Chapter 1

Introduction

1.1 Introduction

One can face many situations with the problem of studying extreme and rare events. Sometimes, it extends to the problems coping with extreme events in a multivariate setting. e.g., risks in equity markets, extreme environmental events, etc. As a conventional approach in the study of the multivariate domain, we can split this problem into two sub-problems:
First, investigate the marginal distributions of multivariate extremes.
Second, explore the dependence between extreme variables.
Thanks to Extreme Value Theory, there are elegant tools for studying the marginal distributions; However, dependence structures have not been sufficiently noticed by researchers so far. In this thesis, we are mainly trying to learn this dependence structure precisely. To do so, we need to find statistical tools to discover the underlying dependence structure for extremes. ”Estimating this dependence in higher dimensions is particularly challenging because the number of extreme observations k_n is by definition much smaller than the number of all samples in a dataset.”[Engelke, Volgushev 2021] Moreover, ”statistical modeling in the field of extreme value theory has been limited to moderate dimensions so far, partly owing to complicated likelihood and lack of understanding probabilistic structure.” [Engelke, Hitz 2020] Therefore, it motivates us to explore statistical methods which are able to model these dependencies straightforwardly. This fact encourages us to use the vital tools including conditional independence, probabilistic graphical models, along the Extreme Value Theory. Engelke and Hitz have shown a new definition of conditional independence and graphical models for extremes. Consequently, it makes factorization and sparse dependence structure possible. The sparse structure learning of extremes done by Engelke and Volgushev enables us to recover the true underlying extremal tree consistently; however, the assumption of having trees as a connected graph is restrictive. One can obtain a more parsimonious model by allowing one to model the dependence structure using an unconnected but plain sub-class of graphs like a forest. This thesis estimates the underlying forest structure of extremes and shows that the provided algorithm consistently recovers the true underlying forest structure. Eventually, we apply the algorithm to a real dataset and assess the algorithm’s performance for a real finite dataset.
In the following, we provide a summary of the concepts, which are fundamental building blocks for our novel contribution.

Chapter 2

Literature review

2.1 Graphical Models

Graphs are great visual tools that can facilitate and standardize the discussions around multivariate dependence. The probabilistic graphical models are naturally modular, so the complex problem of studying the multivariate dependence structure can be described and handled by a careful combination of simple elements.[Lauritzaen 1996?] The critical question is, how do we link the dependence associations to a graph structure? The answer is that the graphs can keep track of conditional independence. The notion of conditional independence explains the dependence structure between random variables with a much weaker assumption than independence. This weaker assumption provides remarkable flexibility for statistical modeling. To compare these assumptions, "for random variables X, Y, and Z, we have expression $X \perp\!\!\!\perp Y$ so-called independence of X and Y, stating that reading Y is irrelevant for reading X. However, expression $X \perp\!\!\!\perp Y|Z$ so-called conditional independence of X and Y knowing Z has the interpretation of knowing Z, reading Y is irrelevant to reading X."[Lauritzaen 1996?] To go further, we should explain how graphs illustrate the structure of conditional independence? The answer key is in Markov properties on undirected? graphs.
Markov properties state: let V is a set of vertices in a graph $\mathcal{G} = (V, E)$ corresponds to a collection of random variables $(X_\alpha)_{\alpha \in V}$, then given some particular conditional independence relationships between the random variables, i.g., nodes, one can specify E set of edges in $\mathcal{G}$ which is storing the dependence structure of those random variables.

Definition 1. Markov properties: Let $V = \{X_1, X_2, ..., X_d\}$ is the set of vertices and we have a collection of random variabels $(X_\alpha)_{\alpha \in V}$ which are associated with an undirected graph $\mathcal{G} = (V, E)$, where $E \subset V \times V$.
Then:

- The pairwise Markov property, realtive to $\mathcal{G}$, if for any pair (i, j) of non-degenerate vertices:

$$X_i \perp\!\!\!\perp X_j | V \setminus \{X_i, X_j\}$$

- The global Markov property, relative to $\mathcal{G}$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $\mathcal{G}$:

$$A \perp\!\!\!\perp B | S$$

Remark. These two properties are equivalent where the joint density of all variables with respect to a product measure is positive and continuous which is our case in the following context.

One can see how these propertiese enable us to connect a collection of random variables and their conditional independence to vertices and edges of a graph in order.

Definition 2. Graphical Models: The random vector X is said to be a probabilistic graphical model on the graph $\mathcal{G} = (V, E)$, where $X = (X_i)_{i \in V}$ takes values in the Cartesian product $\mathcal{X} = \times_{i \in V} \mathcal{X}_i$ if its positive and continuous distribution satisfies either pairwise or global Markov properties.

Conditional independence and graphical models are intimately related to factorizing a multivariate joint distribution to marginal distribution [Lauritzaen 1996?]. Using Markov properties, we can enjoy the advantages of linking factorization to the graph structure. The Hammersly-Clifford theorem connects these two concepts.

Theorem 1. *A probabilistic graphical model X with postive and continuous density f with respect to a product measure μ satisfies the pairwise Markov property with respect to an undirected graph $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$:*

$$f_X(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C), \qquad x \in \mathcal{X} \tag{2.1}$$

where $\mathcal{C}$ is a set of cliques on the graph $\mathcal{G}$, and ψ_C is a suitable function on $\times_{i \in C} \mathcal{X}_i$.

Corollary 1.1. *If the graph $\mathcal{G}$ is decomposable, then this factorization can be rewritten in terms of marginal densities:*

$$f_X(x) = \frac{\prod_{C \in \mathcal{C}} f_C(x_C)}{\prod_{D \in \mathcal{D}} f_D(x_D)}, \qquad x \in \mathcal{X} \tag{2.2}$$

where $\mathcal{D}$ is a multiset containing intersections between cliques called separator set.

The modularity provided by graphical models and factorization facilitates the study of the multivariate complex models. Trees are a simple and sparse sub-class of graphs; one can obtain a more parsimonious model after factorization. So they are popular in literature. This glimpse gives us brilliant insight into applying graphical models on multivariate extremes.

2.2 Graphical Models for Extremes

The paper by Engelke and Hitz[2020] ”Graphical models for extremes” is the first principeled attempt to define the notion of conditional independence for general multivariate extreme values models that naturally extends to the graphical models, sparse structures, and factorization of densities. It estabilishes the foundation of extreme value theory on graphical models through these steps:

1. It explains the implementation of graphical models brings sparsity and parsimony to modelling of dependence structure of extremes.

2. It asserts that the problem of unreliable estimation using a few data points as extremes is an incentive to implement Extreme Value Theory in order to estimate rare events by extrapolation of tail distribution of a d-dim X. Where d is larger than 2, the result of this extrapolation is tightly depends on the strength of extremal dependence of X.
   In Extreme Value Theory, there are two assymptotically justified approaches:

   (a) **Max-stable distribution** which is the limit of normalized maxima of independent copies of X.

   (b) **multivariate Pareto distribution** which describes the random vector X conditioned on the events that at least one component exceeds a high threshold.

3. In the next step, it provides a well-defined notion of conditional independence for sparse multivariate extreme models. Doing this task is not straightforward for extremes and tail distributions. Mainly, it is not feasible to define a notion of conditional independence for a multivariate max-stable model; because of two reasons:

   (a) Papastotathopolus and Strokorb[2016] have shown that for a max-stable random vector with a joint positive and continuous density the notion of conditional independence is equivalent to independence. In other words, we cannot distinguish these two.

   (b) Moreover, Gissibl and Klüppelberg[2018] say that the conditional independence is only available for a max-stable with discrete spectral measure, but the problem with that is the fact that they do not admit densities and it excludes most of parametric families that we can use for statistical modeling.

   Hence, these drawbacks for defining conditional independence on max-stable distributions were their motivation for introducing a new notion of conditional independence on multivariate Pareto distribution, i.g., $Y_A \perp_e Y_C | Y_B$
   This notion is different from the classic one since the domain of Y is not a product space anymore. In this setting, a d-dim multivariate Pareto distribution is conditioned on the event that at least one component exceeds a high threshold, constructing a $\mathcal{L}$-shape sapce. Therefore, the absence of these exceedance on $d-1$ components implies the exceedance over the high threshold for the remaining component. So, the classical notion of conditional independence doesn’t work on a space that is not a product space. It is more clear when $d = 2$. In this case the absence of exceedance in one compenent impose the exceedance of other component. So knowing this information on one component we have information about the other component. It imposes dependece of these two variables although they might not dependent.
   In the paper it has been shown that homogeneity of Y as a multivariate Pareto distribution can be used to show that the new notion of conditional independece is well-defined.

4. We know that conditional independence and graphical models are tightly linked with eachother. Engelke and Hitz[2020] provide a modified version of Markov properties for

a multivariate Pareto distribution in a $\mathcal{L}$-shape space, and link the conditional independence of extremes to extreme graphical models. This brings simplicity and sparsity for the dependence structure of extreme variabels. In next step, they provide a Hammersly-Clifford type theorem to factorize a complex joint density to simpler modules.
The extremal graphical models whose undrlying graph is a tree have a particularly simple multiplicative stochastic representation in terms of extremal function, which is useful in simulation. To estimate the true underlying tree structure and sparsity structure, they suggest implementing a minimum spanning tree approach using censored likelihood as weights. Moreover, it has been discussed that one can find sparsity using the zero pattern of extremal variogram of a multivariate extremes from the family of Hüsler-Reissdistribution.

These four core stages provide an elegant tool to explain the conditional independence structure of multivariate extremes. We can exploit probabilistic graphical models, and factorization to simplify dependence structure of a joint multivariate extreme by parsimonious modules.

2.3 Structure Learning for Extremal Tree Models

Using the new notion of conditional independence provided by Engleke and Hitz[2020] we are able to exploit benefits of graphical models and factorization. The resulted sparsity and simplicity provide an elagant tool for us to study the dependence structure of extremes.

To use the benefits of factorization one needs to be able to estimate univariate and multivariate marginal densities. Moreover, since we usually do not have prior information about dependence structure, one should learn this conditional independence structure and its corresponding factorization.

Typically in a multivariate setting there are two major problem. First, the marginal behavior of random variables. Second, the dependence between random variables. Implementing the powerful tool of Extreme Value Theory we can properly estimate the marginal distributions. The only remaining problem is to estimate the conditional dependence structure. So we need a "data-driven" approach in order to detedt conditional independence relations and to estimate a sensible graph structure.

The simplicity and sparsity of the trees' structure is encouraging for using this sub-class of graphs for the first attempt of modeling this dependence structure. The approches on structure learning for tree are mostly based on the notion of minimum spanning tree. The idea of minimum spanning tree is based on minimization of the overall sum for some predifined distances. Given the set of distances, so called "weights", there exist greedy algorithms that construct this tree. These greedy algorithm are order-based and they do not care about the value of these weights. They are able to provide a minimum spanning tree, T_{mst}, based on these weights; however, for statistical inference it is required to choose weights in a way that gauranties T_{mst} consistently recovers the true underlying tree structure.

A common and conventional appproach is to choose the weights in a way that maximizes the likelihood for a given parametric model on dependencies. The proposed weight in Engelke and Hitz[2020] is negative maximized bivariate likelihood. Although it is based on the conventional approach of likelihood maximization, it has two disacvantages: First, it is asymptotically costly for high dimensions; Second, it needs parametric assumptions for the parametric bivariate modules.

In the paper by Engelke and Volgushev, two other types of weights, extremal correlation and extremal variogram have been introduced. The advantages of these types of weights is that they do not require any further parametric assumptions on the existance of densities and this originated from the homogeneity property of multivariate Pareto distribution and elagant stochastic representation of extremal tree models.

They have shown, using consistent estimators of extremal correlation and extremal variogram, one can non-parametrically and consistently learn the true underlying tree structure of dependencies. They also studied the asymptotic behavior of these structure learning methods and found a non-asymptotic bound for the probability of recovery of the true trees.

By simulation it has been shown that structure learning with extremal variogram based weights overperform the learning methods using extremal correlation as weights. It is realted to the fact that extremal variogram is an additive tree metric and it provides larger loss when the greedy algorithm chooses a wrong edge; however, it is not the case for the learning method using extremal correlation.

However, proposed censored likelihood by Engelke and Hitz[2020] and the extremal variogram provided almost similar and best efficiencies, the lower computational cost and absence of parametric assumptions in the later one are great motivators for us to prefere extremal variogram over censored likelihood weights. They have shown the performance of all methods decreases at the boundaries of dependence, i.e., complete dependence or full independence. They also discussed that efficiency of all approaches deteriorates when the the dimensions increases.

In this thesis and the following chapters, we try to extend the connected sub-class of graphs, tree, to an unconnected sub-class, forests, and build a theoretical foundation for the novel structure learning approach. Compare to trees, forests have sparser structures and they provide a more realistic models when we have weak dependencies between variables. Using the high sparsity of forest we can factorize a multivariate extreme joint density in a most parsimonious way. These advantages encouraging us for doing Structure Learning for Extremal Forest Models But before that we need to adjust the notion of conditional independence in order to include complete independence and so unconnected graphical models.

2.4 Extremal Independence Old and New

In the new contribution for structure learning we try to extend the previous sub-class of graphs, trees, to forests. A forest entails one or several trees (including isolated nodes or trivial nodes). A remarkable difference between forest and tree is that forest might have unconnected nodes. In the language of graphical models and conditional independence it means that we should extent the definition of $Y_A \perp_e Y_C | Y_B$ to the case that $B = \emptyset$ meaning that Y_A and Y_C are independent the novel introduced notion of extremal independece by Strokorb[2020] allows for a meaningful interpretation of disconnected graphs and independence.
WRITE AN INTUITION ABOUT THE PROOF!
It has been proved that this new approach is equivalent to the old one. Given that, we have the desired theoretical foundation for defining a forest as underlying structure of a graphical model. Therefore, we are able to introduce a new algorithm for structure learning for extremal forest model.

Chapter 3

Background and notations

3.1 Extreme Value Theory

3.1.1 Introduction to Extreme Value Theory

There are many problems in which we face with risk evaluation question for occurrence of a rare event. For instance, Risk evaluation for financial crises and natural disasters like extreme flooding require the quantification of small occurrence probabilities. The intrinsic property of these events is the low number of occurrence. Therefore, empirical estimations are not reliable and we need theoretical tools to model occurrence of rare events studying the tail behaviour of distributions. Extreme Value Theory(EVT) provides this theoretical foundation.

A discussion about tail behaviour in EVT

In this domain, we usually focus on the distribution which provides ”high” probability or ”high” risk of extreme events. This high probability for occurrence of extremely large events results in ”heavy” tail distribution. Based on the convention, the distribution of X which has heavier tail than the exponential distribution is considered as a desired ”heavy tail” distribution. Hence, our focus is mainly on Heavier Tail than Exponential distributions, so-called ”HTE”, rather than Exponential distribution(”Exp”) or Lower Tail than Exponential distribution(”LTE”).

Remark. In EVT, the heavy tail concept is not same as heavy tail concept that is based on kurtosis and comparision between kurtosis of a distribution with kurtosis of a normal distribution. On the contrary, it is based on the comparision with exponential tail distribution.

FIGURE for tail behaviour

What is EVT going to do?

In the classical theory, one is often interested in the behaviour of the mean or average. This average will then be described through the expected value $E(X)$ of the distribution. On the basis of Law of Large Numbers, ”LLN”, the sample mean $\bar{X}$ is used as a consistent estimator of $E(X)$. Furthermore, the Central Limit Theorem, ”CLT”, yields the asymptotic behaviour of the sample mean. This asymptotic can be used to provide a confidence interval for $E(X)$ in the case the sample size is sufficiently large, which is the vital condition for using CLT. [Beirlant]
Therefore, for mean of the samples we had CLT which explains that for some distribution, F, the mean variable asymptotically converges to a variable which follows a $\mathcal{N}ormal$ distribution. Using some scale and location standardization, we can obtain a $\mathcal{N}ormal$ standard distribution:

$$\lim_{n\to\infty} \frac{\bar{X}-\mu}{\frac{\sigma_X}{\sqrt{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \tag{3.1}$$

But it only works when the distribution is not very fat tail. Similarly, we need kind of limit theorem which provides asymptotic distribution for extremes.
In the problem of finding an asymptotic distribution for a desired statsitics over a sequence of random variabels with identical distribution F two main questions arise:

1. What is the limiting distribution for desired statsitics
2. For which F any such limit is attained.

In the following we precisely explain how does EVT deal with these problems.

What are extremes in math language?

But before going to fundamental blocks for EVT, we must specify how we define these extremes. There could be many approches for this purpose, but we mainly focus on three of them:

1. Block Maxima:
   It is the fundamental and most classic approach in the extreme value theory. In the Block Maxima one is supposed to have observe the maximum values of some quantities over a number of "blocks" as extremes. Typically, a "block" is a period of time. So in this approach, one recognizes maximum over blocks as extreme events. FIGURE for Block Maxima

2. Peaks Over Threshold:
   In the Peaks Over Threshold(POT) method one is instead supposed to have observed all values which are larger than some suitable thresholds. FIGURE for POT

3. Poisson Point Process:
   In the Poisson Point Process approach, one is interested to evaluate the probability of extreme events' occurrence in a specified extreme region. The idea of poisson process is to model probability of number of rare events happening during an interval by a $\mathcal{Poisson}$ distribution. Typically, a $\mathcal{Poisson}$ randome variable is used to describe the distribution of rare events in a large population. Poisson process is charachterized by intensity measure of process. The idea of using intensity measure, provides an elegant tool for interpretation of "Block Maxima" and "POT" method in a multivariate setting. FIGURE for PPP

In the following, we briefly explain each approach.

3.1.2 Block Maxima approach

Consider a random sample $\{X_1, X_2, ..., X_n\}$ where $X1, X2, ..., X_n$ is a sequence of random variables from a distribution F. In the Block Maxima method, one studies the statistical behaviour of $M_n = \max\{X_1, X_2, ..., X_n\}$. M_n represents the maximum value of n observations. When we talk about a maximum value we must specify that the observation is maximum among some specific number of observations constructing a "Block" of observations.
Particularly one is intereserd to specify the disitribution of M_n as a repesentative for extreme event or equivalently, find $\mathbb{P}\{M_n \leq z\}$ where $X1, X2, ..., X_n$ are independent and identical random variables from a distribution F:

$$\begin{aligned} \mathbb{P}\{M_n \leq z\} &= \mathbb{P}\{X_1 \leq z, X_2 \leq z, ..., X_n \leq z\}; \\ &= \mathbb{P}\{X_1 \leq z\} \cdot \mathbb{P}\{X_2 \leq z\} \cdot \ ... \cdot \ \mathbb{P}\{X_n \leq z\}; \\ &= F^n(z) \end{aligned} \tag{3.2}$$

But, since F is unknown tgis method fails. Moreover, we have $\mathbb{P}\{M_n \leq z\} = F^n(z)$. Hence, any deviation from true estimation of F results in high deviation for $\mathbb{P}\{M_n \leq z\}$ which is not acceptable. On the other hand, similar to approximated $\mathcal{Normal}$ distribution for sample means which has been asymptotically justified by CLT, we can directly evaluate the asymptotic behaviour of F^n when $n \to \infty$. The problem is that by definition a distribution function F is in $[0,1]$. For any z where $z < z^+$ (z^+ is the upper end point of distribution F) $F(z) < 1$ so as "n" goes to infinity $F^n(z)$ goes to zero. and the distribution so-called degenerates.

Definition 3. Degenerate distribution: In mathematics, a degenerate distribution is the probability distribution of a discrete random variable whose support consists of only one value.

The degenerate distribution is localized at a point z^+ on the real line. The probability mass function is given by

$$\mathbb{P}\{M_n = z^+\} = 1. \tag{3.3}$$

The cummulative distribution function of degenerate distribution is then:

$$\mathbb{P}\{M_n \leq z\} = \begin{cases} 0 & \text{if } M_n < z^+ \\ 1 & \text{if } M_n \geq z^+ \end{cases} \tag{3.4}$$

FIGURE for Degeneration

In the classical theory, where LLN provides a consistent estimator for mean. However, the resulted distribution degenerates and it is the main incentive for introducing CLT to find an asympotic non-degenerate for sample mean. CLT solves this problem by doing a type of renormalization and controlling the speed of convergence.
Similarly, In EVT the degenerating distribution problem is solved by allowing a linear renormalization of M_n:

$$M_n^* = \frac{M_n - b_n}{a_n} \tag{3.5}$$

with proper choices for $\{a_n\}$ and $\{b_n\}$ (scale and location stabilizer repectively) a stable distribution appears as a limit for $\lim_{n\to\infty} \mathbb{P}\{M_n^* < z\}$. So regardless of the distribution of F one can find limit distribution for stabilized Block Maxima M_n^*.

Again in analogue with asymptotic distribution problem for sample mean, in extreme domain we face with the fundamental questions:

1. **Extremal limit problem:** find all possible non-degenerate distribution G that can appear as a limit in:

$$\lim_{n\to\infty} \mathbb{P}\{\frac{M_n - b_n}{a_n} \leq z\} = \lim_{n\to\infty} \mathbb{P}\{M_n^* \leq z\} = G(z) \tag{3.6}$$

2. **Domain of attraction problem:** Suppose that G is a possible limit distribution for the sequence $a_n^{-1}(X_n - b_n)$. What are the necessary and sufficient condition on the distribution F, a_n, and b_n to get that limiting distribution function F. Roughly speaking, one is interested to find a family of distribution F for which the limit G is acheivable. This family or set of distributions are known as Domain of attraction.

   Definition 4. Domain of attraction: Domain of attraction $\mathcal{D}(G)$ is the set of such distribution F, which are asymptotically attracted toward the limit distribution G

In the following, we consider both the problems and specify the possible limits and its domain of attraction for the convergence:

$$\lim_{n\to\infty} \mathbb{P}\{\frac{M_n - b_n}{a_n} \leq z\} = G(z) \tag{3.7}$$

FIGURE Cool figure for possible limit and domain of attraction

Convergence in distribution

Before exploring these two problems in detail we need to explain what are the main approaches for studying distribution convergence. We use weak distribution convergence in this domain.

Definition 5. Weak convergence: M_n^* converges to Z with the distribution function G if the distribution function of M_n^* converges pointwise to the distribution function of Z, G.

$$M_n^* \xrightarrow{\mathcal{D}} Z \tag{3.8}$$

To investigate this convergence we have two methods:

1. **Convergence of the characteristics function:** $\varphi_{M_n^*}(t)$ converges to $\varphi_Z(t)$

$$\lim_{n\to\infty} \quad \varphi_{M_n^*}(t) = \varphi_Z(t) \tag{3.9}$$

2. **Convergence of the expectations:(Helly-Bray theorem)** $\varphi_{M_n^*}(t)$ converges to $\varphi_Z(t)$ if and only if for all real bounded and continuous $\mathcal{I}$

$$\lim_{n\to\infty} \quad E[\mathcal{I}(M_n^*)] = E[\mathcal{I}(Z)] \tag{3.10}$$

Although the first approach is the building block for convergence in CLT, the second approach is the main one in EVT. for more detail see [Beirlant]

Extremal limit problem

There are two main approaches to solve this problem:

1. The classical extremal families
2. The Generalized Extreme Value distribution(GEV)

In the classic approach the limit function can be in any classes of three families. This limit function is any of Gumbel, Fréchet , and Weibull or type I, type II,and type III respectively. The later modern approach introduces a new general approach which is able to model all classic families using some modifications in the shape parameter.
To be precise, from the classical families we only introduce Fréchet family which is required in the following.

$$G(z) = \begin{cases} 0 & \text{if } z \le b \\ \exp(-(\frac{z-b}{a})^{-\alpha}) & \text{if } z > b \end{cases} \tag{3.11}$$

which has the standard version:

$$G(z) = \begin{cases} 0 & \text{if } z \le 0 \\ \exp(-\frac{1}{z}) & \text{if } z > 0 \end{cases} \tag{3.12}$$

Theorem 2. *Let X_1, X_2, ..., X_n be a sequence of independent and identically distributed random variables with distribution function F, and let,*

$$M_n = max\{X_1, X_2, ..., X_n\} \tag{3.13}$$

denote any arbitrary X_i as X; suppose that X is in the max domain of attraction of random vector Z. So,

$$\mathbb{P}\{M_n \le z\} = G(z) \tag{3.14}$$

Then, the Generalized Extreme Value(GEV) distribution has this form:

$$G_\xi(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})^{(-\frac{1}{\xi})}]\} \tag{3.15}$$

Where ξ is the shape parameter or so-called extreme value index(EVI). μ and σ are location and scale parameter respectively.

The later approach overcomes two weakness of classic one:

1. A technique is required to choose which of the three families is the most appropriate for the data;
2. Once such a desicion is made, subsequent inferences presume this choice to be correct, and do not allow for the uncertainity that such a selection involves; eventhough this uncertainity may be substantial.

Max Domain of Attraction problem

The second approach of convergence (Helly-Bray theorem) imposes some conditions for the set of distributions attracting to G and also for proper choices of $\{a_n\}$ and $\{b_n\}$. But before studying these conditions, we need to define some tools.

Definition 6. Inverse distribution function: The inverse function of a distribution function F of random variable X with the probaility space $\Omega = \mathbb{R}$ is a function $F^{\leftarrow} : [0,1] \to \mathbb{R}$ where:

$$F^{\leftarrow}(y) = \inf\{x \ : \ F(x) \geq y\} \tag{3.16}$$

Definition 7. Tail function: The tail function of random variable X on the probaility space $\Omega = \mathbb{R}$ having distribution function F is a function $U : [1,\infty] \to \mathbb{R}$ where:

$$U(y) = F^{\leftarrow}(1 - \frac{1}{y}) \tag{3.17}$$

Remark. Trivially, for x^+ as upper point of distribution function F we have:

$$x^+ := U(\infty) \tag{3.18}$$

,and for the lower point of distribution function F we have:

$$x^- := U(1) \tag{3.19}$$

So any point in the probability space $\mathbb{R}$ can be obtained by $U(\frac{n}{v})$ where v varies from 0 to n

Therefore, the convergence problem imposes:

$$\lim_{n\to\infty} E[\mathcal{I}(a_n^{-1}(M_n - b_n))] = \int_0^1 \mathcal{I}(v)dG_\xi(v) \tag{3.20}$$

Using the later definitions for the left side of equation we have:

$$\lim_{n\to\infty} E[\mathcal{I}(a_n^{-1}(M_n - b_n))] = \lim_{n\to\infty} \int_0^n \mathcal{I}(a_n^{-1}(U(\frac{n}{v}) - b_n)) \cdot e^{-v} dv \tag{3.21}$$

So,

$$\lim_{n\to\infty} \int_0^n \mathcal{I}(a_n^{-1}(U(\frac{n}{v}) - b_n)) \cdot e^{-v} dv = \int_0^1 \mathcal{I}(v)dG_\xi(v) \tag{3.22}$$

As one can observe the right hand side is barely function of v so the left hand side should be a function of v. Roughly speaking, $\lim_{n\to\infty} a_n^{-1}(U(\frac{n}{v}) - b_n))$ must be a function of v.
Since we have this result for all positive v and trivially it is valid for $v = 1$ then $b_n = U(n)$ is an appropriate choice.

Definition 8. Extremal Domain of Attraction condition $\mathcal{C}_\xi(a)$: The natural and crucial condition to be imposed is that for some positive a and any $v > 0$ there exists function h where,

$$h(\frac{1}{v}) := \lim_{n\to\infty} \frac{U(\frac{n}{v}) - U(n)}{a(n)} \tag{3.23}$$

To Satisfy this condition it is required that $\{a_n\}$ in $\mathcal{C}_\xi(a)$ to be a regularly varying function.

Definition 9. Regularly varying function: A function $a(x)$ is a regularly varying function (r.v.) where:

$$\lim_{x\to\infty} \frac{a(ux)}{a(x)} = u^\xi \tag{3.24}$$

Remark. We can also define a regularly varying function as $a(x) := x^\xi l(x)$ where $l(x)$ is a slowly varying function.

Definition 10. Slowly varying function: A function $l(x)$ is a slowly varying function (s.v.) where:

$$\lim_{x\to\infty} \frac{l(ux)}{l(x)} = 1 \tag{3.25}$$

Remark. For instance, the log function is a s.v. function.

Roughly speaking, regularly varying function are those function which asymptotically behave like a power function.[Resnick]

Given these restrictive conditions one can define the specific domain of attraction for each case of $\xi > 0$, $\xi = 0$, and $\xi < 0$ corresponding to the classic families Fréchet , Gumbel, and Weibull. For more detail see [Beirlant]

Block size, Bias-Variance trade-off

For the "Block Maxima" approach we have shown the limiting distribution and corresponding domain of attraction set. The remaining important point is the selection of block size. The proper choose of block size is a bias-variance trade-off. Larger block size provides unbiased estimation of maxima because we are finding the extreme event in a larger group of observation. However, since the number of block and respectively number of extreme observations declines, the estimator has higher variance. In the contrary, smaller block size results in higher bias in selection of maximum as extreme observation in each block, although, it increases the number of blocks and respectively the number of extreme observations and results In lower variance in estimation.

FIGURE Cool figure for bias variance trade-off

Order statitics, a soloution for data waste

The initial problem of extremes is linked with the study on the tail of distributions. However, in "Block Maxima" we just assume that only maximum of a sample contains valuable information about the tail of a distribution, and we somehow waste valuable inforamtion in the other observations.

Now, we want to use other observations having valuable information about extremes. The idea is to use order statistics.

Definition 11. Order Statitics: The k^{th} order statistic of a statistical sample is equal to its k^{th}-smallest value.

$$X_{1,1} \le X_{2,n} \le ... \le X_{k,n} \le ... \le X_{n-1,n} \le X_{n,n} \tag{3.26}$$

For instance n^{th} order statistics for a sequence of observations with size n is the maximum observation. So for the estimation of tail distribution, if we only use the order statistics close to the maximum, then only few order statistics are used and our estimator for the tail distribution shows large variance and low bias. If we run away from the maximum, the variation of estimators diminishes, however, we penalized by larger bias since we are using values which are not ”extremes”.

For the bias-variance trade-off it is cruical to find the optimal k. Since tail estimation on low number of data results in high variance in order to have a low asympotic variance k should be allowed to tend to ∞ together with the sample size n. However, on the other hand $\frac{k}{n}$ should be kept small to avoid high asymptotic bias. The optimal choice of k depends on the second order behaviour in the extremal domain of attraction condition ($\mathcal{C}_\xi$). For more detail see [Beirlant].

3.1.3 Peaks Over Threshold approach

A remarkable issue in modeling with ”Block Maxima” approach is that we only use one observation in each block and it is wasting the other valuable onservations which can help us in tail estimation.

Although, order statistics is a better alternative, this method can be wasteful since it is possible to have more extreme events in one block than the others. In this case we just use the same number of order statsitics on every blocks. So it wastes the extreme events in the mentioned block. Therefore as an alternative to the ”Block Maxima” approach we introduce another approach, called ”Peaks Over Threshold”.

Let $X_1,\ X_2,\ ...,\ X_n$ be a sequence of independent and identically distributed random variables with distribution function F. It is intuitive to consider extreme events as events that X_i peaks over some high threshold u. Then one can say the tail distribution is specified as,

$$\mathbb{P}\{X > u + y|X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \qquad y > 0 \tag{3.27}$$

FIGURE figure for threshold exceedance

However, in practice we usually do not know the distribution F. Hence, similar to GEV, we introduce the Generalized Pareto distribution which approximate the tail distribution. The following theorem explains the asymptotic characterization of this approach and the link between ”Block Maxima” and ”Peaks Over Threshold” approaches.

Theorem 3. *Let $X_1,\ X_2,\ ...,\ X_n$ be a sequence of independent and identically distributed random variables with distribution function F, and let,*

$$M_n = max\{X_1,\ X_2,\ ...,\ X_n\} \tag{3.28}$$

denote any arbitrary X_i as X; suppose that X is in the max domain of attraction of random vector Z. So,

$$\mathbb{P}\{M_n \leq z\} = G(z) \tag{3.29}$$

where base on GEV,

$$G_\xi(z) = \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})^{(-\frac{1}{\xi})}]\} \tag{3.30}$$

for some μ, $\sigma > 0$, and ξ. Then for large enough u, the distribution function of $(X - u)$, conditional on $X > u$ is approximately

$$H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-\frac{1}{\xi}}, \tag{3.31}$$

where, $\{y :\ y > 0,\ 1 + \frac{\xi y}{\tilde{\sigma}} > 0\}$ and $\tilde{\sigma} = \sigma + \xi(u - \mu)$

So this theorem justified $H(y)$ as a limiting distribution for the distribution function of $(X-u)$, conditional on $X > u$ when u increases toward the upper end point of F.

The family of distribution denoted by Equation(3.31) is called the Generalized Pareto distribution (GPD). The Theorem(3) shows that GEV distribution of "Block Maxima" approach and GPD on "Peaks Over Threshold" approach are bijectively associated to each other.

Generalized Pareto distribution is characterized with two main properties:

1. Exceedances over a high threshold asymptotically have a multivariate Generalized Pareto distribution if and only if maxima are asymptotically GEV distributed.

2. The GPD is the only distribution which is preserved under change of exceedance or equivalently it is stable under change of threshold.

The stability

In the classical theory we also have stability concept. For example CLT provides a mean-stable distribution for sample mean, where any change in sample size does not change the family of convergent distribution. Mean-stability is associated with the properties of $\mathcal{N}ormal$ distribution. A normal random vector remains normal after applying any location-scale transformation.

Parallelly in the extreme value theory, GEV provides a max-stable distribution. i.e., such that any change in block size only leads to a change of location and scale parameters of the distribution . μ and σ respectively. But it does not change the shape parameter ξ.

Moreover, GPD introduces a threshold-stable distribution, i.e. a family of distributions remain unaltered after changing the threshold. A change in the threshold of GPD barely results in a scale transformation of the original distribution.

Thresold selection, Bias-Variance trade-off

We faced with the issue of block size selection in Block Maxima approach and the bias-variance trade-off in selection of optimal block size. Similarly finding a proper threshold for Peaks over Thresold approach is vital. Too low a threshold is probably leading to high bias since the number of observations which are not extreme and are not in the tail increases. However, if the threshold is high then the estimator considers a few observations as extremes leading to a high variation for estimation. For more detail in threshold selection see [Coles]

3.1.4 Poisson Point Process approach

This approach enables us to have an elegant interpretation about extreme value distributions and connect later approaches to a more intuitive point of view. In the Poisson Point Process approach attempts to model the tail distribution using probability of occurrence of scatter point events in a "extreme" space.

Let $\mathcal{A}$ be a subset of an Euclidean space. In the following, we consider a subset of $[0,+1)$, $\mathbb{R}$, or $\mathbb{R}^d$. We distribute points randomly in $\mathcal{A}$ and consider a simple notation allowing to count the number of points that lie in a closed set A where $A \subset \mathcal{A}$. If $\mathcal{A} = \mathbb{R}^d$, then A is a sub-space in $\mathbb{R}^d$.

Suppose that $\{X_t : t \geq 0\}$ stands for consecutive points in the state space $\mathcal{A}$. If we define the discrete measure

$$\varepsilon_{X_t}(A) = \begin{cases} 0 & \text{if } X_t \in A \\ 1 & \text{if } X_t \notin A \end{cases} \tag{3.32}$$

then summing on t, we get the total number of points X_t in A

If we define the counting measure N by $N(\cdot) = \sum_t \varepsilon_{X_t}(\cdot)$, then $N(A) = \sum_t \varepsilon_{X_t}(A)$ is the random number of points that lie in the set A. N is called a point process and $\{X_t\}$ are called points. The intensity of N, or mean measure of N is defined by $\Lambda(A) = E[N(A)]$, which corresponds to the expected number of points in A. [Scaillet] One can observe for all $A \subset \mathcal{A}$ in a measurable space $\mathcal{A}$,

$$N(A) \sim \text{Poi}(\Lambda(A)) \tag{3.33}$$

Using the notion of intensity measure we are able to parametrize the probability of occurrence of extreme events happened in a subset of $\mathcal{A}$ where is an "extreme" region. There are two interpretation for this subset corresponding to two approaches "Block Maxima" and "Peaks Over Thresold". Here we simply stay with one-dimensional extreme variabel.

1. Block Maxima: The cummulative distribution function for renormalized maxima M_n^* corresponds to

   $$\mathbb{P}[M_n^* \leq z] = \mathbb{P}[\frac{M_n - b_n}{a_n} \leq z] \tag{3.34}$$

   Clearly, the event $\{M_n^* : M_n^* \leq z\}$ is equivalent to the event {for a Point process on M_n^* : $N(A_z) = 0\}$ where $A_z := (0,1) \times (z, \infty)$ Roughly speaking the event of having normalized maxima smaller than some value z is equivalent to the event in which the number of occurrence of the normalized maxima in the region from z to ∞ is none. so,

   $$\begin{aligned} \mathbb{P}[M_n^* \leq z] &= \mathbb{P}[\frac{M_n - b_n}{a_n} \leq z] \\ &= \mathbb{P}[N(A_z) = 0] \\ &= \exp(-\Lambda(A_z)) \\ &= \exp\{-[1 + \xi(\frac{z-\mu}{\sigma})^{-1/\xi}]\} \end{aligned} \tag{3.35}$$

2. Peaks Over Threshold: The distribution of exceedance of renormalized maxima M_n^* over a value z given the information that it is already in the "extreme" region defined by a threshold u can be evaluated by Point process. Here the extreme region is defined by a threshold and not by a area corresponding to occurrence of maxima points.

   Suppose that we factorize intensity measure $\Lambda(A_z)$ to two measure Λ_1 and Λ_2,

   $$\Lambda(A_z) = \Lambda_1([t_1, t_2]) \times \Lambda_2([z, \infty)) \tag{3.36}$$

   Where,

   $$\begin{aligned} \Lambda_1([t_1, t_2]) &= (t_2 - t_1) \\ \Lambda_2([z, \infty)) &= (1 + \frac{\xi(z-\mu)}{\sigma})^{-1/\xi} \end{aligned} \tag{3.37}$$

   Then,

   $$\begin{aligned} \mathbb{P}\{\frac{M_n - b_n}{a_n} > z \mid \frac{M_n - b_n}{a_n} > u\} &= \frac{\Lambda_2[z,\infty)}{\Lambda_2[u,\infty)}, \\ &= \frac{n^{-1}[1 + \xi(z-\mu)/\sigma]^{-1/\xi}}{n^{-1}[1 + \xi(u-\mu)/\sigma]^{-1/\xi}}, \\ &= [1 + \frac{\xi(z-\mu)/\sigma}{1 + \xi(u-\mu)/\sigma}]^{-1/\xi}, \\ &= [1 + \xi(\frac{z-u}{\tilde{\sigma}})]^{-1/\xi}, \qquad \tilde{\sigma} = \sigma + \xi(u - \mu) \end{aligned} \tag{3.38}$$

3.1.5 Introduction to Multivariate Extreme Value Theory

One Might face with a risk evaluation problem in a multivariate setting. In these type of problems, the univariate marginals can be modeled by the provided univariate models in later approches. Although, one might be interested in the potential information that one of these extreme events gives about the others. Roughly speaking, in addition to studying marginal behaviour we must study the dependence structure in a multivariate extreme setting.In this section we extend the probability theory of univariate approaches in extreme value theory to multivariate setting.

3.1.6 Bivariate Block Maxima

For simplicity we start from a bivariate case. Let $(X_1,\ Y_1)$, $(X_2,\ Y_2)$, ..., $(X_n,\ Y_n)$ is a sequence of vectors that are versions of a random vector $(X,\ Y)$ having distribution $F(x,\ y)$. We know that the "Block Maxima" for each of X and Y can be defined as:

$$M_{x,n} = \max_{i=1,\dots,n} \{X_i\} \qquad \text{and} \qquad M_{y,n} = \max_{i=1,\dots,n} \{Y_i\} \tag{3.39}$$

However we need to define what exactly makes a multivariate observation extreme? Is it sufficient that a single coordinate attains an extreme value or all coordinates must be extreme simultaneously. We have a new notion for "Block Maxima" in the multivariate setting which is the vector of componentwise maxima.

$$M_n = (M_{x,n},\ M_{y,n}) = (\max_{i=1,\dots,n} \{X_i\},\ \max_{i=1,\dots,n} \{Y_i\}) \tag{3.40}$$

This artificial point is not necessary a point in the observation set. If we assume that marginal distribution are Fréchet distribution then:

$$\mathbb{P}[\frac{M_{x,n}}{n} < z] = \mathbb{P}[\frac{M_{y,n}}{n} < z] = \exp(-\frac{1}{z}) \tag{3.41}$$

It is valid for all n, since all members of GEV including Fréchet are max-stable. i.g., they are stable under change of block size.

Let $M_n^* = (M_{x,n}^*,\ M_{y,n}^*) = (\frac{\max_{i=1,\dots,n}\{X_i\}}{n},\ \frac{\max_{i=1,\dots,n}\{Y_i\}}{n})$ where the X and Y are independent vectors with standard Fréchet marginal distribution then,

$$\mathbb{P}(M_n^* < z) = \mathbb{P}(M_{x,n}^* < z_1,\ M_{y,n}^* < z_2) \xrightarrow{\mathcal{D}} G(z_1, z_2) \tag{3.42}$$

where G is a non-degenerate distribution function, G has the form

$$G(z_1, z_2) = \exp(\Lambda(z_1, z_2)), \qquad z_1 > 0, \quad z_2 > 0 \tag{3.43}$$

where,

$$\Lambda(z_1, z_2) = 2 \int_0^1 \max(\frac{\omega}{z_1}, \frac{1-\omega}{z_2})\ dH(\omega) \tag{3.44}$$

and H is a distribution function on $[0,1]$ satisfying the mean constriant

$$\int_0^1 \omega\ dH(\omega) = \frac{1}{2} \tag{3.45}$$

Any bivariate extreme value distribution $G(z_1, z_2)$ is in one-one correspondence with the set of distribution H on $[0,1]$ where H is satisfying the mean constraint. When H is differentiable we can write $dH(\omega)$ as $h(\omega)d\omega$ in Equation (3.44).In some special but important cases H is not differentiable.

Example 1. *Let $H(\omega)$ has this form:*

$$H(\omega) = \begin{cases} 0 & \text{if } \omega < 0, \\ \frac{1}{2} & \text{if } 0 \leq \omega < 1, \\ 1 & \text{if } 1 \leq \omega. \end{cases} \tag{3.46}$$

or equivalently,

$$dH(\omega) = \begin{cases} \frac{1}{2} & \text{if } \omega = 0, \\ \\ \frac{1}{2} & \text{if } \omega = 1. \end{cases} \tag{3.47}$$

FIGURE $H(\omega)$ independent

One can validate that $H(\omega)$ is a distribution function satisfying mean constraint.

Trivially,

$$\Lambda(z_1,z_2) = 2 \times [\max(\frac{0}{z_1}, \frac{1}{z_2}) \times \frac{1}{2} + \max(\frac{1}{z_1}, \frac{0}{z_2}) \times \frac{1}{2}] \tag{3.48}$$

where, $z_1, z_2 > 0$

So,

$$\Lambda(z_1,z_2) = 2 \times [\frac{1}{z_2} \times \frac{1}{2} + \frac{1}{z_1} \times \frac{1}{2}] = z_1^{-1} + z_2^{-1} \tag{3.49}$$

$$\begin{aligned} G(z_1,z_2) &= \exp\{-(z_1^{-1} + z_2^{-1})\}, \\ &= \exp\{-z_1^{-1}\} \times \exp\{-z_2^{-1}\}, \\ &= G(z_1) \times G(z_2) \end{aligned} \tag{3.50}$$

where $G_1(z_1)$ and $G_2(z_2)$ are standard marginal Fréchet .

Example 2. *Let $H(\omega)$ is a measure that put entire unit mass on $\omega = 0.5$:*

$$H(\omega) = \begin{cases} 0 & \text{if } \omega < \frac{1}{2}, \\ 1 & \text{if } \omega \geq \frac{1}{2}. \end{cases} \tag{3.51}$$

or equivalently,

$$dH(\omega) = \begin{cases} 0 & \text{if } \omega \neq \frac{1}{2}, \\ 1 & \text{if } \omega = \frac{1}{2}. \end{cases} \tag{3.52}$$

FIGURE $H(\omega)$ complete dependence

Agian, one can validate that $H(\omega)$ is a distribution function satisfying mean constraint.

Trivially,

$$\Lambda(z_1,z_2) = 2 \times [\max(\frac{1}{2z_1}, \frac{1}{2z_2}) \times 1] \tag{3.53}$$

where, $z_1, z_2 > 0$

So,

$$\Lambda(z_1,z_2) = 2 \times [\frac{1}{2}\max(\frac{1}{z_1}, \frac{1}{z_2})] = \max({z_1}^{-1}, {z_2}^{-1}) \tag{3.54}$$

$$G(z_1,z_2) = \exp\{-\max({z_1}^{-1}, {z_2}^{-1})\} \tag{3.55}$$

indicating the complete dependence.

The later examples are two extreme cases of $H(\omega)$ which result in complete independence and complete dependence respectively.

One can see the measure function Λ has homogeneity property for any $a > 0$

$$\Lambda(a^{-1}z_1, a^{-1}z_2) = 2\int_0^1 \max(\frac{a\omega}{z_1}, \frac{a\omega}{z_2})dH(\omega) \tag{3.56}$$

since $a > 0$,

$$\Lambda(a^{-1}z_1, a^{-1}z_2) = 2\int_0^1 a\max(\frac{\omega}{z_1}, \frac{\omega}{z_2})dH(\omega) \quad = a\Lambda(z_1, z_2) \tag{3.57}$$

So, when we have a bivariate extreme value distribution Λ is a homogenuous order-" $-$ 1". One can see if (X, Y) has a bivariate extreme value distribution function G, then M_n has the distribution function,

$$\begin{aligned}\mathbb{P}(M_n \leq z) &= \mathbb{P}\ \{(X_1, Y_1) \leq (z_1, z_2),\ (X_2, Y_2) \leq (z_1, z_2),\ ...,\ (X_n, Y_n) \leq (z_1, z_2)\} \\ &= G^n(z_1, z_2) \\ &= (\exp\{-\Lambda(z_1, z_2)\})^n \\ &= \exp\{-n\Lambda(z_1, z_2)\} \\ &= \exp\{-\Lambda(\frac{z_1}{n}, \frac{z_2}{n})\} \\ &= G(n^{-1}z_1, n^{-1}z_2)\end{aligned} \tag{3.58}$$

Which shows that G posses a bivariate version of the property of max-stability, i.e. the distribution of M_n preserved in the same family apart from re-scaling by inverse of block's size, n^{-1}.

The class of possible limits as G is wide. The best approach to find a possible class for G is to choose H from a parametric distribution family leading to a correspondance G. Given that, one can obtain different families for G like *logistic*, *Hüsler-Reiss*, and etc.

3.1.7 Multivariate Block Maxima

Now, we can extend the notion of max-stable componentwise maxima distribution to multivariate case.

Theorem 4. *Let* $\mathbf{X_i} = (X_{i1}, X_{i2}, ..., X_{id})$, $i = 1, 2, ..., n$ *have been sampled from* $\mathbf{X} = (X_1, X_2, ..., X_d)$ *and denote componentwise maxima* $\mathbf{M_n} = (M_{1n}, M_{2n}, ..., M_{dn})$ *where* $M_{in} = \max_{j=1}^{n} X_{ij}$
Then, there are sequence of $b_{jn} \in \mathbb{R}$ *and* $a_{jn} > 0$, $j = 1, 2, ..., d$ *satisfying regular variation, such that,*

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\frac{M_{jn} - b_{jn}}{a_{jn}} < z_j) &= G_j(z_j) \\ &= \exp\{-(1 + \xi_j z_j)^{-1/\xi_j}\}, \qquad z_j \in \mathbb{R}\end{aligned} \tag{3.59}$$

So, one can model marginals by GEV. To focus on the dependence structure one can normalize marginals to the standard Fréchet distribution.
Then the standardized vector X is said to be in the max-domain of attraction of the random variabl $\mathbf{Z} = (Z_1, Z_2, ..., Z_d)$ *if for any* $\mathbf{z} = (z_1, z_2, ..., z_d)$,

$$\lim_{n\to\infty} \mathbb{P}(\frac{\max_{i=1,2,...,n} X_{i1}}{n} \leq z_1, ..., \frac{\max_{i=1,2,...,n} X_{id}}{n} \leq z_d) = \mathbb{P}(\mathbf{Z} \leq \mathbf{z}) \tag{3.60}$$

where $\mathbf{z} \in \mathcal{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$. $\mathbf{Z}$ *is a multivariate max-stable distribution with standard Fréchet marginals,*

$$\mathbb{P}(Z_j \leq z_j) = \exp(-\frac{1}{z_j}), \qquad z_j \geq 0 \tag{3.61}$$

and,

$$\mathbb{P}(\mathbf{Z} \le \mathbf{z}) = \exp(-\Lambda(z)), \qquad \mathbf{z} \in \mathcal{E} \tag{3.62}$$

where the ***exponent measure*** *Λ is a Radon measure on the cone $\mathcal{E} = [0,\infty)^d \setminus \{\mathbf{0}\}$ and $\Lambda(\mathbf{z})$ is shorthand for $\Lambda(\mathcal{E} \setminus [\mathbf{0},\mathbf{z}])$*

Beforehand, in Equation(3.57) we have shown the exponent measure has homogeneity property of order-" -1" in the biavariate case. One can shown the exponent measure has homogeneity property in higher dimension as well. WHAT IS THE ORDER OF HOMOGENEITY Therefore, not only does G have a max-stable marginals, it is itself a max-stable distribution as well. So it is in its own domain of attraction.

If Λ is absolutely continuous with respect to a Lebesgue measure on $\mathcal{E}$ its Radon-Nikodym derivative denoted by λ, has the following properties:

- Max-stability: Homogeneity of order-$(d+1)$, i.e.

$$\lambda(ty) = t^{d+1}\lambda(y) \text{ for any } t > 0 \text{ and } y \in \mathcal{E}() \tag{3.63}$$

- Standard Fréchet marginals: Normalized marginals, i.e. for any $i = 1, 2, \dots, d$

$$\int_{y \in \mathcal{E}; y_i > 1} \lambda(y) dy = 1 \tag{3.64}$$

In parallel with bivariate case, we can define parametric densities for $\lambda(y)$ and find correspondence G.

3.1.8 Bivariate Peaks Over Threshold

For univariate case we found a class of approximations to the tail of the distribution function $Y = X - u$ conditional on $X > u$,

$$H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi} \tag{3.65}$$

defined on $\{y : y > 0 \text{ and } (1 + \frac{\xi y}{\tilde{\sigma}}) > 0\}$ where $\tilde{\sigma} = \sigma + \xi(u - \mu)$
where μ, σ, and ξ are location, scale, and shape parameters of the Block Maxima distribution.

Then for a bivariate case one can extend the concept of extreme observation to an observation which is extreme in at least one component. Roughly speaking, the extreme observation at least in one component exceeds over a high threshold.

Suppose that $(x_1,\ x_2)$ is independently and identically sampled from (X_1, X_2). for n times. Like before, one can define the bivariate block maxima M_n like,

$$\mathbf{M_n} = (M_{x_1,n}\ ,\ M_{x_2,n}) \tag{3.66}$$

and $\mathbf{M_n^*}$ as,

$$\mathbf{M_n^*} = \frac{\mathbf{M_n} - \mathbf{b_n}}{\mathbf{a_n}}, \qquad \text{where } \mathbf{a_n} \text{ is } r.v. \tag{3.67}$$

One can show that,

$$\begin{aligned}\mathbb{P}(\mathbf{M_n^*} < \mathbf{z}) &= \mathbb{P}(M^*_{x_1,n} < z_1\ ,\ M^*_{x_2,n} < z_2)\\ &= \mathbb{P}(\frac{M_{x_1,n} - b_{1,n}}{a_{1,n}} < z_1\ ,\ \frac{M_{x_2,n} - b_{2,n}}{a_{2,n}} < z_2)\\ &= \mathbb{P}\Big\{ \{\frac{X_{1,1} - b_{1,n}}{a_{1,n}} < z_1, \frac{X_{2,1} - b_{2,n}}{a_{2,n}} < z_2\} \cap \ldots \cap \{\frac{X_{1,n} - b_{1,n}}{a_{1,n}} < z_1, \frac{X_{2,n} - b_{2,n}}{a_{2,n}} < z_2\} \Big\}\\ &= \mathbb{P}\Big\{ \frac{X_1 - b_{1,n}}{a_{1,n}} < z_1, \frac{X_2 - b_{2,n}}{a_{2,n}} < z_2 \Big\}^n\\ &= F^n_{\frac{\mathbf{X-b_n}}{\mathbf{a_n}}}(\mathbf{z})\end{aligned} \tag{3.68}$$

We have shown that when X in the max domain of attraction Z, then

$$\mathbb{P}(\mathbf{M_n^*} < \mathbf{z}) \xrightarrow{\mathcal{D}} G(\mathbf{z}) \tag{3.69}$$

so,

$$F^n_{\frac{\mathbf{X-b_n}}{\mathbf{a_n}}}(\mathbf{z}) \xrightarrow{\mathcal{D}} G(\mathbf{z}) \tag{3.70}$$

Now for a common and increasing threshold $u := o(n)$. We consider $\mathbf{u} = (u, ..., u)$, $\mathbf{b_n} = \mathbf{u}$ and $\mathbf{a_n} = \mathbf{u}$ one can show these choices are compatible with required conditions $\mathcal{C}_\xi(a)$ in EVT.

$$F^n_{\frac{\mathbf{X-u}}{\mathbf{u}}}(\mathbf{z}) \xrightarrow{\mathcal{D}} G(\mathbf{z}) \tag{3.71}$$

Trivially, $\mathbb{P}(\frac{\mathbf{X-u}}{\mathbf{u}} < \mathbf{z}) = \mathbb{P}(\mathbf{X} < \mathbf{uz} + \mathbf{u})$ where $\mathbf{uz}$ is a componentwise multiplication, (uz_1, uz_2) So,

$$F^n_X(\mathbf{uz} + \mathbf{u}) \xrightarrow{\mathcal{D}} G(\mathbf{z}) \tag{3.72}$$

With a log we have:

$$n \log(F_X(\mathbf{uz} + \mathbf{u})) \xrightarrow{\mathcal{D}} \log G(\mathbf{z}) \tag{3.73}$$

$$-n \log(F_X(\mathbf{uz} + \mathbf{u})) \xrightarrow{\mathcal{D}} -\log G(\mathbf{z}) \tag{3.74}$$

From taylor series we have:

$$f(x) \simeq f(a) + f'(a)(x-a) \tag{3.75}$$

Now we want to use taylor approximation where $u \to \infty$ or equivalently $F_X(\mathbf{uz} + \mathbf{u}) \to 1$. Consequently,

$$-\log(F_X(\mathbf{uz} + \mathbf{u})) \simeq -\log(1) - (\frac{1}{1})(F_X(\mathbf{uz} + \mathbf{u}) - 1) \tag{3.76}$$

$$-\log(F_X(\mathbf{uz} + \mathbf{u})) \simeq 1 - F_X(\mathbf{uz} + \mathbf{u}) \tag{3.77}$$

$$-\log(F_X(\mathbf{uz} + \mathbf{u})) \simeq \bar{F}_X(\mathbf{uz} + \mathbf{u}) \tag{3.78}$$

So,

$$-n\bar{F}_X(\mathbf{uz} + \mathbf{u}) \xrightarrow{\mathcal{D}} -\log G(\mathbf{z}) \tag{3.79}$$

Now, for the exceedance over a threshold u we have,

$$\begin{aligned}\mathbb{P}(\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \leq \mathbf{z}|\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{0}) &= \frac{\mathbb{P}\Big[\{\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \leq \mathbf{z}\}\cap\{\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{0}\}\Big]}{\mathbb{P}\{\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{0}\}} \\ &= \frac{\mathbb{P}(\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{z}\wedge\mathbf{0}) - \mathbb{P}(\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{z})}{\mathbb{P}(\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{0})}, \qquad \text{for } z \nleq 0\end{aligned} \tag{3.80}$$

FIGURE for showing (80)

Since,

$$\begin{aligned}\mathbb{P}(\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{z}) &= \mathbb{P}(\mathbf{X} \nleq \mathbf{uz}+\mathbf{u}) \\ &= \bar{F}(\mathbf{uz}+\mathbf{u})\end{aligned} \tag{3.81}$$

base on Equation(3.80) we have,

$$\begin{aligned}\mathbb{P}(\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \leq \mathbf{z}|\frac{\mathbf{X}-\mathbf{u}}{\mathbf{u}} \nleq \mathbf{0}) &= \frac{\bar{F}(\mathbf{u}(\mathbf{z}\wedge\mathbf{0})+\mathbf{u}) - \bar{F}(\mathbf{uz}+\mathbf{u})}{\bar{F}(\mathbf{u0}+\mathbf{u})} \\ &\xrightarrow{\mathcal{D}} \frac{\log G(\mathbf{z}\wedge\mathbf{0}) - \log G(\mathbf{z})}{\log G(\mathbf{0})}\end{aligned} \tag{3.82}$$

where $\mathbf{z} \in \mathcal{E} = [0,\infty)^d \setminus \{0\}$

Consequently,

$$\mathbb{P}(\frac{\mathbf{X}}{\mathbf{u}} \leq \mathbf{z}+\mathbf{1}|\frac{\mathbf{X}}{\mathbf{u}} \nleq \mathbf{1}) \xrightarrow{\mathcal{D}} \frac{\log G((\mathbf{z}+\mathbf{1})\wedge\mathbf{1}) - \log G(\mathbf{z}+\mathbf{1})}{\log G(\mathbf{1})} \tag{3.83}$$

or for $\mathbf{z}' = \mathbf{z}+\mathbf{1}$

$$\mathbb{P}(\frac{\mathbf{X}}{\mathbf{u}} \leq \mathbf{z}'|\frac{\mathbf{X}}{\mathbf{u}} \nleq \mathbf{1}) \xrightarrow{\mathcal{D}} \frac{\log G(\mathbf{z}'\wedge\mathbf{1}) - \log G(\mathbf{z}')}{\log G(\mathbf{1})} \tag{3.84}$$

where $\mathbf{z}' \in \mathcal{L} = \{\mathbf{z} \in \mathcal{E} : ||\mathbf{z}||_\infty > 1\}$

We know that if $G(\mathbf{z})$ has standard Fréchet marginals, then,

$$\mathbb{P}(\mathbf{Z} \leq \mathbf{z}) = \exp\{\Lambda(\mathbf{z})\}, \qquad \mathbf{z} \in \mathcal{E}, \tag{3.85}$$

where the exponent measure Λ is a Radon measure on the cone $\mathcal{E} = [0,\infty)^d \setminus \{0\}$, and $\Lambda(\mathbf{z})$ is a shorthand for $\Lambda(\mathcal{E} \setminus [\mathbf{0},\mathbf{z}])$.

Then from Equation(3.84) the multivariate distribution of threshold exceedance of X satisfies

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{z}') = \mathbb{P}(\frac{\mathbf{X}}{\mathbf{u}} \leq \mathbf{z}'|\frac{\mathbf{X}}{\mathbf{u}} \nleq \mathbf{1}) \xrightarrow{\mathcal{D}} \frac{\Lambda(\mathbf{z}'\wedge\mathbf{1}) - \Lambda(\mathbf{z}')}{\Lambda(\mathbf{1})}, \qquad \mathbf{z}' \in \mathcal{L} \tag{3.86}$$

consequently,

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{z}) = \begin{cases} 0 & \text{if } \mathbf{z} \in \mathcal{E} \setminus \mathcal{L} = [0,1]^d \setminus \{\mathbf{0}\}, \\ \dfrac{\Lambda(\mathbf{z}\wedge\mathbf{1}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})} & \text{if } \mathbf{z} \in \mathcal{L} \end{cases} \tag{3.87}$$

trivially,

$$\begin{aligned}\frac{\Lambda(\mathbf{z}\wedge\mathbf{1}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})} &= \frac{\Lambda(\mathbf{z}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})} \\ &= 0 \qquad\qquad \mathbf{z} \in [0,1]^d \setminus \{\mathbf{0}\}\end{aligned} \tag{3.88}$$

so, using Equations(3.87) and (3.88):

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{z}) = \mathbb{P}(\frac{\mathbf{X}}{\mathbf{u}} \leq \mathbf{z}|\frac{\mathbf{X}}{\mathbf{u}} \nleq \mathbf{1}) \xrightarrow{\mathcal{D}} \frac{\Lambda(\mathbf{z}\wedge\mathbf{1}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})}, \qquad \mathbf{z} \in \mathcal{E} \tag{3.89}$$

3.1.9 Multivariate Peaks Over Threshold

Now, one can use the sketch of what we have introduced in the bivariate Peaks Over Threshold for a multivariate case.

Theorem 5. *Let* $\mathbf{X} = (X_1, X_2, ..., X_d)$ *and* $\mathbf{X_i} = (X_{i1}, X_{i2}, ..., X_{id}), \quad i = 1,...,d,$ *be independent copies of* $\mathbf{X}$ *and denote the componentwise maximum by* $\mathbf{M_n} = (M_{1n}, M_{2n}, ..., M_{dn})$ *where* $M_{jn} = \max_{i=1}^{n} X_{ij}$ *and vector* $\mathbf{X}$ *has been normalized to standard Pareto marginals.*
The standardized vector $\mathbf{X}$ *is said to be in the max-domain of attraction of the random vactor* $\mathbf{Z} = (Z_1, ..., Z_d)$ *if for any* $\mathbf{z} = (z_1, ..., z_d)$

$$\lim_{n\to\infty} \mathbb{P}(\max_{i=1,...,n} X_{i1} \leq nz_1, ..., \max_{i=1,...,n} X_{id} \leq nz_d) = \mathbb{P}(\mathbf{Z} \leq \mathbf{z}) \tag{3.90}$$

In this case, Z is max-stable with standard Fréchet marginals $\mathbf{P}(Z_j \leq z) = \exp(-1/z), z \geq 0,$ *and*

$$\mathbb{P}(\mathbf{Z} \leq \mathbf{z}) = \exp\{-\Lambda(\mathbf{z})\}, \qquad z \in \mathcal{E} \tag{3.91}$$

where the exponent measure Λ *is a Radon measure on the cone* $\mathcal{E} = [0,\infty)^d \setminus \{0\}$*, and* $\Lambda(\mathbf{z})$ *is a shorthand for* $\Lambda(\mathcal{E} \setminus [\mathbf{0},\mathbf{z}])$*.*

Then, the mutivariate Pareto distribution of exceedance of $\mathbf{X}$ *satisfies*

$$H_Y(z) = \mathbb{P}(\mathbf{Y} \leq \mathbf{z}) = \mathbb{P}(\frac{\mathbf{X}}{\mathbf{u}} \leq \mathbf{z} | \frac{\mathbf{X}}{\mathbf{u}} \nleq \mathbf{1}) \xrightarrow{\mathcal{D}} \frac{\Lambda(\mathbf{z} \wedge \mathbf{1}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})}, \qquad \mathbf{z} \in \mathcal{E} \tag{3.92}$$

The distribution of the limiting random vector Y is called a multivariate Pareto distribution.

For proof and more details see [Rootzén Tajvidi(2006)]

Remark. $\mathbf{Z}$ and $\mathbf{Y}$ are one-one correspondance and their distribution mutually determine eachother.

One assumes that the distribution of $\mathbf{Y}$ has a positive and continuous density of $f_{\mathbf{Y}}(\mathbf{y})$ on $\mathcal{L}$, which is,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{\partial^d}{\partial y_1 ... \partial y_d} \mathbb{P}(\mathbf{Y} \leq \mathbf{y}) \\ &= \frac{\frac{\partial^d}{\partial y_1 ... \partial y_d} \mathbf{\Lambda}(\mathbf{y} \wedge \mathbf{1}) - \frac{\partial^d}{\partial y_1 ... \partial y_d} \mathbf{\Lambda}(\mathbf{y})}{\mathbf{\Lambda}(\mathbf{1})}, \qquad \mathbf{y} \in \mathcal{L} \end{aligned} \tag{3.93}$$

Since $\mathbf{y} \in \mathcal{L}$, then $\exists j: \ y_j > 1$ so along at least one coordinate $y_j \wedge 1 = 1$. Therefore, $\Lambda(\mathbf{y} \wedge \mathbf{1})$ is always constant along at least one coordinate for $\mathbf{y} \in \mathcal{L}$. Hence, $\frac{\partial^d}{\partial y_1 ... \partial y_d} \mathbf{\Lambda}(\mathbf{y} \wedge \mathbf{1}) = 0$.
Moreover, since we have defined $\Lambda(\mathbf{y})$ as a shorthand for $\Lambda(\mathcal{E} \setminus [\mathbf{0},\mathbf{y}])$; we have a negative sign in $-\frac{\partial^d}{\partial y_1 ... \partial y_d} \mathbf{\Lambda}(\mathbf{y}) = \lambda(\mathbf{y})$

So,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{\partial^d}{\partial y_1 ... \partial y_d} \mathbb{P}(\mathbf{Y} \leq \mathbf{y}) \\ &= \frac{\lambda(\mathbf{y})}{\mathbf{\Lambda}(\mathbf{1})}, \qquad \mathbf{y} \in \mathcal{L} \end{aligned} \tag{3.94}$$

Rootzén Tajvidi(2006) have shown this definition of multivariate Pareto distribution is compatible with two desired characteristics:

- Exceedance of suitably coordinated levels asymptotically have a multivariate generalized Pareto distribution if and only if componentwise maxima are Extreme Value distributed asymptotically. (one-one correspondance with Extreme Value distribution)
- The multivariate generalized Pareto distribution is the only which is preserved under a suitably coordinated change of exceedance levels. (threshold-stability, max-stability)

3.1.10 Multivariate Poisson Point Process

In the univariate Extreme value theory we explained how extreme events have been linked to a poisson point process. Now, we want to extend this point of view to the multivariate case. The idea behind point process approach is to evaluate the probability of absence of extreme events as scatter points in an "extreme" region. For instance, we have a field which is suspected to be a minefield. Obviously the probability of finding mine in each unit area of the field is too low. So finding mine is a rare or extreme event. Using point process one can assess the probability of number of mines in an arbitrary area. Trivially, one can evaluate the probability of absence of mines in the area. In parallel, in Extreme value theory, Poisson Point process aproach defines an area as "extreme" region and assess the probability of having any scatter point, i.g. extreme event, in that area.

Bivariate Poisson Point Process approach

This approach has been introduced in the thesis in order to provide a nice interpretation and an excellent link between other approaches. Therefore, we (start with) barely stay on bivariate case to avoid complexity.

Suppose that (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) is a sequence of independent bivariate observations from (X, Y) having a joint distribution with standard Fréchet marginals that satisfies the convergence for componentwise maxima.

$$\mathbb{P}(M^*_{x,n} \leq z_1,\ M^*_{y,n} \leq z_2) \xrightarrow{\mathcal{D}} G(z_1, z_2) \tag{3.95}$$

see FIGURE for the area defined by $M^*_{x,n} > z_1,\ M^*_{y,n} > z_2$ From the block maxima point of view the "extreme" region is the region that componentwise maximum exceeds over some values. So for $A_{\mathbf{z}} = \{(0,\infty) \times (0,\infty) \setminus (0, z_1) \times (0, z_2)\}$ from point process properties we have,

$$\begin{aligned} \mathbb{P}(M^*_{x,n} \leq z_1,\ M^*_{y,n} \leq z_2) &= \mathbb{P}(N(A_z) = 0) \\ &= \exp\{-\Lambda(A_z)\} \end{aligned} \tag{3.96}$$

We know that $\Lambda(A_z)$ is an exponent measure on A_z which has homogeneity of order-"-1". The homogeneity property of the exponent measure Λ yeilds a versatile representation in terms of psuedo-polar coordinates. Roughly speaking, One can transform the exponent measure $\Lambda(.)$ of cartesian space on A_z, to a psuedo-polar measure, $H(.)$, on A_z. To define the radial and angular parts of the new psuedo-polar one can use any arbitrary norm p where $1 \leq p < \infty$ or max-norm, i.g. ∞-norm. Usually exploiting two norms $||.||_1$ and $||.||_2$ are more common. Here, we barely use $||.||_1$ on $\mathbb{R}^2$.

Now let $\mathbb{S}_1 = \{\omega \in \mathbb{R}^2 : ||\omega||_1 = 1\}$ be the unit circle with respect to the norm $||\omega||_1$. Define the mapping T from $\mathbb{R}^2 \setminus \mathbf{0}$ to $(0,\infty) \times \mathbb{S}_1$ by

$$T(\mathbf{z}) = (r, \omega) \qquad \text{where} \quad r = ||\mathbf{z}||_1 \quad \text{and} \quad \omega = \frac{\mathbf{z}}{||\mathbf{z}||_1} \tag{3.97}$$

r is the radial part and ω is the angular part of z. For a bivariate case and $\mathbf{z} \geq \mathbf{0}$, trivially we have:

$$T(z_1, z_2) = (r, \omega) \qquad \text{where} \quad r = ||\mathbf{z}||_1 = z_1 + z_2 \quad \text{and} \quad \omega = \frac{\mathbf{z}}{||\mathbf{z}||_1} = \frac{(z_1, z_2)}{z_1 + z_2} \tag{3.98}$$

T is a one-to-one mapping, so $T(\mathbf{z}) = (r, \omega)$ if and only if $z = \dfrac{r\omega}{||\omega||_1} = T^{-1}(r, \omega)$;or equivalently Now define a measure $H(.)$ on $\Delta = \mathbb{S}_1 \cap [0, \infty)$ by

$$H(A_z) = \Lambda(\{\mathbf{z} \in [\mathbf{0}, \infty) : ||\mathbf{z}||_1 > 1, \mathbf{z}/||\mathbf{z}||_1 \in A_z\}) \tag{3.99}$$

for Borel subset A_z of Δ. The measure H is called the spectral measure. It is determined uniquely by the exponent measure Λ. We know that Λ has homogeneity of order-"-1". So the homogeneity implies that

$$\Lambda(\{\mathbf{z} \in [\mathbf{0}, \infty) : \|\mathbf{z}\|_1 > r, \mathbf{z}/\|\mathbf{z}\|_1 \in A_z\}) = r^{-1} H(A_z) \tag{3.100}$$

where $0 < r < \infty$ and A_z is a borel subset of Δ.

Intuitively, with polar coordinate (r, ω) we factorize the measure into a product of two measures, one in the radial coordinate that is **always** equal to $r^{-2}dr$, and one in the angular coordinate, equal to the spectral measure H. This property is usually written as:

$$\Lambda(T^{-1}(dr, d\omega) = r^{-2} dr H(d\omega)) \tag{3.101}$$

which is called the spectral decomposition of the exponent measure.

Given that, we return to the initial problem of probability of extreme events occurrence in a defined space A_z,

$$A_\mathbf{z} = \{(0, \infty) \times (0, \infty) \setminus (0, z_1) \times (0, z_2)\} \tag{3.102}$$

and

$$\mathbb{P}(M^*_{x,n} \leq z_1,\ M^*_{y,n} \leq z_2) = \mathbb{P}(N_n(A_z) = 0) \tag{3.103}$$

Hence by the Poisson Point Process limit

$$\begin{aligned} \mathbb{P}(M^*_{x,n} \leq z_1,\ M^*_{y,n} \leq z_2) &\rightarrow \mathbb{P}\{N(A_z) = 0\} \\ &= \exp\{-\Lambda(A_z)\} \end{aligned} \tag{3.104}$$

So with using norm $\|\mathbf{z}\|_1$ for radial and angular part one can define,

$$r = z_1 + z_2 \tag{3.105}$$

$$\omega = \frac{z_1}{z_1 + z_2} \tag{3.106}$$

Remark. It is important to notice that since in general case we define a d- dimensional space with $r \in \mathbb{R}$ and $\omega \in \mathbb{R}^d$ one can say knowing r, we only need $d-1$ dimensions of ω to specify any vector. For instance any vector from origin in a $2d$ space is specified by a radial component and an angular component. Any angle defined by other coordinate does not provide new inforamtion.

So, here we define ω as $\omega = \dfrac{z_1}{z_1 + z_2}$

$$r = \frac{z_1}{\omega} \tag{3.107}$$

or

$$r = \frac{z_2}{1 - \omega} \tag{3.108}$$

and

$$\lambda(r, \omega) = 2\frac{dH(\omega)}{r^2} \tag{3.109}$$

Then one can obtain:

$$\begin{aligned} \Lambda(A_z) &= \int_{A_z} 2\frac{dH(\omega)}{r^2} \\ &= \int_{\omega=0}^{1} \int_{r=\min(\frac{z_1}{\omega}, \frac{z_2}{1-\omega})}^{\infty} 2\frac{dr}{r^2} dH(\omega) \\ &= 2\int_0^1 \max(\frac{\omega}{z_1}, \frac{1-\omega}{z_2}) dH(\omega) \end{aligned} \tag{3.110}$$

which is totally compatible with the exponent measure defined for a bivariate max-stable distribution in Equation(3.44).

As mentioned beforehand, using the notion of ”extreme” region in Poisson Point Process we have better intuition about connection of Block Maxima and Threshold Exceedance approaches. In parallel to the ”extreme” region defined for Maxima, one can define ”extreme” region using threshold exceedance. In this case the ”extreme” region is an area where at least one of coordinates exceeds over a high threshold. Moreover, we can extend this idea to situations where we are using spectral measure. To do so, the ”extreme” region from threshold exceedance approach and from spectral point of view is an area where radial term exceeds over a high threshold.

$$r > u \tag{3.111}$$

where r is defined by any arbitrary norm. Figure(?) shows two approaches of threshold exceedance in both exponent measure and spectral measure methods are intuitively related to eachother.

FIGURE for comaprision of different norms In spectral measure case

Figure(?) illustrates that one can see the common threshold exceedance approach ,which define ”extreme” region as an area where at least one of coordinates exceeds over a high threshold, is a special case where we use norm-max or norm-∞ for r.

One can observe that the lower norm we use for r the higher number of points considered as ”extremes”. Hence, the with smaller norm for r the tail estimator has larger bias and smaller variance.

The transformation from exponent measure to spectral measure, and respectively from Cartesian to psuedo-polar coordiantes provides elegant tool for dependence structure. It is easier to show this interpretation for bivariate case where

$$\lambda(r,\omega) = 2\frac{dH(\omega)}{r^2} \tag{3.112}$$

It implies that the intensity of the limiting process N factorizes across radial and angular components. In other words, the angular spread of points in Poisson Point Process is determined by angular measure H and is independent of radial distance. Suppose that angular measure H is differentiable with differentiation h. Since $\omega = \frac{z_1}{z_1 + z_2}$, one can say ω is the relative size of z_1 and z_2. Then h measures the relative frequency of events of different relative size. When extremes are close to independence one expects that observations with large values in one coordiante occure with small values on the other coordiante. Therefore, in this case $h(\omega)$ is large in two situations:

1. When ω is close to 0 showing that extreme observations happened in X coordiante; however, small observation occurred in Y coordinate.($z_1 \gg z_2$)

2. When ω is close to 1 showing that extreme observations happened in Y coordiante; however, small observation occurred in X coordinate. ($z_2 \gg z_1$)

In the contrary, if X and Y are extremely dependent to each other, we expect that to observe simultaneous extremes in both coordiantes. So for the standardized version of X and Y one expects to see that they are likely to have similar extreme values which results in high $h(\omega)$ where $\omega = \frac{1}{2}$. ($z_1 \simeq z_2$)

FIGURE

We can extend this idea to the case where we use other norms for definition of ω. For instance if one defines ω using norm-2, $\omega = \frac{z_1}{\sqrt{z_1^2 + z_2^2}}$, then the high frequency of $h(\omega)$ in

$\omega = 0$ and $\omega = \frac{\pi}{2}$ correspond to independence, and the high frequency of $h(\omega)$ in $\omega = \frac{\pi}{4}$ and correspond to dependence.

FIGURE

In two extreme situations ”complete dependence” and ”complete independence” the measure function H is not differentiable. For the fully independent case the entire mass lies in subfaces, and for fully dependent case with ω defined in norm-1 the entire mass lies on the direction where $\omega = \frac{1}{2}$. The later one for ω defined in norm-2 is the direction where $\omega = \frac{\pi}{4}$.

FIGURE

In conclusion, the dependence of extreme variables only specified by the angular measure.

To sum up, using Poisson Point Process approach we are able to justify and link other approaches with the notion of ”extreme” region.

TABLE: comparision of different approaches

3.2 Graph Theory

In this section, we introduce some topics in Graph Theory which are providing fundamental building blocks for defining graphical models for extremes and our contribution on Structure learning for extremal forest models.

3.2.1 Definitions

We introduce some basic definitions in graph theory which is required for next sections.

Definition 12. Directed graph: Let V be a finite and non-empty set and $E \subset V \times V$ then, the pair $\mathcal{G} = (V, E)$ is called a directed graph on V. where V is the set of vertices or nodes and E is the set of directed edges.

Definition 13. Undirected graph: Let V be a finite and non-empty set and E is a set of unordered pairs of vertices in V, then, the pair $\mathcal{G} = (V, E)$ is called an undirected graph on V. where V is the set of vertices or nodes and E is the set of undirected edges.

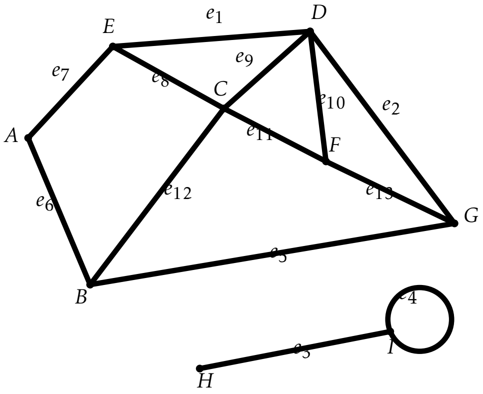


Figure 3.1: Undirected graph

Remark. Let $i, j \in V$. any pair $\{i, j\} \in E$ is an unordered pair if both ordered pairs (i, j) and (j, i) are in E.

In general, if a graph $\mathcal{G}$ is not specified as directed or undirected graph it is assumed to be undirected. From visual aspect, a graph $\mathcal{G} = (V, E)$ is a visual object where dots represent vertices A line joining i and j represents an undirected edge $\{i, j\} \in E$, whereas an arrow from i to j represents a directed edge $(i, j) \in E$.

Definition 14. Subgraph: Let $\mathcal{G} = (V, E)$ then a subgraph $\mathcal{S}$ is a graph $\mathcal{S} = (V_{\mathcal{S}}, E_{\mathcal{S}})$ where $V_{\mathcal{S}} \subset V$ and $E_{\mathcal{S}} \subset E$.

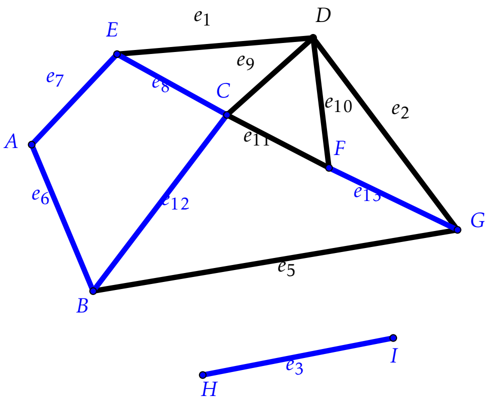


Figure 3.2: Subgraph

Definition 15. Spanning subgraph: Let $\mathcal{G} = (V, E)$ then a spanning subgraph $\mathcal{S}$ is a graph $\mathcal{S} = (V_{\mathcal{S}}, E_{\mathcal{S}})$ where $V_{\mathcal{S}} = V$ and $E_{\mathcal{S}} \subset E$.

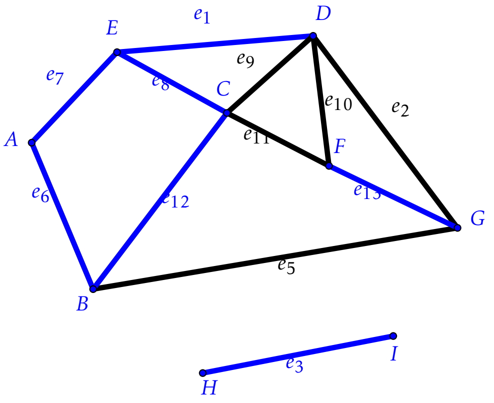


Figure 3.3: Spanning subgraph

Definition 16. Incidence: Let $\mathcal{G} = (V, E)$, then for any edge $(i, j) \in E$ the edge is incident with the vertices i and j.

Definition 17. Adjacent vertices: Let $\mathcal{G} = (V, E)$, then for any directed edge (i, j), i is adjacent to j and j is adjacent from i. we denote it by $i \to j$. Moreover, for any undirected edge $\{i, j\}$, i and j are adjacent or neighbor vertices. we denote it by $i \sim j$

For any directed edge (i, j), i is the source or origin of the edge and j is the terminatory vertex.

Definition 18. Parent and Child: Let $\mathcal{G} = (V, E)$ If there is a directed edge from i to j then i is said to be a parent j and j be a child of i. The set of parents of j is denoted by $pa(j)$ and the set of children of j is denoted by $chi(j)$

Definition 19. Isolated vertex: The vertex i has no incident edges is called isolated vertex.

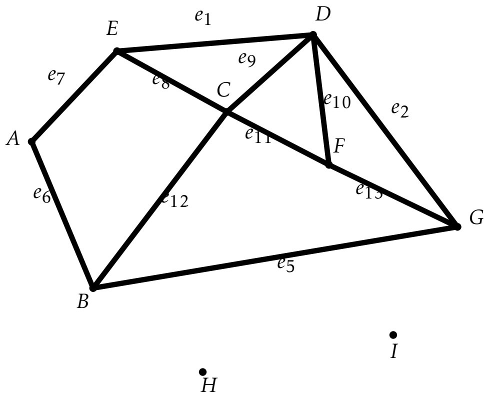


Figure 3.4: Isolated vertex

Definition 20. Path: Let $\mathcal{G} = (V, E)$ and $i, j \in V$. Then a path of length n from i to j is a sequence $s_1 = i$, ..., $s_n = j$ of distinct vertices such that $(s_k, s_{k+1}) \in E$ for all $k = 1$, ..., $n-1$. If there is a path from i to j we say that i leads to j and denote it by $i \mapsto j$.

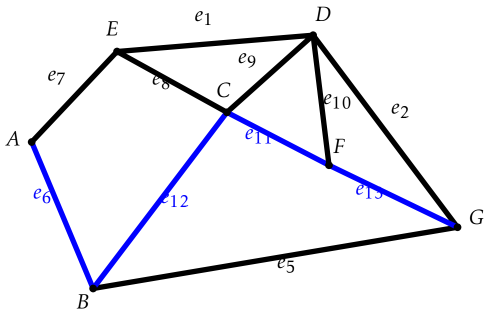


Figure 3.5: Path

Definition 21. Connected nodes Let $\mathcal{G} = (V, E)$ and $i, j \in V$. If both i and j leads to each other, i.e. $i \mapsto j$ and $j \mapsto i$ then we say that i and j are connected and it is denoted by $i \rightleftharpoons j$.

Connectivity ”$\rightleftharpoons$” is an equivalence relation. Connectivity component of $\mathcal{G} = (V, E)$, $\mathcal{C}(i)$, is the equivalence class $[i]_V$ where $j \in [i]_V$ is equivalent to $i \rightleftharpoons j$.

Definition 22. Connected graph: A connected graph is a graph where all vertices are in an equivalence class, i.e. for all $i, j \in V$ is $i \rightleftharpoons j$.

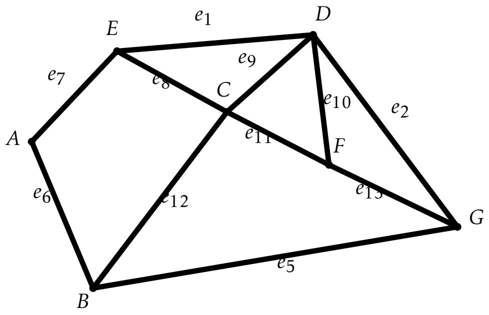


Figure 3.6: Connected graph

Definition 23. Chain: A chain of length a from i to j is a sequence $s_0 = i, ..., s_n = j$ of distinct vertices such that $s_{k-1} \rightarrow s_k$ or $s_k \rightarrow s_{k-1}$ for all $k = 1, ..., n$

Definition 24. Loop: A Loop is a path from a node i to itself.

Definition 25. Loop-free graph: A graph which does not contain any loop is a loop free graph.

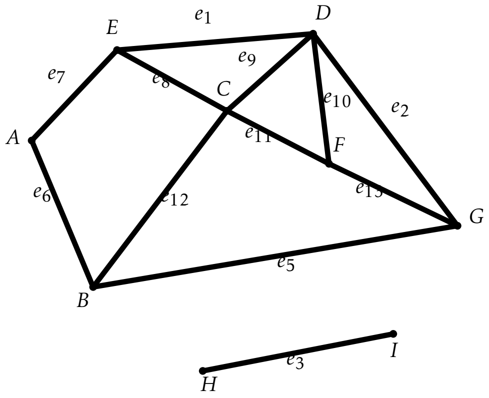


Figure 3.7: Undirected loop-free graph

Definition 26. Cycle: An n–cycle, C, is a path of length n where $s_0 = s_n$ means that the path begins and ends in a same node.

Remark. A cycle is directed if it has a directed edge.

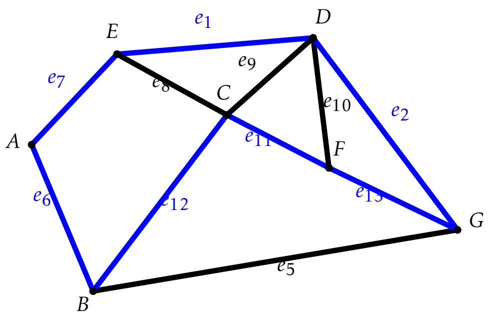


Figure 3.8: Cycle

Definition 27. Acyclic graph: A graph is acyclic when it has no cycle.

Definition 28. Tree: An undirected, connected, acyclic, and loop-free graph is a tree, i.e. there is a unique path between any two vertices in a tree.

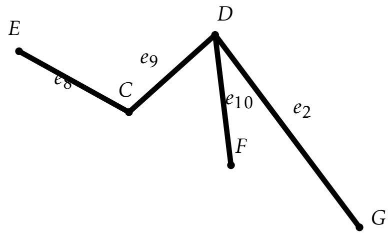


Figure 3.9: Tree

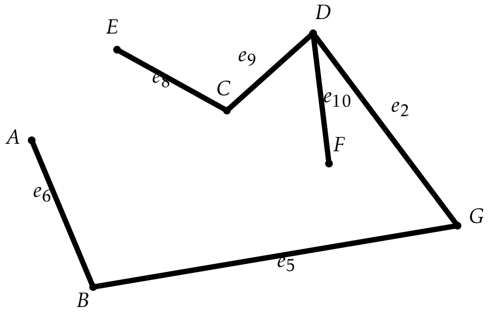


Figure 3.10: Spanning Tree

Definition 29. Rooted tree: A directed, connected, acyclic and loop-free graph is a rooted tree.

Definition 30. Forest: An undirected acyclic graph, all of whose connected components are trees, i.e. any two vertices are connected by at most one path. Roughly speaking, the graph consists of a disjoint union of trees.

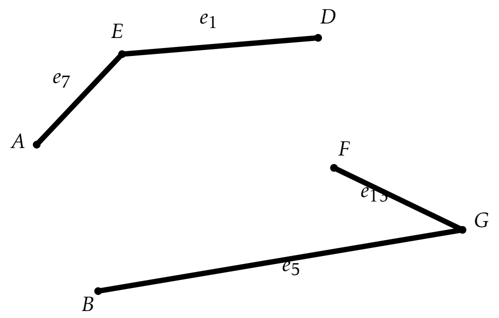


Figure 3.11: Forest

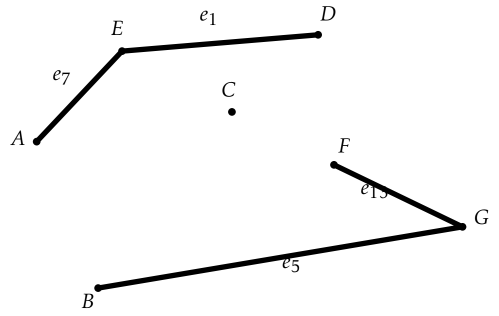


Figure 3.12: spanningForest

Definition 31. Complete graph: A graph $\mathcal{G} = (V, E)$ is complete if all vertices are adjacent, i.e. $E = V \times V$.

Definition 32. Complete subgraph: A subgraph $\mathcal{G}_A = (V_A, E_A)$ where all vertices in A are adjacent.

Definition 33. Clique: A complete subgraph that is maximal with respect to "$\subset$" is a clique, i.e. a clique $\mathfrak{C}(A, E_A)$ is a complete subgraph of graph $\mathcal{G} = (V, E)$ where if we add vertice i, $i \in V \setminus A$ and all edges $(j, i) \in E$ where $j \in A$ then the new subgraph is not complete. we denote the set of Cliques for graph $\mathcal{G}$ by $\mathfrak{C}(\mathcal{G})$.

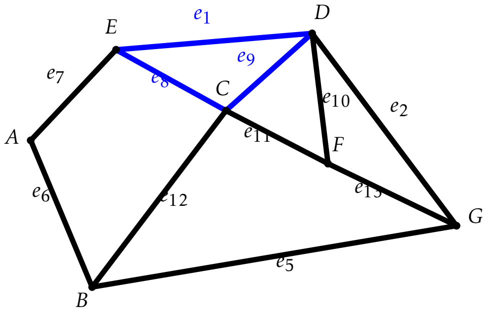


Figure 3.13: Clique

Definition 34. Chord: Let C be a cycle in graph $\mathcal{G} = (V, E)$, then an edge $e \in E$ which is not in the cycle C but connects two vertices of the cycle C is a chord.

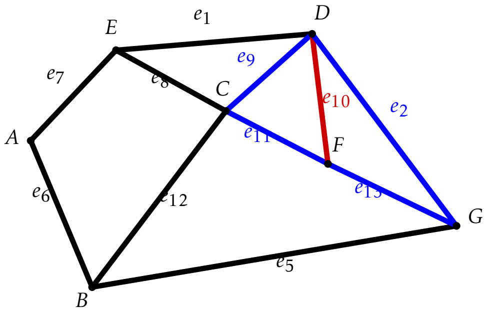


Figure 3.14: Chord

Definition 35. Decomposable graph: A graph in which all $n-$cycles where $n \geq 4$ have a chord is a decomposable graph.

Definition 36. Separator: Let $\mathcal{G} = (V, E)$. A subset $S \subset V$ is said to be An $(i, j)-$separator if all paths from i to j intersects S. Thus, in an undirected graph, S is an $(i, j)-$separator if and only if

$$[i]_{V \backslash S} \neq [j]_{V \backslash S}$$

the subset S is said to separate the subset of A from B where A, $B \subset V$ if it is an $(i, j)-$separator for all $i \in A$, $j \in B$

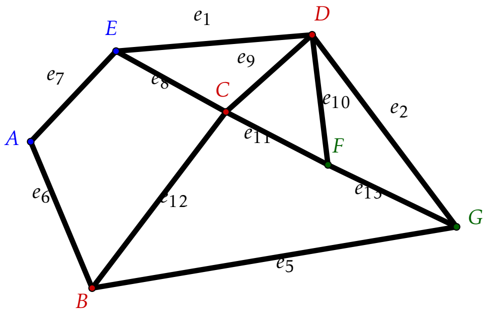


Figure 3.15: Separator

Definition 37. Block graph: A connected and decomposable graph $\mathcal{G} = (V, E)$ with cliques set $\mathfrak{C}(\mathcal{G})$ and separators set $\mathfrak{D}(\mathcal{G})$ where all separators in $\mathfrak{D}$ are single vertices (singleton separators) are known as block graph.

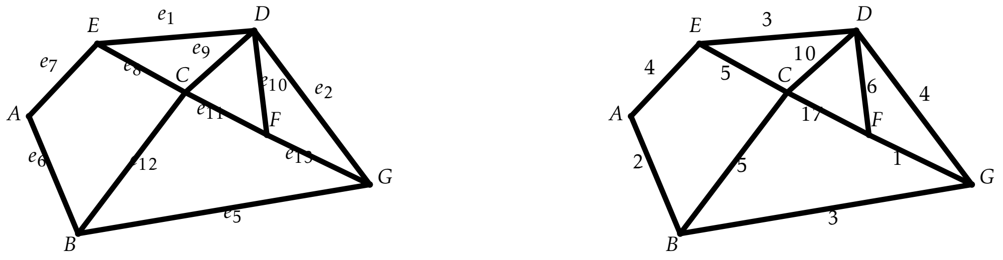
3.2.2 Optimization problems on graphs

Sometimes we can translate a mathematical problem to an optimization problem on graphs. In these problems, usually we have a weighted graph $\mathcal{G} = (V, E)$ where for any $e \in E$, $w(e)$ is a function $w : E \longrightarrow \mathbb{R}^+ \cup \{0, \infty\}$. The attributed value to each edge correspond to some information about that edge. Then optimize, i.e. minimize or maximize, an objective function over these weights with respect to some constraints. To solve these optimization problems we develope algorithmic manner to facilitate the implementation of solutions on computers.

In this contribution and as one can find in the following chapters, we link the minimal dependence structure between extreme variables with an optimiztion problem. So we minimize an objective function (i.e. overal sum of some weights corresponding to extremal dependence) by selecting a subgraph representing the minimal dependence structure with a constraint on the type of selected subgraph.

3.2.3 Minimum Spanning Tree

We have many options for the type of spanning subgraph which minimally represent the dependence structure. To be parsimonious, for the moment, we use ”trees” having a simple structure. So now, we would like to find a spanning subgraph which is a tree such that sum of the edges’ weights in the tree is minimal. These optimal spanning trees can be found by using the Kruskal’s and Prim’s algorithms. These algorithms are greedy where at each step of the algorithms a minimal choice is made from the remaining available data.



(a) Graph $\mathcal{G}$

(b) Weighted graph $\mathcal{G}$

Figure 3.16: Optimization Problem Minimum Spanning Tree

Minimum Spanning Tree Problem: Let graph $\mathcal{G} = (V, E)$ where $|V| = n$ be a connected and loop-free and each edge $e \in E$ is associated with a positive real number $w(e)$. Then **minimize** the objective function, overal sum of weights on the selected subgraph $\mathcal{T}$ of $\mathcal{G}$ with respect to the **constraint** that $\mathcal{T}$ is a connected and acyclic subgraph, i.e. tree.

3.2.4 Kruskal's algorithm

1. Let $\mathcal{T} = (V_{\mathcal{T}}, E_{\mathcal{T}})$ where $\mathcal{T}$ is a spanning subgraph for $\mathcal{G}$, i.e. $V_{\mathcal{T}} = V$ and $E_{\mathcal{T}} = \emptyset$ Set counter $i = 1$. Select an edge e_1 in $\mathcal{G}$ which has smallest weight $w(e_1)$, and add it to $E_{\mathcal{T}}$.

2. For $1 \leq i \leq n-2$, if edges e_1, e_2, ..., e_i have been selected and added to $E_{\mathcal{T}}$; then select edge e_{i+1} from $E \setminus \{e_1, e_2, ..., e_i\}$ the remaining edges in $\mathcal{G}$ such that e_{i+1} satisfies the following conditions and add it to $E_{\mathcal{T}}$.

   - $w(e_{i+1})$ is the smallest possible weight among the weights attributed to the remaining edges.
   - The subgraph of $\mathcal{G}$ determined by edges $\{e_1, e_2, ..., e_{i+1}\}$ and the incident vertices is acyclic.

3. Replece i by $i + 1$,
   If $i = n - 1$, the subgraph of $\mathcal{G}$ determined by edges e_1, e_2, ..., e_{n-1} is acyclic graph with n vertices and $n - 1$ edges so it is connected and the subgraph $\mathcal{T}$ is a minimal spanning tree for $\mathcal{G}$. If $i < n - 1$, return to step 2.

Example 3.

The resulting subgraph $\mathcal{T}$ has these properties:

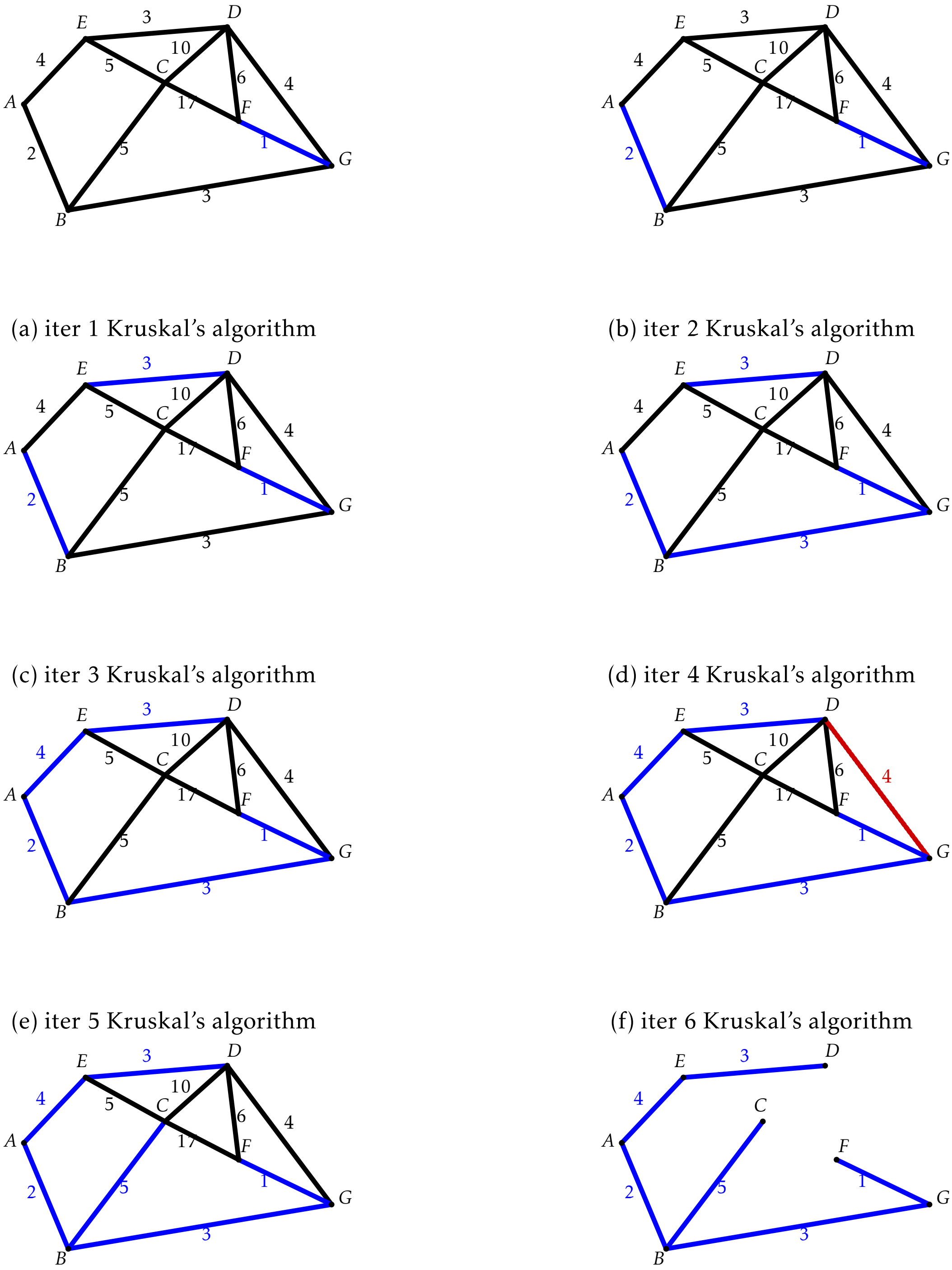
1. It has n vertices and $n - 1$ edges with respect to step 3 of Kruskal's algorithm.

2. It is **acyclic** with respect to part b in step 2 of the algorithm.

3. The resulting subgraph, $\mathcal{T}$, is **spanning subgraph of initial graph** $\mathcal{G}$, i.e. for if $v \in V$ and v is not in $\mathcal{T}$, then we can add an edge e of $\mathcal{G}$ to $\mathcal{T}$ where e is incident with v, and the resulting subgraph of $\mathcal{G}$ still contains no cycle.

4. $\mathcal{T}$ is connected. Otherwise suppose that $\mathcal{T}$ has two components $\mathcal{T}_1$ and $\mathcal{T}_2$. Since $\mathcal{G}$ is connected we could add to $\mathcal{T}$ and edge $\{i, j\}$ from $\mathcal{G}$ where i is in $\mathcal{T}_1$ and j is in $\mathcal{T}_2$ and no cycle would be present in this subgraph. Hence $\mathcal{T}$ is **connected**.

The properties 1 to 4 shows that the resulting subgraph $\mathcal{T}$ of $\mathcal{G}$ is a connected spanning subgraph of $\mathcal{G}$ with no cycles. Consequently, $\mathcal{T}$ is a spanning tree of $\mathcal{G}$.



(a) iter 1 Kruskal's algorithm

(b) iter 2 Kruskal's algorithm

(c) iter 3 Kruskal's algorithm

(d) iter 4 Kruskal's algorithm

(e) iter 5 Kruskal's algorithm

(f) iter 6 Kruskal's algorithm

(g) iter 7 Kruskal's algorithm

(h) Final Minimum Spanning Tree

Figure 3.17: Kruskal's algorithm!

Kruskal's algorithm optimality

One can also show that the greedy Kruskal's algorithm is an optimal solution.

Theorem 6. *Let $\mathcal{G} = (V, E)$ be a loop-free, weighted, connected, and undirected graph. Any spanning tree $\mathcal{T}$ for $\mathcal{G}$ obtained by Kruskal's algorithm is optimal.*

Proof. Suppose that $|V| = n$. Let $\mathcal{T}$ be a spanning tree for $\mathcal{G}$ obtained by Kruskal's algorithm. Clearly, it has $n-1$ edges and its edges are labeled by e_1, e_2, ..., e_{n-1} where e_i is i^{th} edge which is added to the graph by Kruskal's algorithm.
Suppose that $\mathcal{T}'$ is the optimal tree of $\mathcal{G}$. define $d(\mathcal{T}') = k$ as the number of common edges of $\mathcal{T}$ and $\mathcal{T}'$ before that Kruskal's algorithm add a non-optimal edge to $\mathcal{T}$.
so $d(\mathcal{T}') = k$ if k is the smallest positive integer that $\mathcal{T}$ and $\mathcal{T}'$ both contain e_1, e_2, ..., e_{k-1} but $e_k \notin \mathcal{T}'$. Let $\mathcal{T}_1$ be an optimal tree for which $d(\mathcal{T}_1) = r$ is **maximal** if $r = n$ then $\mathcal{T} = \mathcal{T}_1$ and so the tree $\mathcal{T}$ obtained by Kruskal's algorithm is optimal.
Otherwise, if $r \leq n-1$, adding e_r of $\mathcal{T}$ to $\mathcal{T}_1$ produces a cycle C. Since there exists another edge e_{r_1} of $\mathcal{T}_1$ which is not in $\mathcal{T}$ but makes the cycle C.
Start with tree $\mathcal{T}_1$. add e_r to $\mathcal{T}_1$ and delete e_{r_1}. so the resulting graph is a connected acyclic graph with n vertices $n-1$ edges.
So this graph is a spanning tree $\mathcal{T}_2$ where its weights satisfies

$$w(\mathcal{T}_2) = w(\mathcal{T}_1) + w(e_r) - w(e_{r_1})$$

In the Kruskal's algorithm, and in the $r-1$ first iterations we have made a subgraph $\mathcal{H}$ of $\mathcal{G}$ with edges e_1, e_2, ..., e_{r-1}. Given that $d(\mathcal{T}_1) = r$ then the edge e_r is chosen so that $w(e_r)$ is minimal where no cycle results when e_r is added to subgraph $\mathcal{H}$.
But since e_{r_1} also produces no cycle when added to the subgraph $\mathcal{H}$, by minimality of $w(e_r)$ it follows that $w(e_{r_1}) \geq w(e_r)$. So $w(e_r) - w(e_{r_1}) \leq 0$ and trivially, $w(\mathcal{T}_2) \leq w(\mathcal{T}_1)$. But we know that $\mathcal{T}_1$ is optimal. So we must have $w(\mathcal{T}_2) = w(\mathcal{T}_1)$. Therefore, $\mathcal{T}_2$ is also optimal. It has edges e_1, ..., e_{r-1}, e_r in common with $\mathcal{T}$. So $d(\mathcal{T}_2) = r+1 > r = d(\mathcal{T}_1)$. It is contradicting with the selection of $\mathcal{T}_1$ (Since $\mathcal{T}_1$ has highest number of common edges with $\mathcal{T}$ among all optimal trees).
Therefore, $r = n$ and $\mathcal{T}_1 = \mathcal{T}$ so the tree $\mathcal{T}$ obtained by Kruskal's algorithm is optimal.

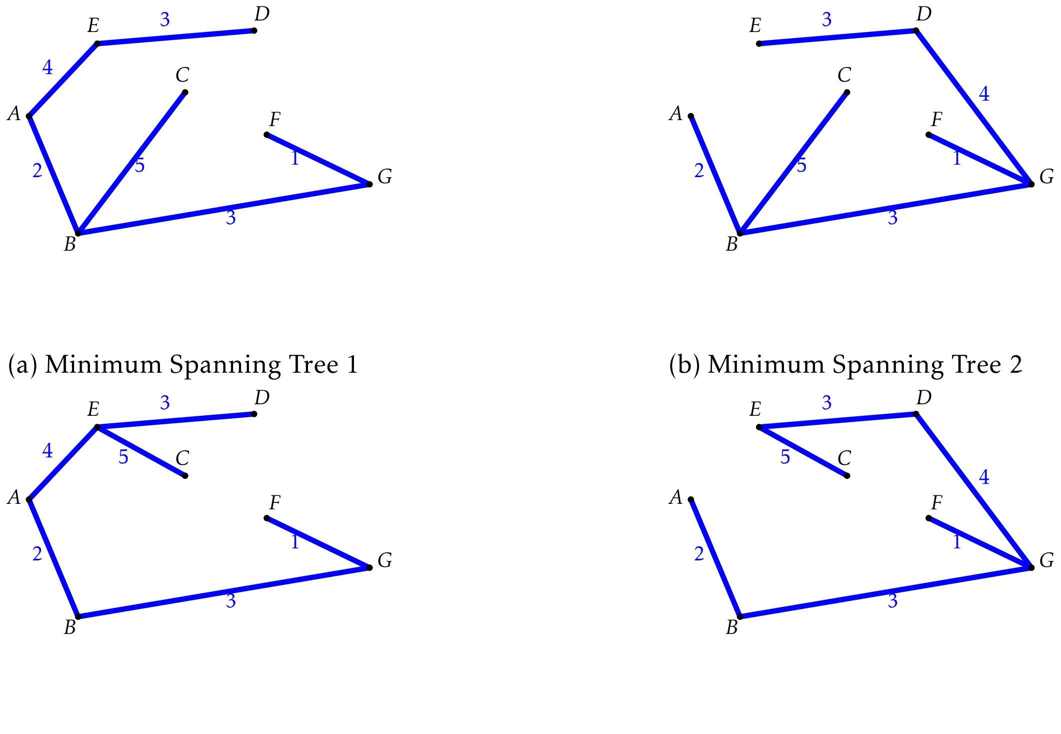
Kruskal's algorithm uniqueness

We should find the necessary conditions required for uniqueness of Kruskal's algorithms' solution. Since, it is proven that Kruskal's algorithm is providing optimal tree. So, the required conditions for uniqueness of Kruskal's algorithm are same as required condition for uniqueness of minimum spanning tree. Before, the discussion about the necessary conditions for uniqueness of minimum spanning tree we define some new terms.

Example 4.

Definition 38. Adjacent spanning trees: two spanning trees are adjacent if each of these spanning tree has exactly one edge that is not in the other spanning tree.

Definition 39. Isolated minimum spanning tree: A minimum spanning tree is an isolated minimum spanning tree if it is not adjacent to another minimum spanning tree.



(a) Minimum Spanning Tree 1

(b) Minimum Spanning Tree 2

(c) Minimum Spanning Tree 3

(d) Minimum Spanning Tree 4

Figure 3.18: Several solutions of Kruskal's algorithm!

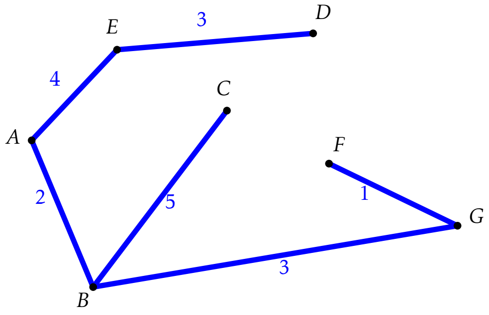


Figure 3.19: Isolated minimum spanning tree

Definition 40. Unique-cycle-heaviest edge: An edge is unique-cycle-heaviest if it is the unique heaviest edge in some cycle.

Definition 41. Non-cycle-heaviest edge: An edge is non-cycle-heaviest if it is never a heaviest edge in any cycle.

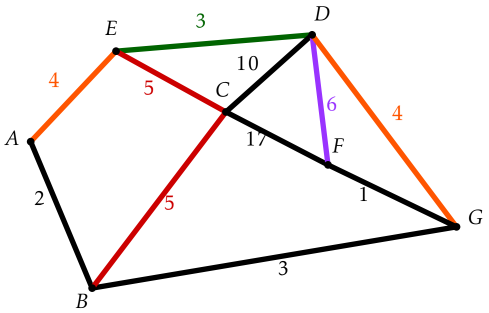


Figure 3.20: Unique-cycle-heaviest and Non-cycle-heaviest edges

Lemma 7. *These properties are equivalent:*

- ***Uniqueness of Minimum Spanning Tree:*** *There is an unique minimum spanning tree.*
- ***One local minimum spanning tree:*** *There is an spanning tree which is lighter than all adjacent spanning trees.*

Proof. ($\Rightarrow$) Let $\mathcal{G} = (V, E)$ and $\mathcal{T}$ is the unique minimum spanning tree for $\mathcal{G}$ then since it is unique, it does not have any adjacent minimum spanning tree. Trivially $\mathcal{T}$ is an isolated minimum spanning tree and since an isolated minimum spanning tree is lighter than all adjacent spanning trees then $\mathcal{T}$ is satisfting one local minimum spanning tree.

($\Leftarrow$) Let $\mathcal{G} = (V, E)$ and $\mathcal{T} = (V_{\mathcal{T}}, E_{\mathcal{T}})$ is an spanning tree which is lighter than all adjacent spanning trees. Suppose that there is another arbitrary spanning tree $\mathcal{T}'$ for $\mathcal{G}$ and it has k edges labeled e'_1, e'_2, ..., e'_k. Consider an arbitrary edge from these k edges and name it e'. This edge is incident with verices (i, j). Since $\mathcal{T}$ is a spanning tree then there is a path, $p(i, j)$ from i to j, in $\mathcal{T}$. By removing one of the edges in path $p(i, j)$ and replacing it with e' we make an adjacent spanning tree $\mathcal{T}_a$ to $\mathcal{T}$. Since $\mathcal{T}$ is an spanning tree which is lighter than all adjacent spanning trees we have $w(\mathcal{T}) < w(\mathcal{T}_a)$. So $w(e') < w(e)$ for all $e \in p(i, j)$. So we can replace e' by one of $e \in p(i, j)$ and the resulted subgraph is lighter than previous one and it is spanning tree.

Since $\mathcal{T} = (V_{\mathcal{T}}, E_{\mathcal{T}})$ is a spanning tree, for the set of edges $\{e'_1, e'_2, ..., e'_k\}$ there is a set $\{e_1, e_2, ..., e_k\}$ in $E_{\mathcal{T}}$ for $\{e'_1, e'_2, ..., e'_k\}$ where $w(e_i) < w(e'_i)$ for $i = 1, ..., k$. One can make find this corresponding edges by doing these algorithm:

1. For every e'_1, e'_2, ..., e'_k find the corresponding alternative path which is in $\mathcal{T}$ and connects the incident vertices the edge. And label them by p_1, p_2, ..., p_k.
2. For each path p_i, c_i is the number of uncommon edges with other paths Sort c_is descending, $c^{(1)}$, $c^{(2)}$, ..., $c^{(q)}$
3. For $1 \leq i \leq k$ with the path $p^{(i)}$ corresponding to $c^{(i)}$ and where $p^{(i)} = p_j$. Now, find the least common edge of $p^{(i)}$ or equivalently p_j, that is not in $\mathcal{T}'$ and name it e_j. Then $w(e_j) < w(e'_j)$ since e_j is on path p_j.

This algorithmic replacement qaurantees that $w(e_j) < w(e'_j)$ for all $j = 1, ..., k$ So $w(\mathcal{T}) < w(\mathcal{T}')$ where $w(\mathcal{T}')$ is an arbitrary spanning tree of $\mathcal{G}$. So $\mathcal{T}$ is the unique minimum spanning tree of $\mathcal{G}$.

Theorem 8. *These properties are equivalent:*

- ***Uniqueness of Minimum Spanning Tree:*** *There is an unique minimum spanning tree.*

- ***Extreme cycle edge:*** *Every edge is either unique-cycle-heaviest or non-cycle-heaviest.*

Proof. (⇒) Using lemma(7) we know that the uniqueness of minimum spanning tree results in one local minimum spanning tree. Hence, we only need to prove that one local minimum spanning tree results in extreme cycle edge. So let $\mathcal{T}$ be an spanning tree that is lighter than all adjacent spanning trees. Then:

- Every edge in $\mathcal{T}$ must be non-cycle-heaviest.
  Proof. Let e be an edge in $\mathcal{T}$. If e does not belong to any cycle in $\mathcal{G}$, we are done. Now suppose e belongs to a cycle C in $\mathcal{G}$. If we remove e from $\mathcal{T}$, $\mathcal{T}$ will be split into two trees, which will be named $\mathcal{T}_1$ and $\mathcal{T}_2$. As a cycle that connects $\mathcal{T}_1$ and $\mathcal{T}_2$ with e, C must have another edge that connects $\mathcal{T}_1$ and $\mathcal{T}_2$. Name that edge e'. Let $\mathcal{T}'$ be the union of $\mathcal{T}_1$, $\mathcal{T}_2$, and e', which must be a spanning tree of $\mathcal{G}$ as well. Since $\mathcal{T}_1$ and $\mathcal{T}_2$ are adjacent, $\mathcal{T}$ is lighter than $\mathcal{T}'$. That means, e is lighter than e'. So e is non-cycle-heaviest.

- Every edge not in $\mathcal{T}$ must be unique-cycle-heaviest.
  Proof. Let e' be an edge not in $\mathcal{T}$. If we add e' to $\mathcal{T}$, we will create a cycle C. Let e be an edge in C that is not e'. Consider the spanning tree $\mathcal{T}'$ made from $\mathcal{T}$ with e replaced by e'. Since $\mathcal{T}$ and $\mathcal{T}'$ are adjacent, $\mathcal{T}$ is lighter than $\mathcal{T}'$. That means, e is lighter than e'. So e' is the unique heaviest edge in $\mathcal{C}$. That is, e' is unique-cycle-heaviest.

(⇐) Let $\mathcal{T}$ be a minimum spanning tree for graph $\mathcal{G} = (V,E)$, Let e be an arbitrary edge in $\mathcal{G}$.

1. If e is non-cycle-heaviest $\mathcal{T}$ must contain it.
   Proof. By contradiction, let e be a non-cycle-heaviest in $\mathcal{G}$ and $\mathcal{T}$ does not contain it. There are two situations for e:

   (a) If the edge e, incident with (i,j), **is not in any cycle.**
   Then e is the only path between i and j. Hence, if $\mathcal{T}$ does not contain e then i and j are not connected which is in contradiction with initial assumption that $\mathcal{T}$ is a spanning tree. So by contradiction, $\mathcal{T}$ must contain any edge e.

   (b) If the edge e, incident with (i,j), **is in some cycles** C_1, ..., C_q **in** $\mathcal{G}$.
   By definition, the e is not the heaviest edge in all of them. So for all cycle C_1, ..., C_q there is at least an edge which is heavier than e. Since $\mathcal{T}$ is a spanning tree; there is at least another path, p, between i and j in graph $\mathcal{T}$. The union of p and e makes one of the circles C_1, ..., C_q, namely C. Then there is at least one edge e' in the path p which is heaviest edge in the cycle C. So $w(e') > w(e)$. Consequently, if we replace e' by e the adjacent spanning tree is lighter, which is in contradiction with the initial assumption about minimality of $\mathcal{T}$.

   Therfore, if e is a non-cycle-heaviest edge in $\mathcal{G}$, then the minimum spanning tree $\mathcal{T}$ must contain it.

2. If edge e is unique-cycle-heaviest $\mathcal{T}$ cannot contain it.
   Proof. By contradiction, let $\mathcal{T}$ be a minimum spanning tree for graph $\mathcal{G}$ and let e be an unique-cycle-heaviest in $\mathcal{G}$ which is contained by $\mathcal{T}$. The edge e, incident with (i,j) is an unique-cycle-heaviest so it is the unique heaviest edge in some cycle of $\mathcal{G}$. So, there is at least another path, p, from i to j on $\mathcal{G}$. The union of p and e is a cycle C on graph $\mathcal{G}$ which entails e as the unique-heaviest-edge in the cycle. So there is an edge e' on that cycle which is trivially not on $\mathcal{T}$ (since $\mathcal{T}$ is a tree and acyclic, it does not have all edges of one cycle) and clearly, $w(e') < w(e)$. So if we replace e' by e the adjacent spanning tree is lighter than $\mathcal{T}$ which is in contradiction with minimality assumption over $\mathcal{T}$

Therefore, if we have the extreme cycle edge property then **every edge** is either unique-cycle-heaviest which is in the edge set of $\mathcal{T}$ or it is non-cycle-heaviest which is not in the edge set of $\mathcal{T}$. Then, $\mathcal{T}$ is uniquely specified by the set of non-cycle-heaviest edges.

Remark. If there is an edge e in $\mathcal{G} = (V,E)$ which is neither unique-cycle-heaviest nor non-cycle-heaviest then $\mathcal{G}$ does not have an **unique** minimum spanning tree. FIGURE

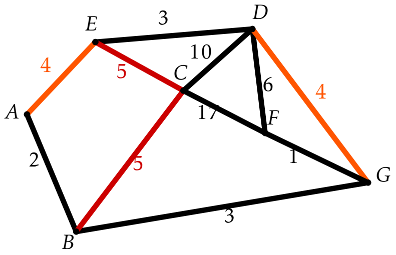


Figure 3.21: Edges which are violating necessary conditions

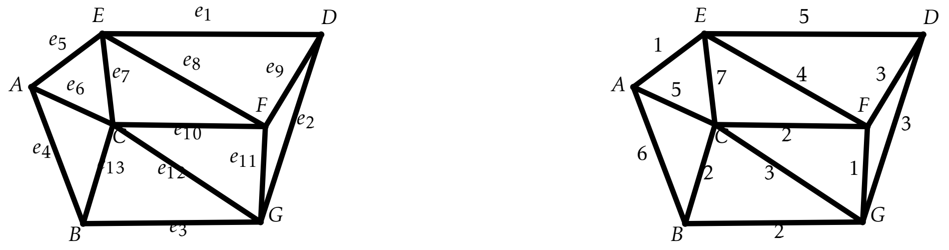
Since the Kruskal's algorithm is the optimal spanning tree, it requires same necessary and sufficient conditions for uniqueness. In parallel, if in the iteration r^{th} of the Kruskal's algorithm we have to choose between u equally weighted edges e_{r_1}, e_{r_2}, ..., e_{r_u} which makes a cycle with the edges e_1, e_2, ..., e_{r-1} that already have chosen by the algorithm then for being acyclic we cannot add all of e_{r_1}, e_{r_2}, ..., e_{r_u}. So we have to choose one or some of them and leave it out from the spanning tree $\mathcal{T}$. So we have different options for selection of this edge which is in contradiction with uniqueness of solution.

One can show these edges e_{r_1}, e_{r_2}, ..., e_{r_u} violate the extreme cycle edge property since they are neither unique-cycle-heaviest nor non-cycle-heaviest.

Although, it is not necessary condition for uniqueness of minimum spanning tree and the solution of Kruskal's algorithm. But it is sufficient condition for $\mathcal{G} = (V,E)$ to have uniquely weighted edges so automatically every edge is either unique-cycle-heaviest or non-cycle-heaviest.

3.2.5 Minimum Spanning Forest

Another option for the type of spanning subgraph which minimally represent the dependence structure is a forest structure. In the following, we show that using the forest structure we are able to represent the complete independence of extreme variables by graph. So now, we would like to find a spanning subgraph which is a forest such that sum of the edges' weights in the forest is minimal. These optimal spanning forests can be found by using the modified version of Kruskal's and Prim's algorithms. These algorithms are greedy where at each step of the algorithms a minimal choice is made from the remaining available data.



(a) Graph $\mathcal{G}$

(b) Weighted graph $\mathcal{G}$

Figure 3.22: Optimization Problem Minimum Spanning Forest

Minimum Spanning Forest Problem: Let graph $\mathcal{G} = (V, E)$ where $|V| = n$ be a connected and loop-free and each edge $e \in E$ is associated with a positive real number $w(e)$. Then **minimize** the objective function, overal sum of weights on the selected subgraph $\mathcal{F}$ of $\mathcal{G}$ with respect to the **constraint** that $\mathcal{F}$ is an acyclic subgraph, i.e. forest.

A discussion about the constraint: Since the minimal subgraph with the present condition can be a spanning subgraph without any edges then we need other restrictions to have a meaningful definition for minimum spanning forest. One can define a minimum spanning forest $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ of $\mathcal{G} = (V, E)$ with one of the following additional constraints.

1. **Size constraint:** $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ is a spanning forest for $\mathcal{G} = (V, E)$, where $|V| = |V_{\mathcal{F}}| = n$ and $|E_{\mathcal{F}}| = q$ and q can be any integer number in $\{0,\ 1, \ldots,\ n-1\}$

2. **Weight constraint:** $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ is a spanning forest for $\mathcal{G} = (V, E)$, where $|V| = |V_{\mathcal{F}}| = n$, for all $e \in E_{\mathcal{F}}$ and known τ, $w(e) < \tau$, and $|E_{\mathcal{F}}|$ is maximal.

In this contribution, we suggest two modified versions of Kruskal's algorithm to find minimum spanning forest. Each version corresponds to one of the later additional constraints.

3.2.6 Modified Kruskal's algorithms for minimum spanning forest

Modified Kruskal's algorithm with respect to size constraint $|E_{\mathcal{F}}| = q$:

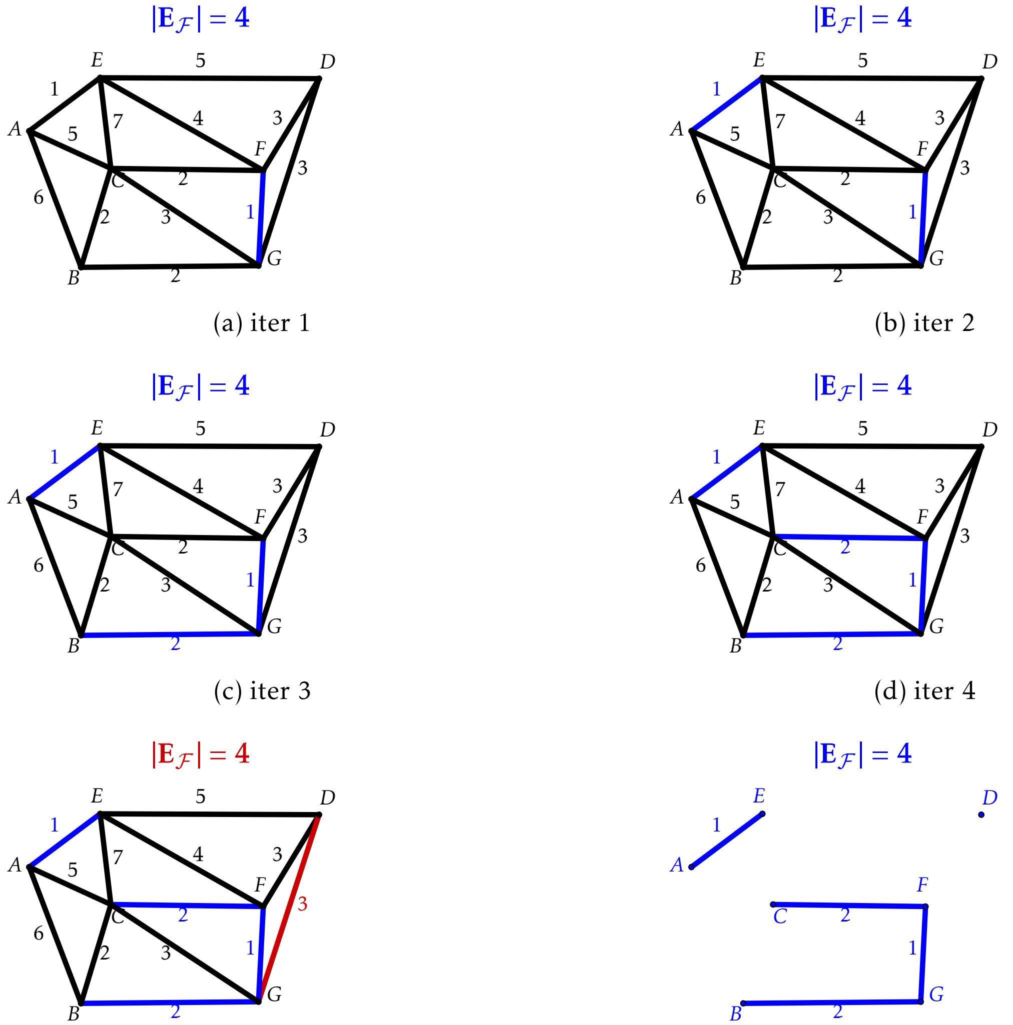
1. Let $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ where $\mathcal{F}$ is a spanning subgraph for $\mathcal{G}$, i.e. $V_{\mathcal{F}} = V$ and $E_{\mathcal{F}} = \emptyset$ Set counter $i = 1$. Select an edge e_1 in $\mathcal{G}$ which has smallest weight $w(e_1)$, and add it to $E_{\mathcal{F}}$.

2. For $1 \leq i \leq n-2$, if edges e_1, e_2, ..., e_i have been selected and added to $E_{\mathcal{F}}$; then select edge e_{i+1} from $E \setminus \{e_1,\ e_2,\ \ldots,\ e_i\}$ the remaining edges in $\mathcal{G}$ such that e_{i+1} satisfies the following conditions and add it to $E_{\mathcal{T}}$.

   - $w(e_{i+1})$ is the smallest possible weight among the weights attributed to the remaining edges.
   - The subgraph of $\mathcal{G}$ determined by edges $\{e_1,\ e_2,\ \ldots,\ e_{i+1}\}$ and the incident vertices is acyclic.

3. Replece i by $i + 1$,
   If $i = q$, the subgraph of $\mathcal{G}$ determined by edges e_1, e_2, ..., e_q is acyclic graph, $\mathcal{F}$,with n vertices and q edges so it is an optimal spanning forest for $\mathcal{G}$ If $i < q$, return to step 2.

Example 5.



(a) iter 1

(b) iter 2

(c) iter 3

(d) iter 4

(e) iter 5

(f) Final Minimum Spanning Forest

Figure 3.23: Modified Kruskal's algorithm with respect to size constraint $|E_{\mathcal{F}}| = 4$

Modified Kruskal's algorithm with respect to weight constraint $\forall e \in E_{\mathcal{F}} \quad w(e) < \tau$ **for a known** τ**:**

1. Let $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ where $\mathcal{F}$ is a spanning subgraph for $\mathcal{G}$, i.e. $V_{\mathcal{F}} = V$ and $E_{\mathcal{F}} = \emptyset$ Set counter $i = 1$. Select an edge e_1 in $\mathcal{G}$ which has smallest weight $w(e_1)$, and add it to $E_{\mathcal{F}}$.

2. For $1 \leq i \leq n-2$, if edges e_1, e_2, ..., e_i have been selected and added to $E_{\mathcal{F}}$; then select edge e_{i+1} from $E \setminus \{e_1, e_2, ..., e_i\}$ the remaining edges in $\mathcal{G}$ such that e_{i+1} satisfies the following conditions and add it to $E_{\mathcal{T}}$.

    - $w(e_{i+1})$ is the smallest possible weight among the weights attributed to the remaining edges.
    - The subgraph of $\mathcal{G}$ determined by edges $\{e_1, e_2, ..., e_{i+1}\}$ and the incident vertices is acyclic.
    - If $w(e_{i+1}) \geq \tau$ then do not add e_{i+1} to the subgraph $\mathcal{F}$ and stop the algorithm. The subgraph $\mathcal{F}$ of $\mathcal{G}$ determined by edges $\{e_1, e_2, ..., e_i\}$ is an acyclic graph with n vertices and i edges which are lighter than τ. So $\mathcal{F}$ is an optimal spanning forest for $\mathcal{G}$.
      If $w(e_{i+1}) < \tau$ then add e to the subgraph $\mathcal{F}$ and continue.

3. Replece i by $i+1$,
   If $i = n-1$, the subgraph of $\mathcal{G}$ determined by edges e_1, e_2, ..., e_{n-1} is acyclic graph, $\mathcal{F}$,with n vertices and q edges so it is an optimal spanning forest for $\mathcal{G}$ If $i < n-1$, return to step 2.

Example 6.

The resulting subgraph $\mathcal{F}$ has these properties:

1. It has n vertices and q edges where $q \in \{1, ..., n-1\}$
2. It is **acyclic** with respect to part b in step 2 of the algorithm.
3. The resulting subgraph, $\mathcal{F}$, is **spanning subgraph of initial graph** $\mathcal{G}$, i.e. we initialize with a spanning subgraph and add some edges.
4. $\mathcal{F}$ is satisfying one of two restrictions namely, size constraint or weight constraint.

The properties 1 to 4 shows that the resulting subgraph $\mathcal{F}$ of $\mathcal{G}$ is an acyclic spanning subgraph of $\mathcal{G}$ satisfying one of the additional constraints. Consequently, $\mathcal{F}$ is a spanning forest of $\mathcal{G}$.

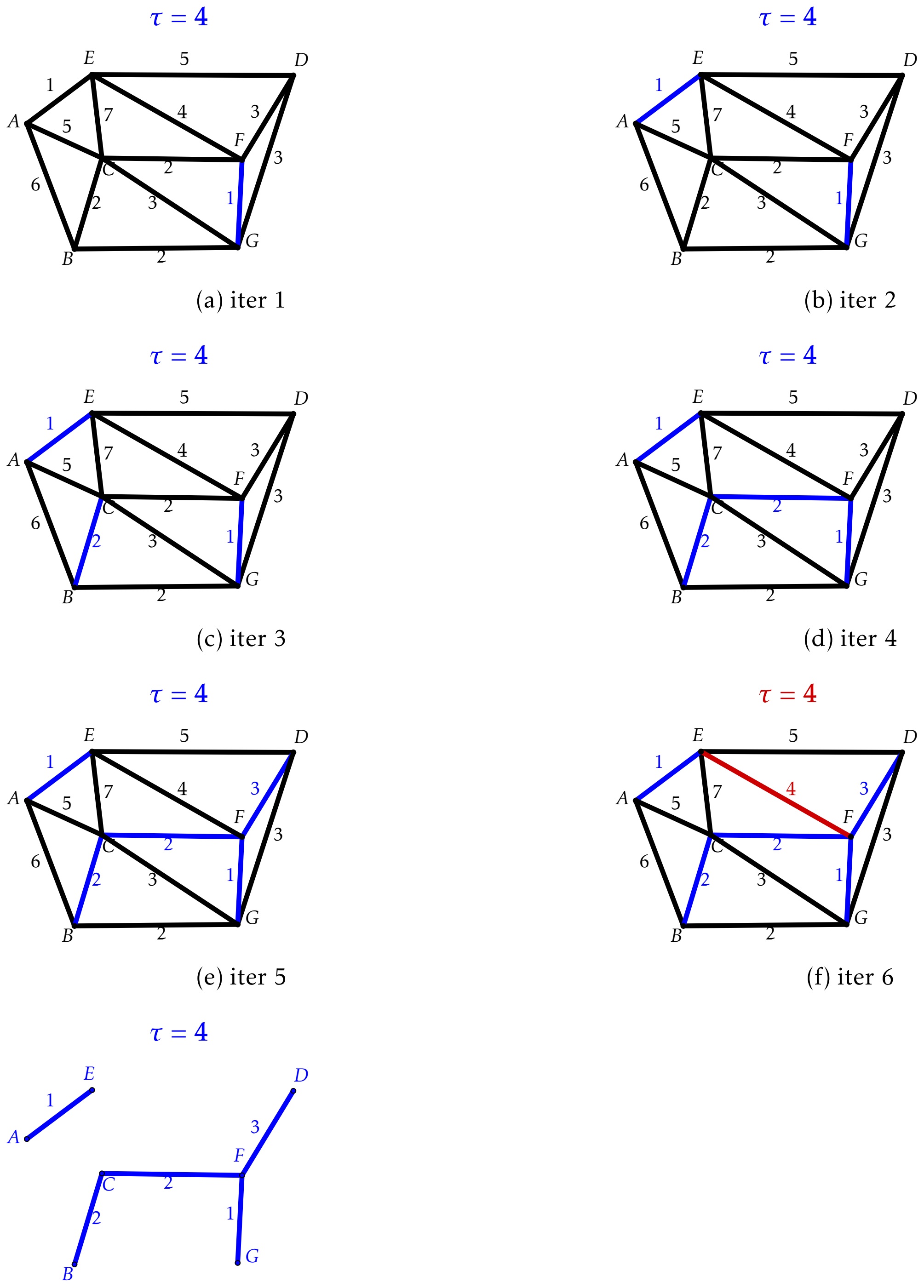
Modified Kruskal's algorithms optimality

One can show that the modified versions of the greedy Kruskal's algorithm are optimal.

Theorem 9. *Let $\mathcal{G} = (V, E)$ be a loop-free, weighted, connected, and undirected graph. Any spanning forest $\mathcal{F}$ for $\mathcal{G}$ obtained by modified versions of Kruskal's algorithm is optimal.*

Size restrcited version of minimum spanning forest:

Proof. Suppose that $|V| = n$. Let $\mathcal{F}$ be a spanning forest for $\mathcal{G} = (V, E)$ obtained by size restricted version of modified Kruskal's algorithm. Clearly, it has q edges and its edges are labeled by e_1, e_2, ..., e_q where e_i is i^{th} edge which is added to the graph by the algorithm.
Suppose that $\mathcal{F}'$ is the optimal forest of $\mathcal{G}$. define $d(\mathcal{F}') = k$ as the number of common edges of $\mathcal{F}$ and $\mathcal{F}'$ before that the algorithm add a non-optimal edge to $\mathcal{F}$. Obviously $k \leq q+1$.
So $d(\mathcal{F}') = k$ if k is the smallest positive integer that $\mathcal{F}$ and $\mathcal{F}'$ both contain e_1, e_2, ..., e_{k-1} but



(a) iter 1

(b) iter 2

(c) iter 3

(d) iter 4

(e) iter 5

(f) iter 6

(g) Final Minimum Spanning Forest

Figure 3.24: Modified Kruskal's algorithm with respect to weight constraint $\forall e \in E_{\mathcal{F}} \quad w(e) < \tau$ for $\tau = 4$

$e_k \notin \mathcal{F}'$. Let $\mathcal{F}_1$ be an optimal forest for which $d(\mathcal{F}_1) = r$ is **maximal** if $r = q+1$ then $\mathcal{F} = \mathcal{F}_1$ and so the forest $\mathcal{F}$ obtained by the algorithm is optimal.
Otherwise, if $r \leq q$, then we have e_{r_1} which is in $\mathcal{F}_1$ and not in $\mathcal{F}$ and e_r which is in $\mathcal{F}$ and not in $\mathcal{F}_1$. By adding e_r of $\mathcal{F}$ to $\mathcal{F}_1$ we face with one of these situations:

1. Adding e_r of $\mathcal{F}$ to $\mathcal{F}_1$ does not make a cycle.
2. Adding e_r of $\mathcal{F}$ to $\mathcal{F}_1$ makes a cycle, C.

For the first case, as we know, there exists another edge e_{r_1} of $\mathcal{F}_1$ which is not in $\mathcal{F}$. Start with forest $\mathcal{F}_1$ add e_r to $\mathcal{F}_1$ and delete e_{r_1}. $\mathcal{F}_1$ is acyclic and adding e_r of $\mathcal{F}$ to $\mathcal{F}_1$ does not make a new cycle, so adding e_r to $\mathcal{F}_1$ does not make a cyclic subgraph. So, if we replace e_{r_1} by e_r in $\mathcal{F}_1$ the resulting subgraph $\mathcal{F}_2$ remains acyclic and it has q edges.

For the second case, start with forest the $\mathcal{F}_1$ add e_r to $\mathcal{F}_1$ and name the created subgraph $\mathcal{F}_2$. As we know, there exists another edge e_{r_1} of $\mathcal{F}_1$ which is not in $\mathcal{F}$ and is in the cycle C. Since before this $\mathcal{F}_1$ was acyclic removing e_{r_1} from the newly created subgraph $\mathcal{F}_2$ makes $\mathcal{F}_2$ acyclic again. So the resulting subgraph $\mathcal{F}_2$ remains acyclic and it has q edges.

So the resulting graph is a spanning forest $\mathcal{F}_2$ where its weights satisfies

$$w(\mathcal{F}_2) = w(\mathcal{F}_1) + w(e_r) - w(e_{r_1})$$

In the algorithm, and in the $r-1$ first iterations we have made a subgraph $\mathcal{H}$ of $\mathcal{G}$ with edges e_1, e_2, ..., e_{r-1}. Given that $d(\mathcal{F}_1) = r$ then the edge e_r is chosen so that $w(e_r)$ is minimal where no cycle results when e_r is added to subgraph $\mathcal{H}$.
But since e_{r_1} also produces no cycle when added to the subgraph $\mathcal{H}$, by minimality of $w(e_r)$ it follows that $w(e_{r_1}) \geq w(e_r)$. So $w(e_r) - w(e_{r_1}) \leq 0$ and trivially, $w(\mathcal{F}_2) \leq w(\mathcal{F}_1)$. But we know that $\mathcal{F}_1$ is optimal. So we must have $w(\mathcal{F}_2) = w(\mathcal{F}_1)$. Therefore, $\mathcal{F}_2$ is also optimal. It has edges e_1, ..., e_{r-1}, e_r in common with $\mathcal{F}$. So $d(\mathcal{F}_2) = r+1 > r = d(\mathcal{F}_1)$. It is contradicting with the selection of $\mathcal{F}_1$ (Since $\mathcal{F}_1$ has highest number of common edges with $\mathcal{F}$ among all optimal forests). Therefore, $r = q+1$ and $\mathcal{F}_1 = \mathcal{F}$ so the forest $\mathcal{F}$ obtained by the size restricted version of the Kruskal's algorithm is optimal.

Weight restrcited version of minimum spanning forest:

Proof. Suppose that $|V| = n$. Let $\mathcal{F}$ be a spanning forest for $\mathcal{G} = (V, E)$ obtained by weight restricted version of modified Kruskal's algorithm. Suppose that $\mathcal{S}$ is a spanning subgraph of $\mathcal{G}$ which contains every edges e where $e \in E$ where $w(e) < \tau$ does not contain any edge $e \in E$ where $w(e) \geq \tau$.
We know that $\mathcal{G}$ is a connected graph, however, after removing the edges which are not lighter than τ from E, the remaining subgraph $\mathcal{S}$ has $\mathcal{C}_\#$ connectivity components where $\mathcal{C}_\# \in \{1, 2, ..., n\}$.

Regarding to weight constraint we know that any optimal forest $\mathcal{F}'$ must be a spanning subgraph of $\mathcal{S}$. Moreover, we have another term in weight where "$|E_\mathcal{F}|$ **is maximal.**" so the desired minimum spanning forest $\mathcal{F}'$ is an acyclic graph with $\mathcal{C}_\#$ connectivity components where $|E_{\mathcal{F}'}|$ is maximal. Therefore $|E_{\mathcal{F}'}| = n - \mathcal{C}_\#$

We know that $\mathcal{F}$ obtained by weight restricted version of modified Kruskal's algorithm. Since in step 2 of the algorithm for any new edge e we study the acyclic property and weight restriction $w(e) < \tau$ then $\mathcal{F}$ should also be an acyclic spanning subgraph of $\mathcal{S}$ so $|E_\mathcal{F}| \leq n - \mathcal{C}_\#$. Now we have to prove that $|E_\mathcal{F}|$ is also maximal.

By contradiction $|E_\mathcal{F}| = u$ where $u < n - \mathcal{C}_\#$ then for sure the number of connected components for $\mathcal{F}$ is smaller than $n - \mathcal{C}_\#$ so there is at least two neighbors i, j in $\mathcal{F}'$ that are connected by e' but are not connected in $\mathcal{F}$. The algorithm is greedy and it stops by violation of the weight restriction , $w(e') < \tau$, and the number of allowed iteration, $i = n-1$.
So we can still add e' without violating above conditions. Moreover, we know that adding e'

does not make a cycle since it will be the only path between i and j in the new graph. Therfore, the created graph is an optimal spanning subgraph with $|E_{\mathcal{F}}| = u+1$ which is in contradiction with initial assumption. So $|E_{\mathcal{F}}|$ is also maximal. $|E_{\mathcal{F}}| = n - \mathcal{C}_{\#}$.

Now we can use the entire steps of proof for optimality of size restrcited version of Kruskal's algorithm by considering $q := n - \mathcal{C}_{\#}$. Therfore, the forest $\mathcal{F}$ obtained by the weight restricted version of the Kruskal's algorithm is optimal.

Modified Kruskal's algorithms uniqueness

In this part, we pursue our studies on the necessary and sufficient conditions for uniqueness of modified versions of Kruksal's algorithm for minimum spanning forest. In parallel to what we state for minimum spanning forests, it is proven that the modified versions provide optimal spanning forest. So, the required conditions for uniqueness of minimum spanning forest are same the necessary conditions for uniqueness of solution of modified algorithms. In the following we prove that the minimum spanning forest defined by either size constraint or weight constraint is unique.

Size restrcited version of minimum spanning forest:

Theorem 10. *Let $\mathcal{G} = (V, E)$. These properties are equivalent:*

- ***Uniqueness of Minimum Spanning Forest defined by size constraint:*** *There is an unique minimum spanning forest $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ for $\mathcal{G} = (V, E)$, where $|V| = |V_{\mathcal{F}}| = n$ and $|E_{\mathcal{F}}| = q$ and q can be any integer number in $\{0,\ 1, \ldots,\ n-1\}$*
- ***Extreme cycle edge - size restricted version:*** *Let $N \subset E$ where for every e which is a non-cycle-heaviest edge in the graph $\mathcal{G}$, we have $e \in N$.*
  Sort edges in N in ascending order of weights. Label them by $N = \{e_N^{(1)},\ e_N^{(2)},\ \ldots,\ e_N^{(k)}\}$.
  N must satisfy two initial conditions:
  - *For the number of non-cycle-heaviest edges k we have $q \leq k$*
  - *The $w(e_N^q)$ is unique.*

  Then Extreme cycle edge - size restricted version is that:
  Every edge in $\mathcal{G}$ is either "unique-cycle-heaviest or non- cycle-heaviest which is heavier than $w(e_N^q)$" or "non-cycle-heaviest which is not heavier than $w(e_N^q)$".

Proof. Similar to what we have done for minimum spanning forest, one can show that if the graph $\mathcal{G} = (V, E)$ satisfies the "Extreme cycle edge - size restricted version" the edges that has the property: "unique-cycle -heaviest or non-cycle-heaviest which is heavier than $w(e_N^q)$" are not in the size constrained minimum spanning forest , and the edges that have the property: "non-cycle-heaviest is not heavier than $w(e_N^q)$" are in the size constrained minimum spanning forest.

Proof

Weight restrcited version of minimum spanning forest:

Theorem 11. *Let $\mathcal{G} = (V, E)$. These properties are equivalent:*

- ***Uniqueness of Minimum Spanning Forest defined by size constraint:*** *There is an unique minimum spanning forest $\mathcal{F} = (V_{\mathcal{F}}, E_{\mathcal{F}})$ for $\mathcal{G} = (V, E)$, where $|V| = |V_{\mathcal{F}}| = n$, for all $e \in E_{\mathcal{F}}$ and known τ, $w(e) < \tau$, and $|E_{\mathcal{F}}|$ is maximal.*
- ***Extreme cycle edge - weight restricted version:*** *Then Extreme cycle edge - weight restricted version is that:*
  Every edge is either "unique-cycle-heaviest or is not lighter than τ" or "non-cycle-heaviest and lighter than τ".

Proof. Similar to what we have done for minimum spanning forest, one can show that if the graph $\mathcal{G} = (V, E)$ satisfies the ”Extreme cycle edge - weight restricted version” the edges that has the property: ”unique-cycle-heaviest or is not lighter than τ” are not in the weight constrained minimum spanning forest, and the edges that have the property: ”non-cycle-heaviest and lighter than τ” are in the weight constrained minimum spanning forest.

Proof

In analogue with uniqueness of minimum spanning tree, although, it is not necessary condition for uniqueness of minimum spanning forest and the solution of modified Kruskal’s algorithms; it is for $\mathcal{G} = (V, E)$ to have uniquely weighted edges so automatically every edge satisfies one of the modified versions of extreme cycle edge property.

Chapter 4

Structure Learning For Extremal Forest Modles

Chapter 5

Applications