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

Table of contents

Introduction

Copulas Basics

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

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

- 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*:

$$f(x) = \sum_{\ell \in \mathcal{L}} \frac{\mathrm{f}_\ell}{\lambda(\ell)} \mathbf{1}_{x \in \ell}$$

- What loss can we use ?
- Will this yield a copula if applied to pseudo-observations ?

Piecewise linear copulas

Definition

Let $\mathbb{I} = [0,1]^d$ be the unit hypercube and $\mathcal{L}$ a partition of $\mathbb{I}$.

Definition (Piecewise linear copula)

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

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

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

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

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)

$$C_{p,\mathcal{L}} \textit{ is a proper copula}$$

$$\iff$$

$$p \in \mathcal{C}_{\mathcal{L}} = \{p \in \mathbb{R}^{|\mathcal{L}|} : \ Bp = g \textit{ and } p \geq 0\}$$

$$
\begin{aligned}
B_1 &= (\lambda_{\ell_i}(u_i))_{(i,u)\in\{1,\dots,d\}\times M_{\mathcal{L}},\ \ell\in\mathcal{L}} && (\text{size } nd \times |\mathcal{L}|)\\
B_2 &= (1)_{\ell\in\mathcal{L}} && (\text{size } 1 \times |\mathcal{L}|)\\
g_1 &= (u_i)_{(i,u)\in\{1,\dots,d\}\times M_{\mathcal{L}}} && (\text{size } nd)\\
B &= (B_1, B_2) && (\text{size } (nd+1) \times |\mathcal{L}|)\\
g &= (g_1, 1) && (\text{size } (nd+1))
\end{aligned}
$$

Where $M_{\mathcal{L}}$ is the set of middle-points of leaves in $\mathcal{L}$.

Kendall τ, Spearman ρ

Definition

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

Property (Piecewise linear class)

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

where we denoted $\ell = (a, b]$ and $k = (c, d]$, $\wedge$ denotes the minimum operator and γ_+ is the upper regularised gamma function.

The Cort Estimator

An integrated square error loss...

We use the integrated square error between densities.

$$\begin{aligned}
\|c_{p,\mathcal{L}} - c\|_2^2 &\approx \|c_{p,\mathcal{L}}\|_2^2 - 2\langle c_{p,\mathcal{L}}, c\rangle && \text{(additive indep.)} \\
&\approx \|c_{p,\mathcal{L}}\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} c_{p,\mathcal{L}}(u_i) && \text{(MC plug-in)} \\
&= \sum_{\ell\in\mathcal{L}} \frac{p_\ell^2}{\lambda(\ell)} - 2\sum_{\ell\in\mathcal{L}} \frac{p_\ell \mathrm{f}_\ell}{\lambda(\ell)} && (\mathrm{f} = \text{emp. freq}) \\
&= p'Ap - 2p'A\mathrm{f} && (A = diag(\lambda(\ell)^{-1}) \\
&= \|p\|_{\mathcal{L}}^2 - 2\langle p, \mathrm{f}\rangle_{\mathcal{L}}
\end{aligned}$$

Where $\langle x, y\rangle_{\mathcal{L}} = \sum_{\ell\in\mathcal{L}} \frac{x_\ell y_\ell}{\lambda(\ell)}$ is, indeed, a scalar product.

… yields a simple quadratic program

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 :

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

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

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

Note that without the constraints, $p^* = \mathrm{f}$.

Joint optimisation fo 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 ℓ splitted on a point x in dimensions D, i.e:

$$L((a,b], x, D) = \bigtimes_{j \in D} \{(a_j, x_j], (x_j, b_j]\} \bigtimes_{j \in \{1,...,d\} \setminus D} \{(a_j, b_j]\}.$$

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 :

$$\underset{\substack{D \in \mathcal{P}(\{1,...,d\}) \\ x \in \mathbb{I} \\ p \in \mathcal{C}_{\mathcal{L}_{x,D}}}}{\arg\min} \quad \|p\|^2_{\mathcal{L}_{x,D}} - 2 \langle p, \mathrm{f}_{\mathcal{L}_{\mathrm{x,D}}} \rangle_{\mathcal{L}_{x,D}}$$

The recursive procedure

1. Solve the *density* problem:

$$\underset{\substack{D \in \mathcal{P}(\{1,\ldots,d\}) \\ x \in \mathbb{I}}}{\arg\min} \quad -\|\mathrm{f}_{\mathcal{L}_{\mathrm{x},\mathrm{D}}}\|^2_{\mathcal{L}_{x,D}}$$

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

2. Recurse on each ℓ in $\mathcal{L}_{x,D}$ by rescaling ℓ to $\mathbb{I}$ and solving the same problem to obtain the final partition $\mathcal{L}$.
3. Then, with $\mathcal{L}$ fixed, solve the projection:

$$\underset{p \in \mathcal{C}_{\mathcal{L}}}{\arg\min} \; \|p\|^2_{\mathcal{L}} - 2\,\langle p, \mathrm{f}_{\mathcal{L}} \rangle_{\mathcal{L}}$$

   via a quadratic programming solver, with inital values $\mathrm{f}_{\mathcal{L}}$.

Finding the splitting dimensions D (for $U \sim C$)

Hypothesis ($\mathcal{H}_j$)

$$(U_j \perp\!\!\!\perp U_{-j})\,|U \in \ell \text{ and } U_j|U \in \ell \sim \mathcal{U}(\ell_j)$$

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

Definition (Test statistic)

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

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

where $\mathcal{L}$, $e_{j,n}$ and $f_{\mathrm{f},\mathcal{L}}^{(n)}$ are stochastic objetcs, depending on U.

Test procedure

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
$$x^* = \underset{x \in \mathbb{I}}{\arg\min} \; -\|\mathrm{f}_{\mathcal{L}_\mathrm{x}}\|^2_{\mathcal{L}_x}$$
2. Compute:
$$\widehat{\mathcal{I}}_j = \sum_{k \in \mathcal{L}_{x^*,\{1,\dots,d\}} \setminus \{j\}} \left(\frac{\mathrm{f}_k^2}{\lambda(k)} + \sum_{\substack{\ell \in \mathcal{L}_{x^*,\{1,..,d\}} \\ \ell \subset k}} \left(\frac{\mathrm{f}_\ell^2}{\lambda(\ell)} - 2\frac{\mathrm{f}_k \mathrm{f}_\ell}{\lambda(k)} \right) \right).$$
Argument : the cut will be on the same x in dimensions other than j wheter 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

Previous result

Ram and Gray [4] gave the consistency of $f_{\mathrm{f},\mathcal{L}}^{(n)}$. Assuming the maximum diameter of leaves goes to 0 as n goes to ∞, we have :

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

Denoting q s.t:

$$\forall\ell\in\mathcal{L},\ q_\ell=\int_\ell c(u)du,$$

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

Furthermore, by construction, $q\in\mathcal{C}$.

Constraint influence

Definition (Integrated constraint influence)

$$\|c_{p,\mathcal{L}}^{(n)} - f_{\mathrm{f},\mathcal{L}}^{(n)}\|_2^2 = d_{\mathcal{L}}(p, \mathrm{f})^2$$

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 :

Property (Asymptotical effect of constraints)

The integrated constraint influence is asymptotically 0.

Proof.

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

Cort consistency

Property (Consistency)

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

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

Proof.

$\|c_{p,\mathcal{L}}^{(n)}-c\|_2^2=d_{\mathcal{L}}(p,q)^2$ *and* $d_{\mathcal{L}}(p,q)^2\leq d_{\mathcal{L}}(\mathrm{f},q)^2$. $\square$

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)

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

$$J_{oob}(c_N) = \|c_N\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} c_{oob}(u_i)$$

$$KL_{oob}(c_N) = \int c(u) \ln\left(\frac{c(u)}{c_N(u)}\right) \approx -\frac{1}{n} \sum_{i=1}^{n} ln(c_{oob}(x_i))$$

A weighted forest

After fitting the trees $c^1, ..., c^N$, 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)

$$c^w_{oob}(u) = \frac{1}{W(u)} \sum_{j=1}^{N} w_j c^{(j)}(u) \mathbf{1}_{\text{u was not in the training set of } c^{(j)}}$$

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

Definition (Optimal weights)

$$w^* = \arg\min_w J_{oob}(c^w_N)$$

Simulation Study

The Cort estimator is implemented in the cort R package, avaliable 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 exibit strong dependency.

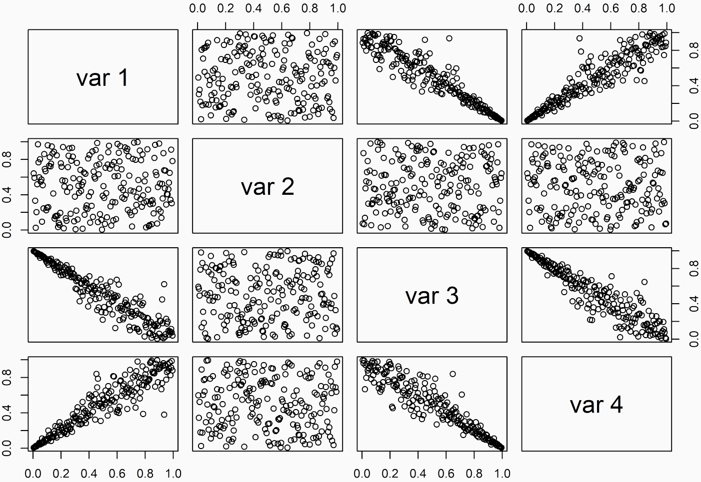

Figure 1: The dataset we will use.

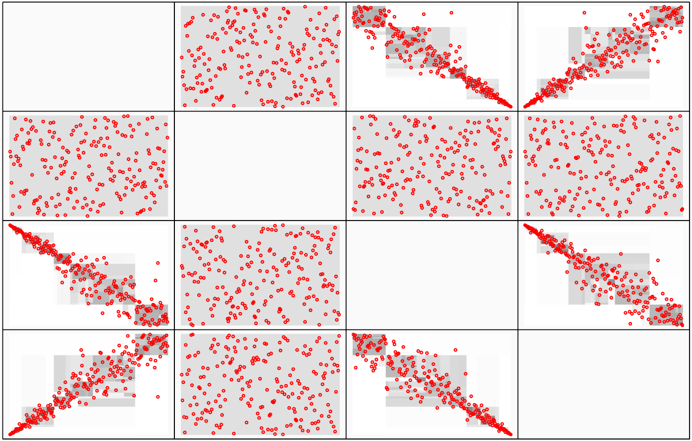

Figure 2: In gray scale, we observe a bivariate histogram of the simulation from the estimated tree. The small red points represent the input data.

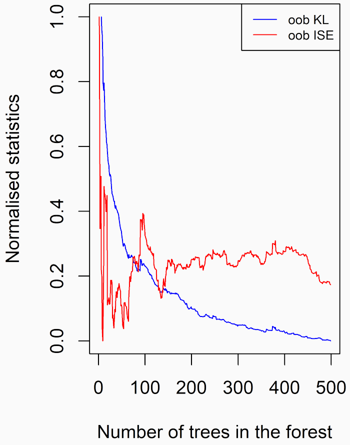

Figure 3: Left : Oob Kullback-leibler and Oob ISE if function of the number of trees; Right : Constraint influence and L2-Norm in function of weights

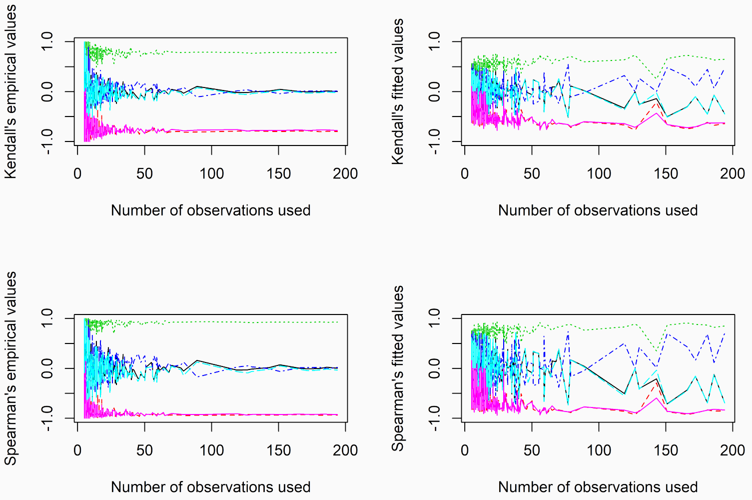

Figure 4: Top row: kendall's taus. Bottom row: Spearman's rho. Left: empirical values from burn-in data. Right : values from the fitted models. The size of the subsamples is in abssissa.

Table 1: Statitics of several models on the Dataset

	Empirical	Cb(m=10)	Cb(m=5)	Beta	Cort	Bagged Cort
Kendall Taus						
$\tau_{1,2}$	-0.01	0.00	0.02	-0.01	0.00	-0.04
$\tau_{1,3}$	-0.80	-0.75	-0.68	-0.80	-0.78	-0.56
$\tau_{1,4}$	0.78	0.73	0.66	0.78	0.71	0.54
$\tau_{2,3}$	0.03	0.02	0.00	0.02	0.00	0.05
$\tau_{2,4}$	-0.03	-0.02	-0.01	-0.04	0.00	-0.05
$\tau_{3,4}$	-0.78	-0.73	-0.65	-0.77	-0.69	-0.55
Spearman Rhos						
$\rho_{1,2}$	-0.02	0.00	0.02	-0.02	0.00	-0.02
$\rho_{1,3}$	-0.93	-0.91	-0.87	-0.93	-0.93	-0.72
$\rho_{1,4}$	0.93	0.90	0.86	0.93	0.87	0.70
$\rho_{2,3}$	0.04	0.02	0.00	0.04	0.00	0.04
$\rho_{2,4}$	-0.05	-0.03	-0.01	-0.06	0.00	-0.04
$\rho_{3,4}$	-0.92	-0.90	-0.85	-0.92	-0.86	-0.71
Bagging Results						
KL_{oob}	Inf	4.48	3.80	-4.55	-5.15	NaN

Conclusion

Take away and potential improvements

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...

References

Gérard Biau, Erwan Scornet, and Johannes Welbl. “Neural Random Forests”. In: (Apr. 25, 2016).

A.W. Bowman. “Density Based Tests for Goodness-of-Fit”. In: *Journal of Statistical Computation and Simulation* 40.1-2 (Feb. 1992), pp. 1–13. ISSN: 0094-9655, 1563-5163.

Leo Breiman. “Out-of-Bag Estimation”. In: (1996), p. 13.

Parikshit Ram and Alexander G. Gray. “Density Estimation Trees”. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’11*. The 17th ACM SIGKDD International Conference. San Diego, California, USA: ACM Press, 2011, p. 627. ISBN: 978-1-4503-0813-7.

Stephan R Sain, Keith A Baggerly, and David W Scott. “Cross-Validation of Multivariate Densities”. In: (1994), p. 12.

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

Kaiyuan Wu, Wei Hou, and Hongbo Yang. “Density Estimation via the Random Forest Method”. In: (2017), p. 28.