Computing Highest Density Regions with Copulae

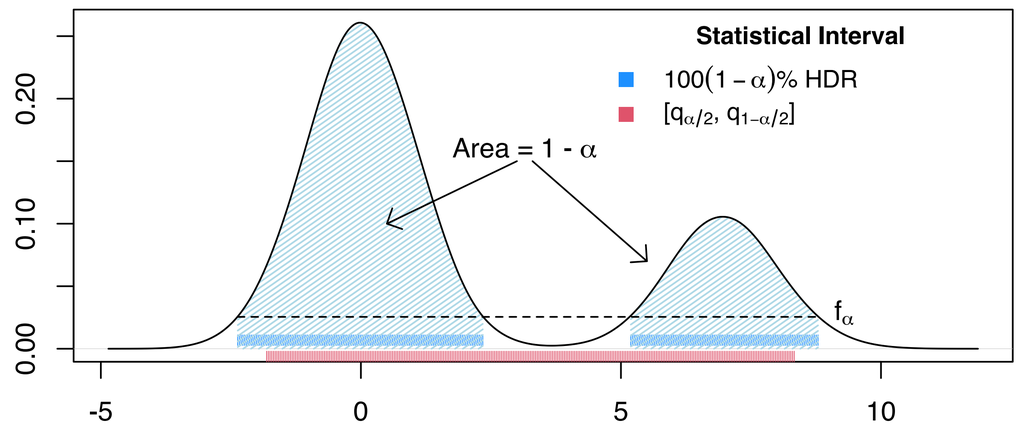
Nina Deliu and Brunero Liseo
MEMOTEF, Sapienza University of Rome

Highest Density Regions [Hyndman, 1996]

Let $f_{\mathbf{X}}$ denote the PDF of a continuous random variable $\mathbf{X} \in \mathbb{R}^d$, with $d \geq 1$. Denote with $1-\alpha$, where $\alpha \in (0,1)$, the required coverage probability. A $100(1-\alpha)\%$ **highest density region** (HDR) is defined as the subset $R(f_\alpha)$ of the sample space of $\mathbf{X}$ such that:

$$R(f_\alpha) = \{\mathbf{x}\colon f_{\mathbf{X}}(\mathbf{x}) \geq f_\alpha\}, \tag{1}$$

with f_α being the largest constant such that $P(\mathbf{X} \in R(f_\alpha)) \geq 1-\alpha$.



- Flexibility and effectiveness "to convey both multimodularity and asymmetry in the density".
- Problems / applications of interest: prediction intervals, anomaly detection, unsupervised and semisupervised classification.
- Interest in Bayesian analysis as well: *highest posterior density regions* and *credibility regions* [Box and Tiao, 2011].

A Copula-based Approach for HDR Estimation

Copulae and Sklar's Theorem [Sklar, 1959]

Given a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ and a d-variate CDF $\mathbf{F_X}$ which can depend on some parameters $\lambda_1, \cdots, \lambda_d$, [Sklar, 1959] showed that there always exists a d-variate function $C = C_\theta\colon [0,1]^d \to [0,1]$, such that:

$$F(x_1, \ldots, x_d; \lambda_1, \cdots, \lambda_d, \theta) = C_\theta(F_1(x_1;\lambda_1), \ldots, F_d(x_d;\lambda_d)),$$

where F_j is the marginal CDF of X_j.

Therefore, in case that the multivariate distribution has a density $\mathbf{f}$, and this is available, it holds further that

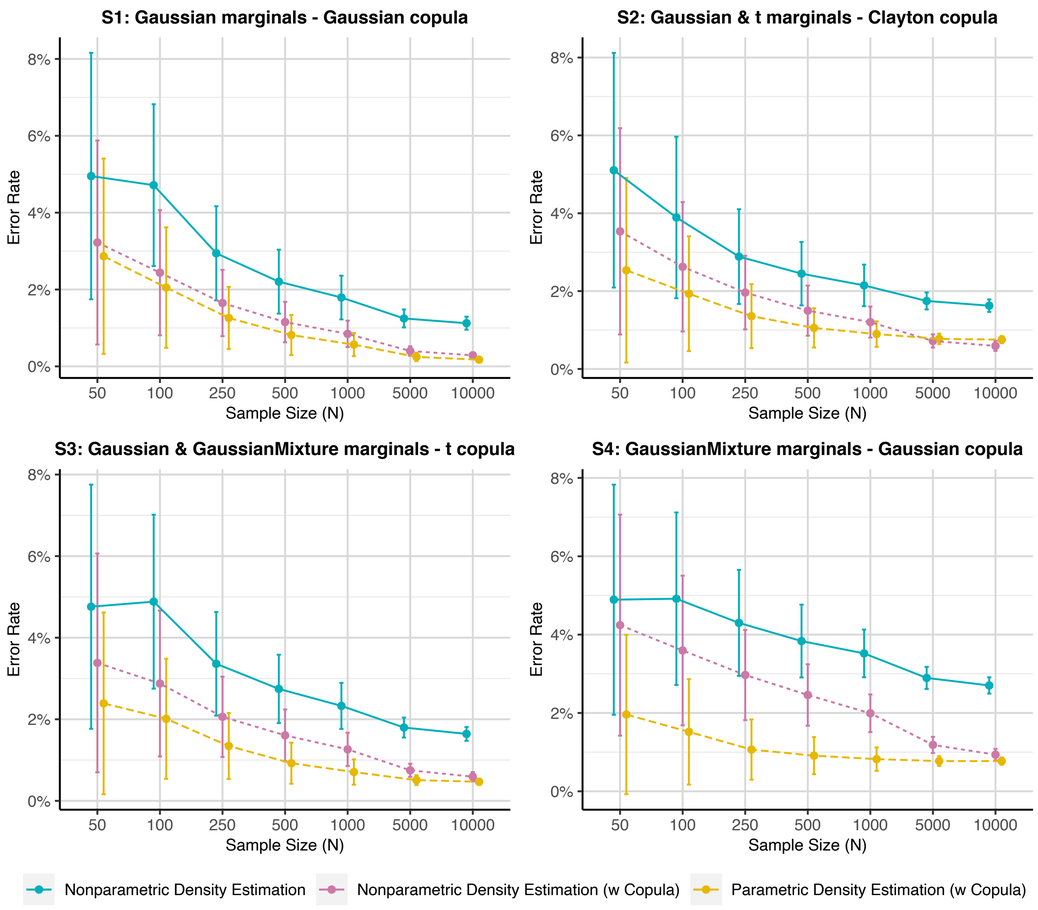
$$\underbrace{\mathbf{f}(x_1, \ldots, x_d)}_{\text{Multivariate density}} = \underbrace{c_\theta(F_1(x_1), \ldots, F_d(x_d))}_{\text{Copula density}} \times \underbrace{f_1(x_1) \times \cdots \times f_d(x_d)}_{\text{Marginal univariate densities}}$$

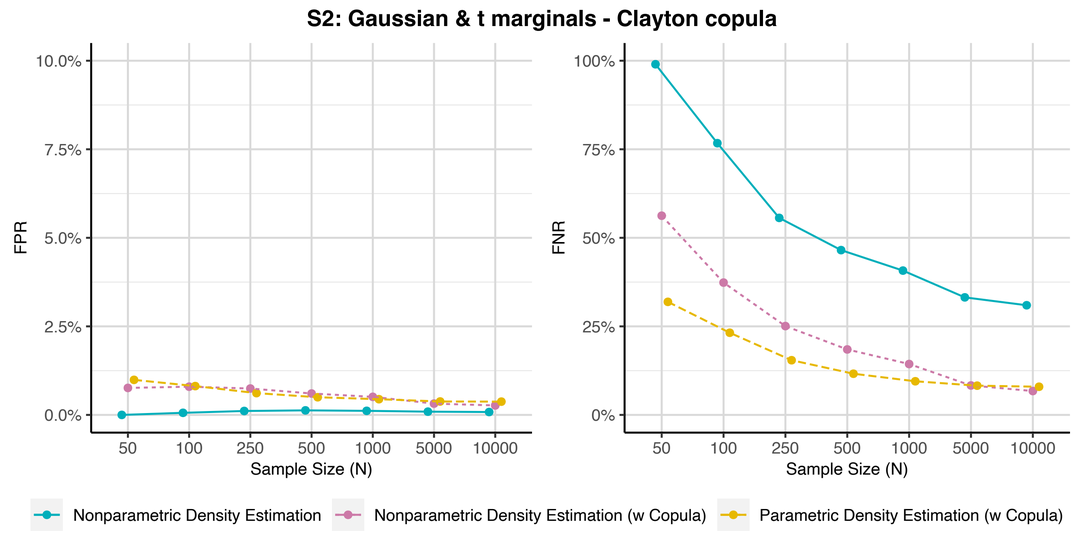
- The typical approach to estimate HDRs is the *density quantile approach*, that is, the sample quantile of a random variable $\mathbf{Y} = f_{\mathbf{X}}(\mathbf{X})$.
- Problem(s): f is generally unknown and estimation methods such as KDE are critical in multidimensional settings (bandwidth selection, algorithmic/computational issues [Liu et al., 2007, Wand and Jones, 1994])
- A copula approach may overcome the multidimensional issue.
- In a bivariate continuous case, with $\mathbf{X} = (X^{(1)}, X^{(2)})$, denoted with $\hat{c}$ an estimator for the copula density, we get an $100(1-\alpha)\%$ HDR estimate as:

$$\hat{R}_n(\hat{f}_\alpha) = \{\mathbf{x}\colon \hat{c}_{\mathbf{X}}(\hat{F}_1(x^{(1)}), \hat{F}_2(x^{(2)}))\hat{f}_1(x^{(1)})\hat{f}_2(x^{(2)}) > f_{\lfloor \alpha n \rfloor}\},$$

with $\hat{f}_j$ and $\hat{F}_j$, $j = 1, 2$, consistent estimators of the marginals and n the size of an iid sample of observations.

Simulation Studies



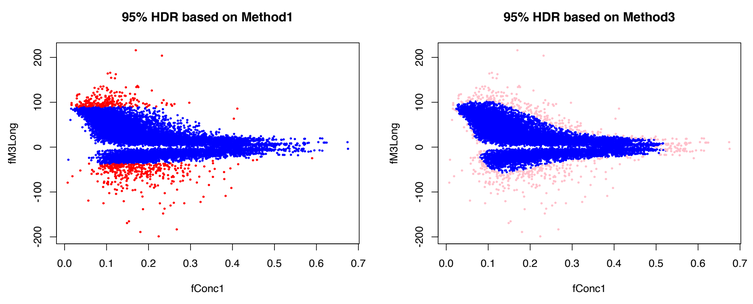


$$\text{FNR} = \frac{FN}{FN+TP}; \quad \text{FPR} = \frac{FP}{FP+TN}; \quad \text{ER} = \frac{FN+FP}{FN+FP+TN+TP}$$

Results

- The copula-based approaches outperform the direct KDE.
- The parametric copula-based approach shows the lowest ER, FPR, FNR.

MAGIC Data



References

[Box and Tiao, 2011] Box, G. E. and Tiao, G. C. (2011). *Bayesian Inference in Statistical Analysis*. John Wiley & Sons.

[Hyndman, 1996] Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126.

[Liu et al., 2007] Liu, H., Lafferty, J., and Wasserman, L. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. In *Artificial Intelligence and Statistics*, pages 283–290. PMLR.

[Sklar, 1959] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231.

[Wand and Jones, 1994] Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. CRC press.