Estimation of a copula through recursive partitioning of the unit hypercube

The Copula Recursive Tree

O Laverny [1, 3] olaverny@scor.com

V. Maume-Deschamps[1] E. Masiello[1] D. Rullière[2]

[1] *Institut camille jordan* [2] *Institut de sciences financières et d'assurance* [3] *SCOR SE*

Introduction

Copulas are multivariate distributions with $\mathcal{U}([0,1])$ margins. They allow to separate dependence and marginals estimations (Sklar 1959). There exist a lot of parametric models, although high-dimensional cases are problematic for them, and few non-parametric ones which usually requires a lot of data.

CORT is a nonparametric recursive copula estimator, inspired by patchwork copulas of (Durante et al. 2015)

Piecewise linear copulas

An histogram with irregular bin shapes and weights...

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

- $\mathcal{L}$ is a partition of $[0,1]^d$
- p are associated weights
- λ the Lebesgue measure

Following (Ram and Gray 2011), we use hyperrectangular leaves. The fitting algorithm finds, recursively, the best breakpoint, like CART does. But here, breakpoints are **multidimensional**, so the model is expressive enough to accept the copula constraints.

...and easy integration of copula constraints

The copula constraints defines a convex, closed, and non-empty domain $\mathcal{C}$, given by:

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

where the matrix B and the vector g depends on $\mathcal{L}$. We solve for the projection of f, the empirical frequencies in the leaves, onto $\mathcal{C}$, with respect to the distance associated to the matrix A.

Fitting procedure

Given the partition $\mathcal{L}$, we use the integrated square error (ISE) loss to optimize the weights p over $\mathcal{C}$:

$$\begin{aligned} I_c(c_{p,\mathcal{L}}) &= \|c_{p,\mathcal{L}} - c\|_2^2 \\ &\propto \|c_{p,\mathcal{L}}\|_2^2 - 2\mathbb{E}(c_{p,L}(U)) \quad (\text{for } U \sim C) \\ &\approx \|c_{p,\mathcal{L}}\|_2^2 - 2n^{-1}\sum_{i=1}^{n} c_{p,L}(u_i) \quad (\text{for data } u) \\ &= p'Ap + 2p'Af \quad (\text{for matrix A and vector f}) \end{aligned}$$

Numerical exemple

Exemple dataset: 200 points, one uniform marginal and three from a Clayton (with one reversed).

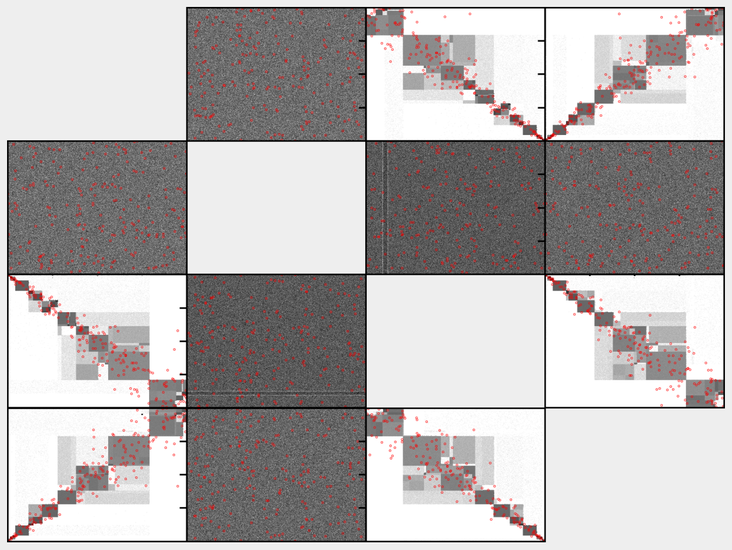

Figure 1: Obtained histogram (gray) against sample data (red) in the tree estimator

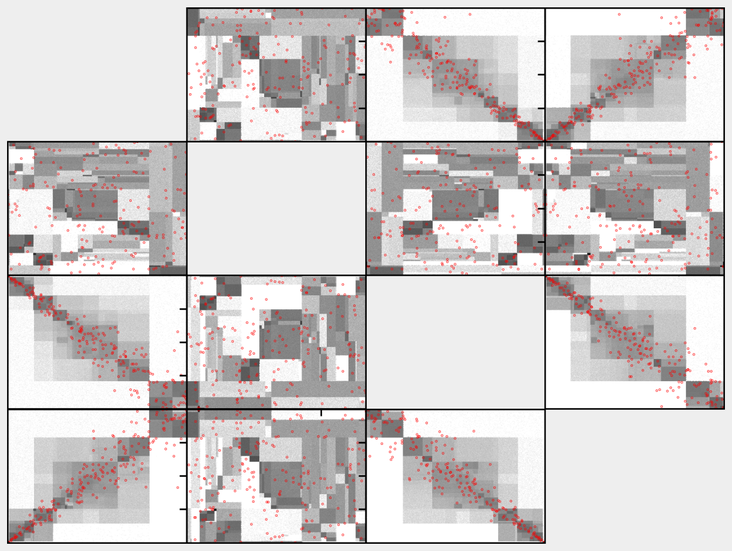

Figure 2: Obtained histogram (gray) against sample data (red) without dimension reduction

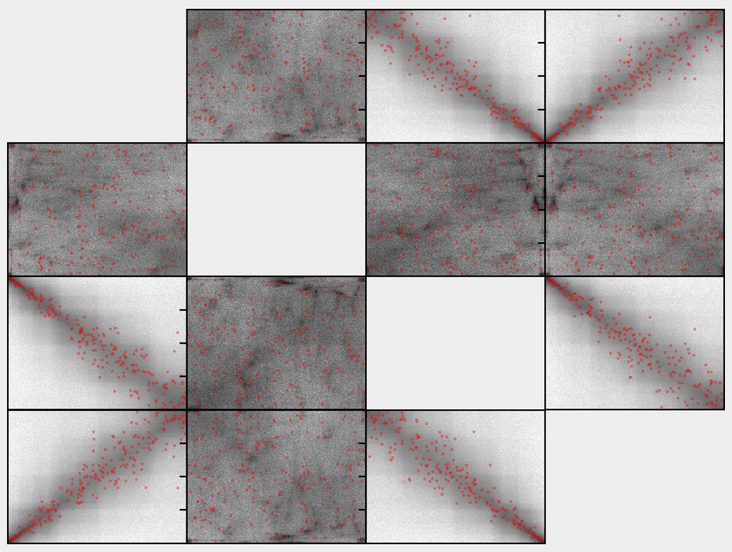

Figure 3: Obtained histogram (gray) against sample data (red) in the forest estimator

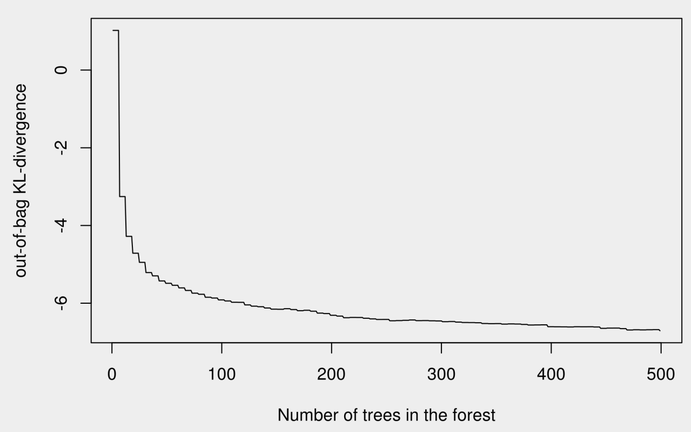

Figure 4: Out-of-bag Kullback-Leibler divergence K_N given by the fitted forest

Refinnements

Dimension reduction: An ISE based test to exclude splitting dimensions. The hypothesis is:

$$\mathcal{H}_j : \quad \left(U_j \perp\!\!\!\perp U_{-j}\right) | U \in \ell \text{ and } U_j | U \in \ell \sim \mathcal{U}(\ell_j).$$

The test statistic is then:

$$\mathcal{I}_j = \|f_{\mathrm{f},\mathcal{L}}^{(n)} - \mathbb{E}\left(f_{\mathrm{f},\mathcal{L}}^{(n)}(x) | \mathcal{H}_j\right)\|_2^2$$

Following (Bowman 1992), the empirical value of the statistic is compared to a monte-carlo distribution. We make further simplifications to estimate $\mathcal{I}_j$.

Consitency: Using a result from (Ram and Gray 2011), we show that the model is consistent, meaning that, as soon as the maximum diameter of the leaves decreases towards 0 with the increasing number of observation, we have:

$$\|c_{p,\mathcal{L}}^{(n)} - c\|_2^2 \xrightarrow[n\to\infty]{a.s} 0$$

Copula random forest

The CORT estimator can be *bagged* easily, following (Wu, Hou, and Yang 2017) for the density bagging theory. Out-of-bag samples permits estimation of an *oob* density, less prone to overfit, which can provide fitting statistics (ISE, KL divergence):

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

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

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

Conclusion

Piecewise linear copulas are handy models since the copula constraints have a nice expression for them. Moreover, fitting piecewise linear distribution function as trees is quite fast. However, the constraints in the weights reduce the degree of freedom, forcing us to use multidimensional splits, making the model harder to fit. Finaly, the CORT esitmator can easily be bagged, boosted, cross-validated, etc.

References

Bowman, A.W. 1992. "Density Based Tests for Goodness-of-Fit." *Journal of Statistical Computation and Simulation* 40 (1-2): 1–13.

Durante, Fabrizio, Juan Fernández-Sánchez, José Juan Quesada-Molina, and Manuel Úbeda-Flores. 2015. "Convergence Results for Patchwork Copulas." *European Journal of Operational Research* 247 (2): 525–31.

Ram, Parikshit, and Alexander G. Gray. 2011. "Density Estimation Trees." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, 627. San Diego, California, USA: ACM Press.

Sklar, A. 1959. "Fonctions de Repartition à N Dimension et Leurs Marges." *Université Paris* 8 (3.2): 1–3.

Wu, Kaiyuan, Wei Hou, and Hongbo Yang. 2017. "Density Estimation via the Random Forest Method," 28.

SCOR The Art & Science of Risk

Math Institut Camille Jordan