Pseudo-Likelihood Information Criteria

Aibat Kossumov, Department of Statistics, Faculty of Math. & Phys., Charles Univ.

Abstract

The Akaike Information Criterion (AIC) based on pseudo-likelihood is known to be not justified for selecting parametric copula models [1]. Alternatives based on leave-one-out cross-validation, such as xv_1 and its approximation $\mathrm{xv}_{\mathrm{CIC}}$, were proposed, but showed only minor differences to AIC [2]. Inspired by [3], we apply leave-n_v-out cross-validation, with $n_v/n \xrightarrow[n\to\infty]{} 1$, to copula model selection and compare it with existing criteria.

Pseudo-Likelihood & AIC

We restrict our attention to the two-dimensional case and copula families with a one-dimensional dependence parameter θ, such as Clayton, Gumbel, Joe, Frank, and Gaussian. Denote by $\mathcal{X}_n = \{\boldsymbol{x}_i\}_{i=1}^n$ a random sample from the joint cdf

$$H(x_1, x_2) = C(F_1(x_1), F_2(x_2)),$$

where C is the copula, and F_1 and F_2 are continuous but unknown marginal cdfs. Also, define

$$\widetilde{\boldsymbol{F}}_n(x_1, x_2) = \left(\widetilde{F}_{n,1}(x_1), \widetilde{F}_{n,2}(x_2)\right),$$

where $\widetilde{F}_{n,k}$ is the $\frac{n}{n+1}$-rescaled empirical cdf of the kth marginal, for $k = 1, 2$. The corresponding pseudo-observations are denoted by ${}^p\mathcal{X}_n = \{{}^p\boldsymbol{x}_i\}_{i=1}^n$, where ${}^p\boldsymbol{x}_i = \widetilde{\boldsymbol{F}}_n(\boldsymbol{x}_i)$. The pseudo-log-likelihood is then defined as

$${}^p\ell_n(\theta) = \sum_{i=1}^n \log[c_\theta({}^p\boldsymbol{x}_i)].$$

As a univariate parameter θ is considered, the AIC is given by

$$\mathrm{AIC} = 2 \cdot {}^p\ell_n(\widehat{\theta}_n) - 2,$$

where $\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\ {}^p\ell_n(\theta)$ is the maximum pseudolikelihood estimator.

Leave-One-Out Copula Information Criterion

The selection procedure is based on the following quantity:

$$\mathrm{xv}_1 = \frac{1}{n}\sum_{i=1}^n \log\left[c_\theta\left(\widetilde{\boldsymbol{F}}_{(-i)}(\boldsymbol{x}_i)\right)\right]_{\theta=\widehat{\theta}_{(-i)}}, \text{ where}$$

- $\widehat{\theta}_{(-i)} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{j\neq i} \log\left[c_\theta\left(\widetilde{\boldsymbol{F}}_{(-i)}(\boldsymbol{x}_j)\right)\right]$,
- $\widetilde{\boldsymbol{F}}_{(-i)}(x_1, x_2) = \left(\widetilde{F}_{(-i),1}(x_1), \widetilde{F}_{(-i),2}(x_2)\right)$, where $\widetilde{F}_{(-i),k}$ is the $\frac{n-1}{n}$-rescaled empirical cdf of the kth marginal, computed from the sample $\mathcal{X}_n$ excluding $\boldsymbol{x}_i$, for $k = 1, 2$.

Full Simulation Study:

Email: kossumov@karlin.mff.cuni.cz

Other Copula Information Criteria

As computing xv_1 is computationally expensive, the authors of [1] recommend using $\mathrm{xv}_{\mathrm{CIC}}$, defined as:

$$\mathrm{xv}_{\mathrm{CIC}} = 2 \cdot \left({}^p\ell_n(\widehat{\theta}_n) - \widehat{p}_n - \widehat{q}_n - \widehat{r}_n\right),$$

where $\widehat{p}_n$, $\widehat{q}_n$, $\widehat{r}_n$ are bias-correcting terms whose explicit analytical forms can be found in Section 4 of [1].

Inspired by [3], we randomly draw, without replacement, a collection $\mathcal{T}_n$ of $b_n = O(n)$ subsets of $\{1, \ldots, n\}$, each of size n_v, such that $n_v/n \xrightarrow[n\to\infty]{} 1$. Here, the n_v observations are used for validation, while the remaining $n_c = n - n_v$ observations are used for parameter estimation. Denote by $s_v \in \mathcal{T}_n$ the set of indices corresponding to the n_v validation observations. Then define the following quantity:

$$\mathrm{xv}_{n_v} = \frac{1}{n_v b_n}\sum_{s_v\in\mathcal{T}_n}\sum_{i\in s_v} \log\left[c_\theta\left(\widetilde{\boldsymbol{F}}_{(-s_v)}(\boldsymbol{x}_i)\right)\right]_{\theta=\widehat{\theta}_{(-s_v)}}, \text{ where}$$

- $\widehat{\theta}_{(-s_v)} = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{j\notin s_v} \log\left[c_\theta\left(\widetilde{\boldsymbol{F}}_{(-s_v)}(\boldsymbol{x}_j)\right)\right]$,
- $\widetilde{\boldsymbol{F}}_{(-s_v)}(x_1, x_2) = \left(\widetilde{F}_{(-s_v),1}(x_1), \widetilde{F}_{(-s_v),2}(x_2)\right)$, where $\widetilde{F}_{(-s_v),k}$ is the $\frac{n_c}{n_c+1}$-rescaled empirical cdf of the kth marginal, computed from the sample $\mathcal{X}_n$ excluding $\{\boldsymbol{x}_i : i \in s_v\}$, for $k = 1, 2$.

Results

In this simulation study, the considered copula families C are parameterized using different values of Kendall's tau τ. For each combination of τ and sample size n, we conducted 1000 replications.

IC	Clayton	Gumbel	Joe	Frank	Gaussian
AIC	90.0 ± 1.9	49.6 ± 3.1	80.1 ± 2.5	59.1 ± 3	49.8 ± 3.1
xv_1	90.0 ± 1.9	49.5 ± 3.1	80.1 ± 2.5	59.1 ± 3	49.8 ± 3.1
$\mathrm{xv}_{\mathrm{CIC}}$	87.0 ± 2.1	45.5 ± 3.1	84.0 ± 2.3	61.2 ± 3	49.1 ± 3.1
xv_{n_v}	89.7 ± 1.9	52.4 ± 3.1	79.8 ± 2.5	59.8 ± 3	47.3 ± 3.1

Table 1: Hit rates ($n = 200$, $\tau = 0.20$) are shown with 95% confidence intervals, and all values are expressed as percentages.

From Table 1, one can see that for the intermediate sample size $n = 200$ and weak dependence $\tau = 0.2$, the considered criteria perform well only when the true copula is Clayton or Joe. In contrast, the Gumbel and Gaussian copulas appear to be more challenging for all criteria. Additionally, AIC and xv_1 perform very similarly in all cases—a pattern that holds across other combinations of τ and n (see Table 2).
For the Clayton copula, AIC and xv_1 outperform $\mathrm{xv}_{\mathrm{CIC}}$, whereas for Joe and Frank copulas, the opposite is true. Interestingly, when the true copula is Gumbel, the proposed xv_{n_v} criterion outperforms the others—a result that also holds for all other simulation scenarios. For more details, see the QR code.

	n	$\tau = 0.05$	$\tau = 0.1$	$\tau = 0.15$	$\tau = 0.2$	All
AIC & xv_1	100	99.74	99.92	99.90	99.96	99.88
AIC & xv_1	200	99.92	99.98	99.98	99.98	99.97
AIC & $\mathrm{xv}_{\mathrm{CIC}}$	100	79.26	86.46	89.10	89.34	86.04
AIC & $\mathrm{xv}_{\mathrm{CIC}}$	200	86.88	91.38	92.90	94.10	91.32
AIC & xv_{n_v}	100	47.76	67.84	80.44	87.40	70.86
AIC & xv_{n_v}	200	59.28	83.14	91.56	94.64	82.16

Table 2: Coincidence of AIC with cross-validation based information criteria, with all values expressed as percentages.

Table 2 shows the coincidence percentages of AIC with other information criteria across all considered copula families. It is seen that under weak dependence, AIC has the highest coincidence with xv_1 and the lowest with xv_{n_v}.

Conclusion

- The proposed method xv_{n_v} was still unable to beat the well-known AIC.
- For larger sample sizes or stronger dependence, all considered criteria are able to select the true copula model reliably.
- All criteria perform poorly under small sample sizes and weak dependence.
- Regardless of the sample size and the value of Kendall's tau, the most challenging copulas to identify for all criteria are Gaussian and Gumbel.
- As an interesting secondary finding, it was shown that for all considered values τ, the closest method to AIC is xv_1.
- Under weaker dependence $\tau \in \{0.05, 0.1, 0.15\}$, $\mathrm{xv}_{\mathrm{CIC}}$ is much closer to AIC than xv_{n_v}.
- In the specific case when the true copula model is Gumbel, the proposed xv_{n_v} outperformed the other criteria (in terms of hit rates and their confidence intervals) for all considered combinations of τ and n.

References

[1] S. Grønneberg and N. L. Hjort, "The copula information criteria," *Scandinavian Journal of Statistics*, vol. 41, pp. 436–459, 2014.

[2] L. A. Jordanger and D. Tjøstheim, "Model selection of copulas: AIC versus a cross validation copula information criterion," *Statistics and Probability Letters*, vol. 92, pp. 249–255, 2014.

[3] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, pp. 486–494, 1993.