Construction of a copula estimator through recursive partitioning of the unit hypercube

O. Laverny $^{1,2}$ V. Maume-Deschamps $^1$ E. Masiello $^1$ D. Rullière $^1$
April 17, 2020

$^1$ Université Lyon 1

$^2$ SCOR SE
1. Introduction

2. Piecewise linear copulas

3. The Cort Estimator

4. Asymptotic behavior

5. Bagging and cross-validation

6. Simulation Study

7. Conclusion
Introduction
Suppose that $X$ is a (continuous) random vector of dimension $d$ with c.d.f $F$ and marginals c.d.f $(F_i)_{i \in \{1, \ldots, d\}}$. Then Sklar’s theorem [6] gives us the copula of $X$ as:

$$C(u) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))$$

- $C$ is a c.d.f with uniform margins on $[0, 1]$.
- It characterises the dependence structure of $F$ in the sense that $F$ is completely characterised by $C$ and the $F_i$’s.

The estimation of the copula is a wide-treated subject: there exists a lot of parametric distributions that can be fitted. Some non-parametric models exists but are facing problems in high dimensions.
Density estimation trees

In regression, the CART algorithm from Breiman [3] selects a covariate and a univariate breakpoint, minimizing a loss, and assign to each leaf the mean response inside the leaf.

In density estimation, the DET from Ram and Gray [4] selects a dimension and a breakpoint minimizing a loss, and assign to each leaf the frequency of observations:

\[ f(x) = \sum_{\ell \in \mathcal{L}} \frac{f_{\ell}}{\lambda(\ell)} \mathbf{1}_{x \in \ell} \]

- What loss can we use?
- Will this yield a copula if applied to pseudo-observations?
Piecewise linear copulas
Let $\II = [0, 1]^d$ be the unit hypercube and $\mathcal{L}$ a partition of $\II$.

**Definition (Piecewise linear copula)**

Let the piecewise linear copula be defined by its distribution function:

$$\forall u \in \II, \ C_{p,\mathcal{L}}(u) = \sum_{\ell \in \mathcal{L}} p_{\ell} \lambda_{\ell}(u)$$

- $\lambda_{\ell}(u) = \frac{\lambda([0,u] \cap \ell)}{\lambda(\ell)}$ where $\lambda$ is the lebesgue measure.
- $p$ is a vector of weights summing to one.

Corresponding density: $c_{p,\mathcal{L}}(u) = \sum_{\ell \in \mathcal{L}} \frac{p_{\ell}}{\lambda(\ell)} 1_{u \in \ell}$
Copula constraints

We restrict the leaves in \( \mathcal{L} \) to be hyperboxes of the form \([a, b]\).

**Property (Copula constraints are linear in the weights)**

\[
C_{p,\mathcal{L}} \text{ is a proper copula } \iff p \in C_\mathcal{L} = \{ p \in \mathbb{R}^{\mid \mathcal{L} \mid} : Bp = g \text{ and } p \geq 0 \}
\]

\[
B_1 = (\lambda_{\ell_i}(u_i))_{(i, u) \in \{1, \ldots, d\} \times M_\mathcal{L}, \ \ell \in \mathcal{L}} \quad \text{(size } nd \times \mid \mathcal{L} \mid) \\
B_2 = (1)_{\ell \in \mathcal{L}} \quad \text{(size } 1 \times \mid \mathcal{L} \mid) \\
g_1 = (u_i)_{(i, u) \in \{1, \ldots, d\} \times M_\mathcal{L}} \quad \text{(size } nd) \\
B = (B_1, B_2) \quad \text{(size } (nd + 1) \times \mid \mathcal{L} \mid) \\
g = (g_1, 1) \quad \text{(size } (nd + 1))
\]

Where \( M_\mathcal{L} \) is the set of middle-points of leaves in \( \mathcal{L} \).
Kendall $\tau$, Spearman $\rho$

**Definition**

$\tau = 4 \int C(u) c(u) \, du - 1$, and $\rho = 12 \int C(u) \, du - 3$, and $K(t) = \mathbb{P}(C(U) \leq t)$

**Property (Piecewise linear class)**

$\tau = -1 + 2 \sum_{\ell \in \mathcal{L}} \prod_{i=1}^{d} (b_i \wedge d_i - a_i \wedge c_i) (b_i \wedge d_i + a_i \wedge c_i - 2c_i) + 2 (d_i - c_i) (b_i - a_i \wedge d_i)$

$\rho = -3 + 6 \sum_{\ell \in \mathcal{L}} p_{\ell} \prod_{i=1}^{d} (2 - b_i - a_i)$

where we denoted $\ell = (a, b]$ and $k = (c, d]$, $\wedge$ denotes the minimum operator and $\gamma_+$ is the upper regularised gamma function.
The Cort Estimator
An integrated square error loss...

We use the integrated square error between densities.

\[ \left\| \mathbf{c}_{p,L} - \mathbf{c} \right\|^2_2 \approx \left\| \mathbf{c}_{p,L} \right\|^2_2 - 2 \langle \mathbf{c}_{p,L}, \mathbf{c} \rangle \]  
(additive indep.)

\[ \approx \left\| \mathbf{c}_{p,L} \right\|^2_2 - \frac{2}{n} \sum_{i=1}^{n} \mathbf{c}_{p,L}(u_i) \]  
(MC plug-in)

\[ = \sum_{\ell \in \mathcal{L}} \frac{p^2_{\ell}}{\lambda(\ell)} \- 2 \sum_{\ell \in \mathcal{L}} \frac{p_{\ell} f_{\ell}}{\lambda(\ell)} \]  
(f = emp. freq)

\[ = \mathbf{p}' \mathbf{A} \mathbf{p} - 2 \mathbf{p}' \mathbf{A} \mathbf{f} \]  
(A = \text{diag}(\lambda(\ell)^{-1})

\[ = \left\| \mathbf{p} \right\|_L^2 - 2 \langle \mathbf{p}, \mathbf{f} \rangle_L \]

Where \( \langle \mathbf{x}, \mathbf{y} \rangle_L = \sum_{\ell \in \mathcal{L}} \frac{x_{\ell} y_{\ell}}{\lambda(\ell)} \) is, indeed, a scalar product.
The weights $p^*$ that minimize the integrated square error for a given partition $\mathcal{L}$ are given by the following:

**Definition (Quadratic program)**

$p^*$ is the solution to the quadratic program:

$$\arg\min_{p \in \mathcal{C}} \|p\|_{\mathcal{L}}^2 - 2 \langle p, f \rangle_{\mathcal{L}}$$

which is the projection of $f$ onto $\mathcal{C}_{\mathcal{L}}$ regarding $\|\cdot\|_{\mathcal{L}}^2$.

We denote $p^* = P_{\mathcal{C}_{\mathcal{L}}}(f)$ this projection.

Note that without the constraints, $p^* = f$. 
Joint optimisation for the breakpoint and the weights

For a set of dimensions $D$ in $\mathcal{P}(\{1, ..., d\})$, let $L(\ell, x, D)$ be the partition of the leaf $\ell$ splitted on a point $x$ in dimensions $D$, i.e:

$$L([a, b], x, D) = \times_{j \in D} \{(a_j, x_j), (x_j, b_j)\} \times \{([a_j, b_j])\}.$$ 

Define then the full partition by:

$$\mathcal{L}_{x, D} = \mathcal{L} \setminus \{\ell(x)\} \cup L(\ell(x), x, D).$$

We will omit the parameter $D$ if $D = \{1, .., d\}$.

**Definition (Final optimisation problem)**

The global optimisation problem we want to solve is:

$$\arg \min_{D \in \mathcal{P}(\{1, ..., d\})} \|p\|^2_{\mathcal{L}_{x, D}} - 2 \langle p, f_{\mathcal{L}_{x, D}} \rangle_{\mathcal{L}_{x, D}}$$

subject to:

- $x \in \mathbb{I}$
- $p \in \mathcal{C}_{\mathcal{L}_{x, D}}$
The recursive procedure

1. Solve the *density* problem:

\[
\arg \min_{D \in \mathcal{P} \{1, \ldots, d\}} \min_{x \in \mathbb{I}} -\|f_{\mathcal{L}_x, D}\|_{\mathcal{L}_x, D}^2
\]

- Find the splitting dimensions \( D \) first
- Minimize greedily on \( x \) via a non-linear programming solver.

2. Recurse on each \( \ell \) in \( \mathcal{L}_{x, D} \) by rescaling \( \ell \) to \( \mathbb{I} \) and solving the same problem to obtain the final partition \( \mathcal{L} \).

3. Then, with \( \mathcal{L} \) fixed, solve the projection:

\[
\arg \min_{p \in C_{\mathcal{L}}} \|p\|_{\mathcal{L}}^2 - 2 \langle p, f_{\mathcal{L}} \rangle_{\mathcal{L}}
\]

via a quadratic programming solver, with initial values \( f_{\mathcal{L}} \).
Finding the splitting dimensions $D$ (for $U \sim C$)

<table>
<thead>
<tr>
<th>Hypothesis ($\mathcal{H}_j$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(U_j \perp \perp U_{-j}) \mid U \in \ell$ and $U_j \mid U \in \ell \sim \mathcal{U} (\ell_j)$</td>
</tr>
</tbody>
</table>

Bowman [2]: Suppose that $\ell = \mathbb{I}$, containing $n$ obs. of the random variable $U \sim F$, for $F$ the restriction of $C$ to $\ell$, rescaled to $\mathbb{I}$. Then:

**Definition (Test statistic)**

Denote by $f_{f,\mathcal{L}}^{(n)}$ the piecewise constant density that will be estimated on data $U_1, \ldots, U_n \sim F$, and $e_{j,n}(x) = \mathbb{E}(f_{f,\mathcal{L}}^{(n)}(x) \mid \mathcal{H}_j)$. The test statistic is given by:

$$I_j = \| e_{j,n} - f_{f,\mathcal{L}}^{(n)} \|_2^2$$

where $\mathcal{L}$, $e_{j,n}$ and $f_{f,\mathcal{L}}^{(n)}$ are stochastic objects, depending on $U$. 
We weakened the test by assuming that the next split is enough to test $\mathcal{H}_j$. This gives a test procedure as follows:

1. Solve

$$x^* = \arg \min_{x \in I} -\|f_{L_x}\|_{L_x}^2$$

2. Compute:

$$\hat{I}_j = \sum_{k \in \mathcal{L}_{x^*}, \{1, \ldots, d\} \setminus \{j\}} \left( \frac{f_k^2}{\lambda(k)} + \sum_{\ell \in \mathcal{L}_{x^*}, \{1, \ldots, d\} \setminus \ell \subset k} \left( \frac{f_\ell^2}{\lambda(\ell)} - 2 \frac{f_k f_\ell}{\lambda(k)} \right) \right).$$

**Argument:** the cut will be on the same $x$ in dimensions other than $j$ whether or not we work under $\mathcal{H}_j$.

3. Compare to a Monte-carlo simulation of its distribution under the null to exclude the dimension $j$ if necessary.
Asymptotic behavior
Ram and Gray [4] gave the consistency of \( f_{f,\mathcal{L}}^{(n)} \). Assuming the maximum diameter of leaves goes to 0 as \( n \) goes to \( \infty \), we have:

\[
\mathbb{P} \left( \lim_{n \to +\infty} \| f_{f,\mathcal{L}}^{(n)} - f \|_2^2 = 0 \right) = 1.
\]

Denoting \( q \) s.t:

\[
\forall \ell \in \mathcal{L}, \; q_{\ell} = \int \ell c(u) du,
\]

this results writes \( d_{\mathcal{L}}(f, q)^2 \to 0 \), a.s.

Furthermore, by construction, \( q \in \mathcal{C} \).
**Constraint influence**

<table>
<thead>
<tr>
<th>Definition (Integrated constraint influence)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|c_{p,\mathcal{L}}^{(n)} - f_{\mathcal{L}}^{(n)}|<em>2^2 = d</em>{\mathcal{L}}(p, f)^2$</td>
</tr>
</tbody>
</table>

This quantity measures how much $f$ and $p$ are far from each other. But since $f$ is closer and closer to $q$, which is in the set that $f$ is projected on to give $p$, we have:

<table>
<thead>
<tr>
<th>Property (Asymptotical effect of constraints)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The integrated constraint influence is asymptotically 0.</td>
</tr>
</tbody>
</table>

**Proof.**

$C$ is convex, closed and non-empty. Hence $p = \mathcal{P}_C(f)$ exist and is unique. Since $q \in C$, we have that $d_{\mathcal{L}}(f, p)^2 \leq d_{\mathcal{L}}(f, q)^2$. □
Property (Consistency)

For $c$ the density of the true copula, assuming the diameter of the leaves goes to 0 as $n$ goes to $\infty$, the estimator $c_{p,L}^{(n)}$ is consistent, i.e:

$$\mathbb{P} \left( \lim_{n \to +\infty} \| c_{p,L}^{(n)} - c \|_2^2 = 0 \right) = 1$$

Proof.

$$\| c_{p,L}^{(n)} - c \|_2^2 = d_L(p, q)^2 \text{ and } d_L(p, q)^2 \leq d_L(f, q)^2.$$
Bagging and cross-validation
A simple forest

In regression: See Breiman [3]

In density estimation, Kernels uses leave-one-out for bandwidths. Sain, Baggerly and Scot [5] formalized the cross-validation process for density estimation. The more involved out-of-bag procedure we propose is inspired by Wu [7].

Definition (Out-of-bag "density" and metrics)

\[ c_{ooob}(u) = \frac{1}{N(u)} \sum_{j=1}^{N} c^{(j)}(u) \mathbf{1}_{u \text{ was not in the training set of } c^{(j)}} \]

\[ J_{ooob}(c_N) = \|c_N\|_{2}^{2} - \frac{2}{n} \sum_{i=1}^{n} c_{ooob}(u_i) \]

\[ KL_{ooob}(c_N) = \int c(u) \ln \left( \frac{c(u)}{c_N(u)} \right) \approx -\frac{1}{n} \sum_{i=1}^{n} ln(c_{ooob}(x_i)) \]
A weighted forest

After fitting the trees \( c^1, \ldots, c^N \), we can assign weights to them minimizing an out-of-bag integrated square error for the forest:

**Definition (Out-of-bag "density" and metrics, weighted case)**

\[
c_{\text{oob}}^w(u) = \frac{1}{W(u)} \sum_{j=1}^{N} w_j c^{(j)}(u) \mathbf{1}_u \text{ was not in the training set of } c^{(j)}
\]

Where \( W(u) \) is the sum of \( w_j \)'s for trees that did not see \( u \). Then:

**Definition (Optimal weights)**

\[
w^* = \arg \min_{w} J_{\text{oob}}(c_N^w)
\]
Simulation Study
The Cort estimator is implemented in the `cort` R package, available on CRAN.

The dataset is as follows:

- Simulation of 200 points from a 3-dimensional clayton copula with $\theta = 8$ for marginals 1, 3 and 4.
- The second marginal is added as independent uniform draws.
- The fourth marginal is flipped, inducing anticomonotonicity.

Marginals 1, 3 and 4 exhibit strong dependency.
Figure 1: The dataset we will use.
Figure 2: In gray scale, we observe a bivariate histogram of the simulation from the estimated tree. The small red points represent the input data.
Figure 3: Left: Oob Kullback-leibler and Oob ISE if function of the number of trees; Right: Constraint influence and L2-Norm in function of weights
Figure 4: Top row: kendall’s taus. Bottom row: Spearman’s rho. Left: empirical values from burn-in data. Right: values from the fitted models. The size of the subsamples is in abssissa.
### Table 1: Statistics of several models on the Dataset

<table>
<thead>
<tr>
<th>Kendall Taus</th>
<th>Empirical</th>
<th>Cb(m=10)</th>
<th>Cb(m=5)</th>
<th>Beta</th>
<th>Cort</th>
<th>Bagged Cort</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{1,2}$</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>$\tau_{1,3}$</td>
<td>-0.80</td>
<td>-0.75</td>
<td>-0.68</td>
<td>-0.80</td>
<td>-0.78</td>
<td>-0.56</td>
</tr>
<tr>
<td>$\tau_{1,4}$</td>
<td>0.78</td>
<td>0.73</td>
<td>0.66</td>
<td>0.78</td>
<td>0.71</td>
<td>0.54</td>
</tr>
<tr>
<td>$\tau_{2,3}$</td>
<td>0.03</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>$\tau_{2,4}$</td>
<td>-0.03</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.04</td>
<td>0.00</td>
<td>-0.05</td>
</tr>
<tr>
<td>$\tau_{3,4}$</td>
<td>-0.78</td>
<td>-0.73</td>
<td>-0.65</td>
<td>-0.77</td>
<td>-0.69</td>
<td>-0.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spearman Rhos</th>
<th>Empirical</th>
<th>Cb(m=10)</th>
<th>Cb(m=5)</th>
<th>Beta</th>
<th>Cort</th>
<th>Bagged Cort</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_{1,2}$</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.02</td>
</tr>
<tr>
<td>$\rho_{1,3}$</td>
<td>-0.93</td>
<td>-0.91</td>
<td>-0.87</td>
<td>-0.93</td>
<td>-0.93</td>
<td>-0.72</td>
</tr>
<tr>
<td>$\rho_{1,4}$</td>
<td>0.93</td>
<td>0.90</td>
<td>0.86</td>
<td>0.93</td>
<td>0.87</td>
<td>0.70</td>
</tr>
<tr>
<td>$\rho_{2,3}$</td>
<td>0.04</td>
<td>0.02</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>$\rho_{2,4}$</td>
<td>-0.05</td>
<td>-0.03</td>
<td>-0.01</td>
<td>-0.06</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>$\rho_{3,4}$</td>
<td>-0.92</td>
<td>-0.90</td>
<td>-0.85</td>
<td>-0.92</td>
<td>-0.86</td>
<td>-0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bagging Results</th>
<th>Empirical</th>
<th>Cb(m=10)</th>
<th>Cb(m=5)</th>
<th>Beta</th>
<th>Cort</th>
<th>Bagged Cort</th>
</tr>
</thead>
<tbody>
<tr>
<td>$KL_{oob}$</td>
<td>Inf</td>
<td>4.48</td>
<td>3.80</td>
<td>-4.55</td>
<td>-5.15</td>
<td>NaN</td>
</tr>
</tbody>
</table>
Conclusion
Some take away points:

- Piecewise linear distribution function are handy models for copula modeling since the copula constraints have a nice expression.
- Fitting piecewise linear d.f with trees is quite simple and fast.
- The main issue is the degree of freedom in weights took away by the copula constraint.
- Such models can easily be bagged, boosted, cross-validated...
Gérard Biau, Erwan Scornet, and Johannes Welbl. “Neural Random Forests”. In: (Apr. 25, 2016).


A Sklar. “Fonctions de Repartition à n Dimension et Leurs Marges”. In: *Université Paris* 8.3.2 (1959), pp. 1–3.