain the improvement of parameter estimation post-clustering using a bootstrapping scheme.

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

4.1 Gaussian copula

Our Gaussian copula simulation design for a single "location", nomenclature chosen to match the spatial nature of our applications, is as follows:

- We generate multivariate data from a bivariate Gaussian copula, where we control the dependence structure through its correlation parameter, ρ_N. **?** show that for a bivariate Gaussian copula, the asymptotic CE parameters are $\alpha = (\rho_{\text{Gauss}}\rho^2_{\text{Gauss}}, \beta = \frac{1}{2}$, indicating extremal independence.

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

- We sample from this copula and transform to GPD margins with shape $\xi = -0.05$, scale $\sigma = 1$, giving us our variables.

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

- Each variable in this multivariate dataset may represent, for example, a meteorological or air pollution variable at a given location.

How do I get the ρ in ggplot to

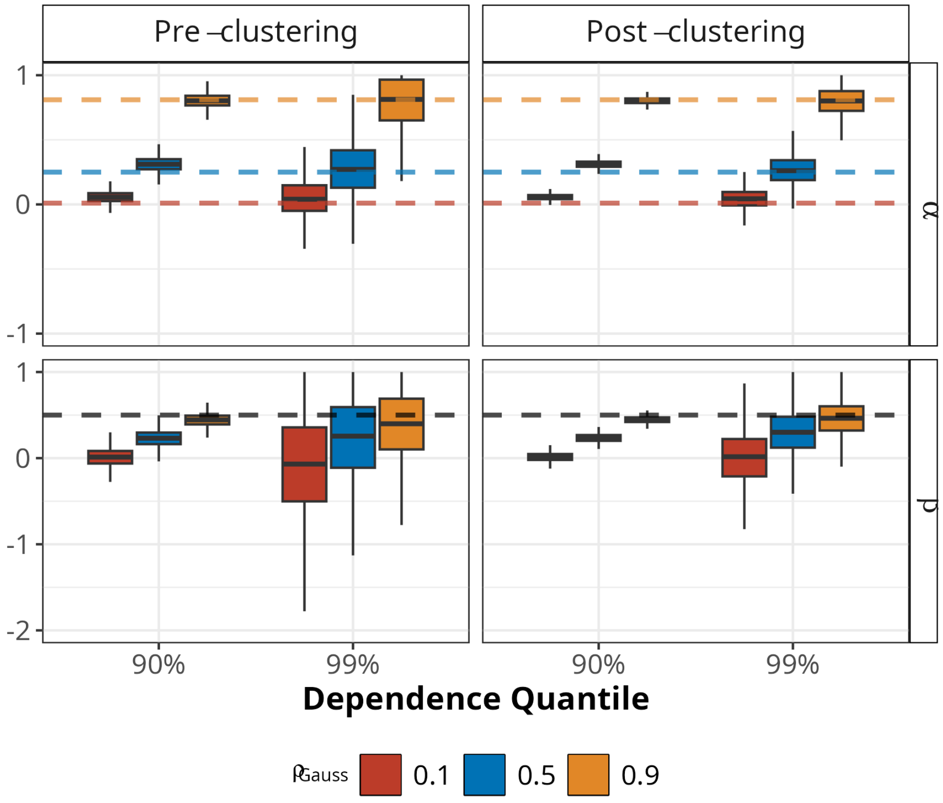
- We can then transform the Laplace margins and use the CE model to estimate the multivariate extremal dependence structure at this location.

- We can easily extend this design by generating multiple locations with the same and different ρ_{Gauss}.
- The knowledge of these parameters provides our "known" clustering solution.
- For example, we could generate data for four locations, two of which have $\rho_{\text{Gauss}} = 0.1$, and the other two have $\rho_{\text{Gauss}} = 0.9$.
- 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.
- These simulations can be repeated many times and we can evaluate the clustering solution using the ARI, and evaluate the bias and variance of our estimates around their theoretical values.
- To this end, we generated data for 12 locations each with 1000 observations of two variables using the above design.
- Three known clusters were designated by having ρ_{Gauss} values of 0.1, 0.5, and 0.9 respectively for four locations each.
- Conditional dependence quantiles of 0.9 and 0.99 were used to estimate the CE model parameters before and after clustering.
- The number of clusters k was ascertained using the method described in ?

reference part of methods that describes this

- Simulations were repeated 500 times, and the results are plotted in figure 1.
- The estimation of k and the ARI for this simple example were perfect across all simulations.
- Several interesting points can be made from this simulation study.
- Pre-clustering, the CE model estimates are more uncertain, especially for β, where the bias is also larger.
- This is especially the case for the higher dependence quantile, where data is scarcer, although this also corresponds to less bias, as we take our data from further into the tail.
- Bias and variance are lower for higher ρ_{Gauss}, which appears to be associated with better convergence to the theoretical asymptotic values.
- Post-clustering, the same patterns emerge, but the variability in the estimates is reduced, as desired.

Quote any specific values?



Stop alpha and beta facet labels from clipping

Figure 1: Boxplots of α *and* β *parameter estimates for the conditional extremes model pre- and post-clustering for a Gaussian copula simulation of 12 datasets, or "locations" belonging equally to three known clusters. Simulations were repeated 500 times. The x-axis represents the conditional quantile used to fit the conditional extremes model, and the boxes are coloured by their Gaussian copula correlation parameter,* ρ_{Gauss}*. Dotted horizontal lines indicate the theoretical asymptotic values of* α *and* β *for a bivariate Gaussian copula, coloured by the same* ρ_{Gauss} *values for* α *and black for* β *to indicate the same value across all clusters, at* $1/2$.

4.2 Mixture models

- In this section, we extend our simulation design to a mixture of Gaussian and t-copulas, where we control the dependence structure through their respective correlation parameters, ρ_{Gauss} and ρ_t, with each t-copula having 3 degrees of freedom.
- The idea behind this design is that 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.
- While there are no theoretical guarantees about our estimates, we can still evaluate our clustering solution using the ARI.
- We can also evaluate the improvement in parameter estimation post-clustering using the bootstrapping scheme described in Heffernan and Tawn [2004].
- Below, we will describe the results of this simulation study in a variety of scenarios.

4.2.1 Comparison to competing methods

Say that we used the max risk function in Vignotto

- We compared our method to the leading competitor in the bivariate case, Vignotto et al. [2021], using the maximum risk function.
- As in 4.1, for 500 simulations we generated data for 12 locations each with 1000 observations of two variables, but this time using a mixture of Gaussian and t-copulas with GPD margins.

Extend to 500 simulations, only ran 200 times

- Two known clusters were defined using different values of ρ_t, keeping ρ_{Gauss} the same for both (but varying across simulations).
- The CE model was fit at the 90th dependence quantile.
- The clustering solutions were evaluated using the ARI, and the results are plotted in figure 2.
- The proposed method is shown to be superior to the leading competitor in the bivariate case, Vignotto et al. [2021], across a range of possible dependence structures.
- Both models perform better when the difference in the t-copula correlation parameters between the two clusters is larger.
- Both models also perform better when Gaussian and t-copula correlation are higher, as the signal of extremal dependence is stronger.

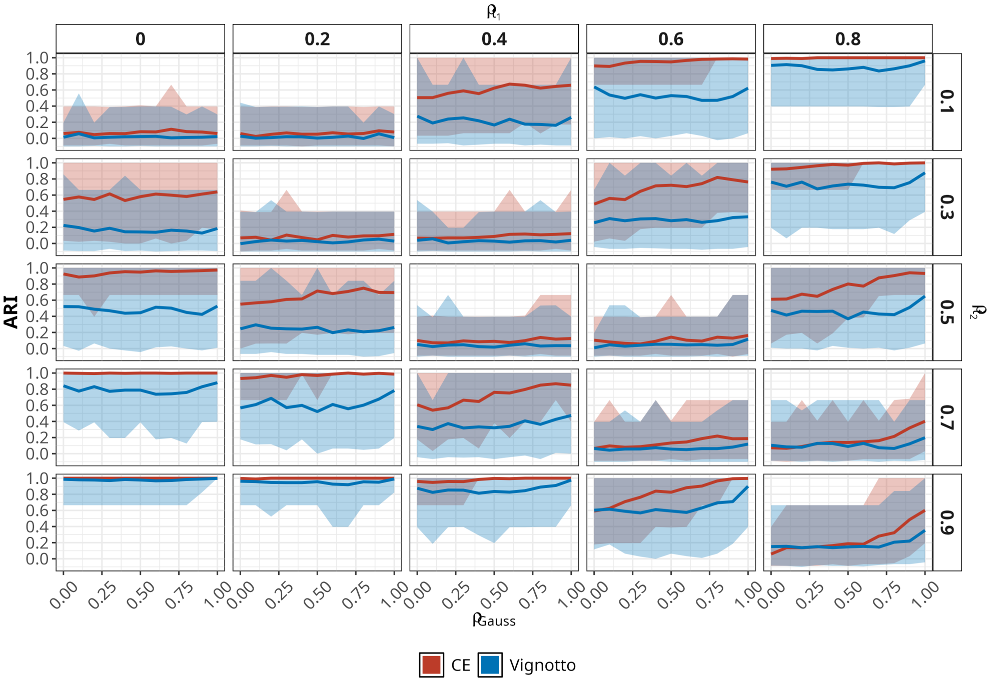


Figure 2: *Comparison of clustering methods for two variables and two equally sized clusters for simulations of 12 "locations" from a mixture of Normal and Gaussian copulas. A grid search was performed, with the x-axis representing the Gaussian correlation parameter used for both clusters, and the facet labels showing the t-copula correlation parameters for each "known" cluster. This grid search was repeated 500 times. The lines show the median of the Adjusted Rand Index for both clustering methods, with uncertainty coming from the 90% credible interval.*

4.2.2 Extension to > 2 dimensions

- While the Vignotto et al. [2021] method is restricted to two dimensions, our method can be extended to three dimensions and beyond.
- To illustrate this, we can repeat this simulation study, but with three variables.
- The results are plotted in figure 3.
- The proposed method is shown to work well for three dimensions, and the same patterns emerge as in the bivariate case.
- The method is shown to actually perform better in three dimensions than in two.
- This comes with the caveat that the variables all have the same GPD marginal distribution, and so including three variables naturally increases the amount of data available to estimate the same dependence structure as in the bivariate case.

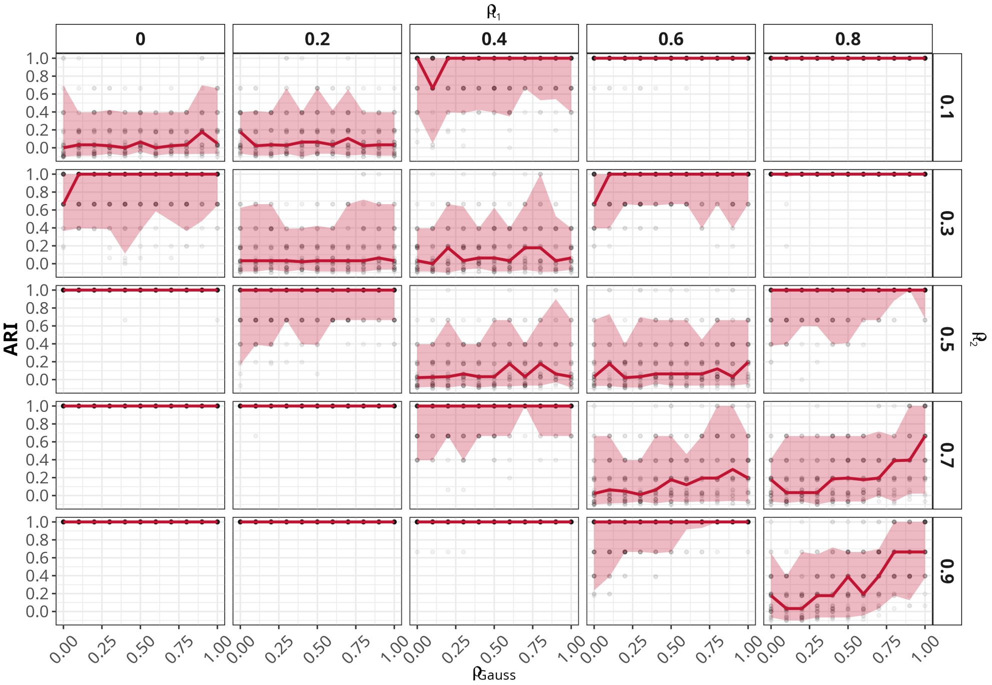


Figure 3: *Evaluation of clustering performance for three variables and two equally sized clusters for simulations of 12 "locations" from a mixture of Normal and Gaussian copulas. A grid search was performed, with the x-axis representing the Gaussian correlation parameter used for both clusters, and the facet labels showing the t-copula correlation parameters for each "known" cluster. This grid search was repeated 500 times. The lines show the median of the Adjusted Rand Index for both clustering methods, with uncertainty coming from the 90% credible interval. Points show individual values of the ARI for a given simulation.*

4.2.3 More realistic example

- We generated a more realistic example to somewhat mimic the structure of the Irish dataset introduced in 2.1.
- We generated data from 60 locations, each with 1000 observations of two variables, using a mixture of Gaussian and t-copulas with GPD margins.

- The 60 locations had three known clusters, with 30, 20, and 10 locations in each cluster respectively.
- A slight perturbation via a uniform sample between −0.05 and 0.05 was added to the correlation parameters of the t-copula for each location, to make the clustering solution more challenging.
- Again, clustering was done on the CE model parameters using the JS divergence, the results of which are shown in figure 4.
- The algorithm is again shown to perform well, particularly where the difference in the t-copula correlation parameters between the clusters is largest.

Check facet and x-axis labels

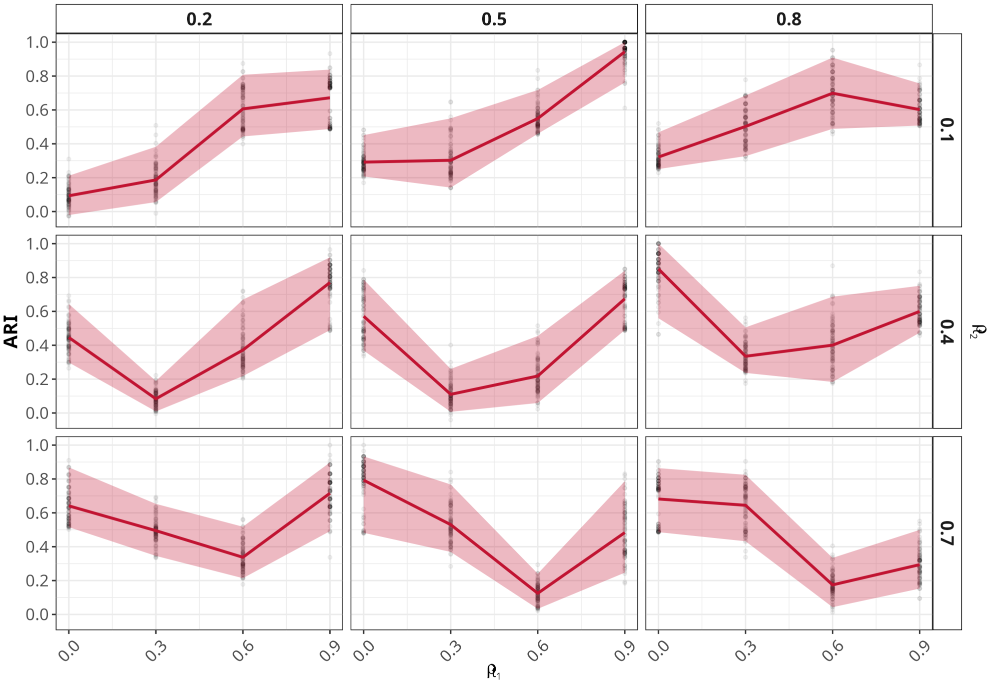


Figure 4: *Evaluation of clustering performance for two variables and three clusters for simulations from a mixture of Normal and Gaussian copulas. 60 locations were simulated with 10, 20 and 30 belonging to each respective cluster. A grid search was performed, with the x-axis and facet labels showing the t-copula correlation parameters for each of the three "known" clusters, for a Gaussian copula correlation of 0.5. Perturbations were added to these parameters. This grid search was repeated 500 times. The line shows the median of the Adjusted Rand Index for both clustering methods, with uncertainty coming from the 90% credible interval. Points show individual values of the ARI for a given simulation.*

4.2.4 Parameter estimation pre- and post-clustering

- Finally, we recall that a reason for clustering might be to reduce the uncertainty in parameter estimates by grouping similar data together.
- To this end, we desired to ascertain whether the estimates of our dependence parameters were less uncertain after clustering.
- We can bootstrap using the scheme described in Heffernan and Tawn [2004] to determine the uncertainty in our parameter estimates.

Describe scheme, how it preserves the dependence structure of the data?

- As we are estimating two parameters, α and β, for which we do not know the true values, it may be that although our uncertainty is diminished post-clustering, the estimates may still be biased, with the parameters unidentifiable.
- There are also many possible combinations of α and β that could give the same conditional estimates, and so the estimates may be biased in this sense as well.
- To avoid this, we adopt the approach of Richards and Wadsworth [2021] in calculating the conditional expectation of one variable given the other is at the 98% marginal quantile u, given by

  $$\mathbb{E}[X_1 | X_2 = u] = \hat{\alpha} u + u^{\hat{\beta}} \hat{\mu}, \tag{4.1}$$

  where $(\hat{\alpha}, \hat{\beta}, \hat{\mu})$ are the estimated CE model parameters.
- We can bootstrap this conditional expectation to quantify the reduction in uncertainty post-clustering.

Tense consistency

- We take an example for which we know the clustering algorithm performs well, namely the same simulation design as in 4.2.1, with $\rho_{\text{Gauss}} = 0.5$, and $\rho_t = 0.1$ and 0.9 for the two clusters.
- We estimate the CE model parameters pre- and post-clustering, and bootstrap the conditional expectation of one variable given the other is at the 98% marginal quantile using equation 4.1.
- The results are plotted in figure 5.
- We can see that the uncertainty in the bootstrapped estimates is vastly reduced post-clustering, as desired.
- The conditional expectation is higher for the higher t-copula correlation cluster, as expected.

Stop clipping for x-axis title

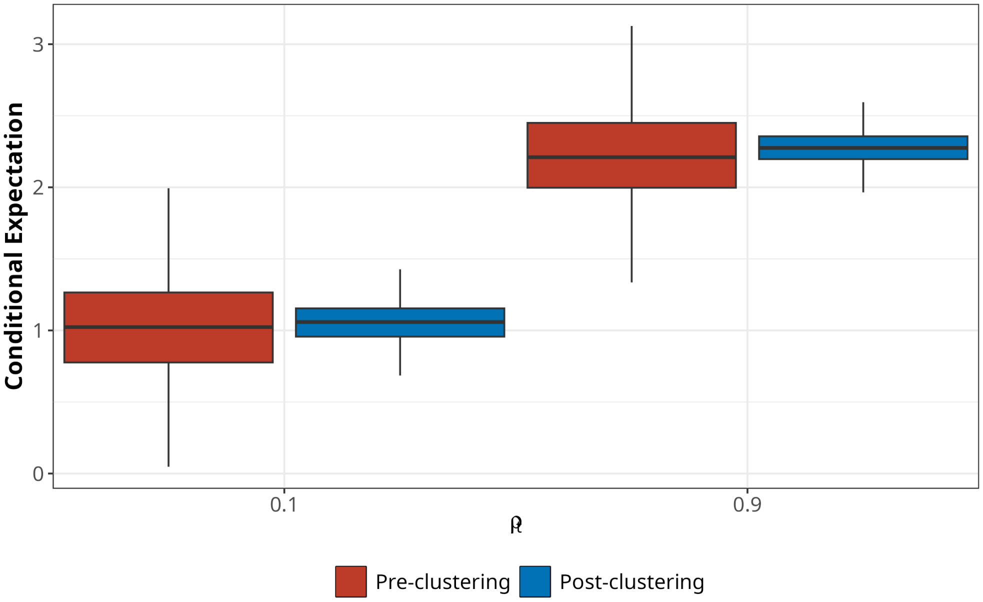


Figure 5: *Boxplots of 500 bootstrapped conditional expectations of one variable given the other is at the* 98% *marginal quantile u, pre- and post-clustering for a simulation of 12 "locations" from a mixture of Normal and Gaussian copulas, with the Gaussian correlation set to 0.5. The t-copula correlation parameters for the two clusters were set to 0.1 and 0.9, as shown on the x-axis.*

5 Applications

5.1 Irish meteorological data

-

5.2 US urban air pollution data

-

6 Discussion

- For simulations, could have looked at LRI as well as ARI.
- Could have looked at spatially varying clustering parameters, to more closely mimic spatial applications.

Code availability

References

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.

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.

Jordan Richards and Jennifer L Wadsworth. Spatial deformation for nonstationary extremal dependence. *Environmetrics*, 32(5), August 2021.