Les Cahiers du GERAD ISSN: 0711–2440

Using BIRCH to Compute Approximate Rank Statistics on Massive Datasets

L. Charest
J.-F. Plante

G–2012–76

November 2012

Using BIRCH to Compute Approximate Rank Statistics on Massive Datasets

Lysiane Charest
Jean-François Plante

GERAD & Department of Management Sciences
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada, H3T 2A7

jfplante@hec.ca

November 2012

Les Cahiers du GERAD

G–2012–76

Abstract

The BIRCH algorithm (Balanced Iterative Reducing and Clustering Hierarchies) handles massive dataset by reading the data file only once, clustering the data as it is read, and retaining only a few clustering features to summarize the data read so far. Using BIRCH allows to analyze datasets that are too large to fit in the computer main memory. We propose estimates of Spearman's ρ and Kendall's τ that are calculated from a BIRCH output and assess their performance through Monte Carlo studies. The numerical results show that the BIRCH-based estimates can achieve the same efficiency as the usual estimates of ρ and τ while using only a fraction of the memory otherwise required.

Key Words: Correlation; Rank statistics; Massive dataset; Kendall's tau; Spearman's rho; BIRCH.

1 Introduction

The BIRCH algorithm (Balanced Iterative Reducing and Clustering Hierarchies) of Zhang [20] was developed to handle massive datasets that are too large to be contained in the main memory (RAM). To minimize I/O costs, every datum is read once and only once. Harrington and Salibián-Barrera [9] have implemented a version of BIRCH that is described in detail therein. Their code is available in the "birch" R package [2] and includes functions to compute approximate solutions to robust inference problems.

Spearman's ρ and Kendall's τ are coefficients of correlation that have desirable properties because they are based on ranks. Embrechts et al. [3] argue that for the estimation of the dependence structure, such multivariate rank statistics should be preferred to Pearson's correlation.

Computing rank statistics on massive datasets can be challenging, especially if the dataset is too large to fit in the main memory. The BIRCH algorithm builds clusters in the data using a limited amount a memory and a linear I/O cost. We use the BIRCH algorithm to compute approximate rank statistics by assuming equal ranks for the data within each of these clusters. The BIRCH output yields numerous ties, we therefore consider estimates of Spearman's ρ and Kendall's τ that are designed to handle ties.

The asymptotic properties of the BIRCH algorithm are not well known. The behavior of the BIRCH-based estimates is thus assessed through an extensive Monte Carlo study. As long as the tuning parameter of BIRCH (called the radius) is set to a reasonable value, we observe that the performance of BIRCH-based estimates is superior to the traditional estimates of ρ or τ calculated on a subsample that uses a larger amount of memory as BIRCH.

Section 2 gives an overview of how BIRCH handles massive datasets by creating compact clusters and how we use its output to compute approximate ranks. We also introduce five estimates of Spearman's ρ and Kendall's τ. Section 3 shows the results of Monte Carlo studies that evaluate the properties and performances of the estimates. We look at the general behavior of the estimates using three widely used families of copula, examine the effect of different margins on the performances of the BIRCH-based estimates, and examine the behavior of the estimates on multivariate datasets and massive datasets. Recommendations on the choice of radius for the BIRCH clustering step are formulated in Section 4. Simulations are used to support these recommendations.

2 Details of the algorithm

As the RAM of a computer is the only storage directly accessible to the CPU (processor), working with a dataset too big to fit here in increases the I/O costs of every calculation on this. To minimize I/O costs, the BIRCH algorithm reads the data file only once, one datum at a time, and allocates each datum to a cluster before reading the next one. The first datum is a cluster by itself, then subsequent steps determine the closest existing cluster based on the Euclidean distance. To decide whether the datum should be allocated to that cluster or become the first member of a new cluster, criteria of closeness and compactness are used. To optimize further memory usage, the clusters are contained in a tree structure. Detailed definitions of the criteria and structures are described by Harrington and Salibián-Barrera [9].

In the general BIRCH setting, the main memory of the computer is deemed insufficient to hold the whole dataset. The dataset can be stored in any other secondary storage such as hard disk drives or external databases. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ be the i^{th} datum from the dataset of size n and $\mathcal{L}_j$ the set of indices of the data in the j^{th} cluster created by BIRCH. To limit memory requirements, only a vector of three elements, the clusering features, is retained for each cluster,

$$CF_j = \left\{ \sum_{i \in \mathcal{L}_j} 1, \sum_{i \in \mathcal{L}_j} \mathbf{X}_i, \sum_{i \in \mathcal{L}_j} \mathbf{X}_i \mathbf{X}_i^\top \right\}.$$

The output of the BIRCH algorithm consists therefore in a list of m clusters with their CF_j. For models such as linear regression, the CF_j provide complete information for the fit. For calculating ranks, however, the loss of individual information means that data within each cluster will have to be treated as ties.

The most widely used approach for treating ties is the attribution of mid-ranks to all tied data, see e.g. [1]. To be more specific, the $N_j = \sum_{i\in\mathcal{L}_j} 1$ points in cluster $\mathcal{L}_j$ will be associated with a vector of ranks whose k^{th} component is equal to $(N_j+1)/2 + \sum_{i=1}^m N_i \mathbb{1}(\bar{X}_{ik} < \bar{X}_{jk})$ where $\bar{X}_{jk}$ is the k^{th} element of the vector of means $\bar{\mathbf{X}}_j = \sum_{i\in\mathcal{L}_j} \mathbf{X}_i/N_j$ and $\mathbb{1}(\cdot)$ is the indicator function. The distribution of the data is assumed continuous, hence the probability of ties between clusters is null.

Let us consider a bivariate dataset. The vector of ranks allocated to the N_i data points in cluster i will be noted (R_i, S_i). If each cluster were a singleton and there were no ties, R_i would correspond to the rank of X_{i1} in an ordered list of $X_{\bullet 1}$ and similarly for S_i and X_{i2}. This is equivalent to the classic problem where the usual estimate of Spearman [16] can be written as

$$\rho = \frac{\sum_{i=1}^n \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum_{i=1}^n \left(R_i - \bar{R}\right)^2}\sqrt{\sum_{i=1}^n \left(S_i - \bar{S}\right)^2}}$$

with $\bar{R} = \bar{S} = n(n+1)/2$ since there are no ties.

Under this classic paradigm, the τ of Kendall [11] can also be expressed as a function of the ranks, but is more often written as

$$\tau = \frac{C - D}{\binom{n}{2}}$$

where

$$\begin{aligned} C &= \sum_{1\le i<j\le n} \mathbb{1}\{\left(\bar{X}_{i1} - \bar{X}_{j1}\right)\left(\bar{X}_{i2} - \bar{X}_{j2}\right) > 0\} \\ D &= \sum_{1\le i<j\le n} \mathbb{1}\{\left(\bar{X}_{i1} - \bar{X}_{j1}\right)\left(\bar{X}_{i2} - \bar{X}_{j2}\right) < 0\} \end{aligned}$$

are the number of concordant and discordant pairs respectively.

Kendall [10] surveys strategies that lead to different estimates of ρ and τ in the presence of ties. Attributed to Woodbury [19], the estimates ρ_W and τ_W, are derived by averaging over all possible permutations of the ties. The estimates ρ_S and τ_S use the mid-rank method without further adjustments and are attributed to Student [18]. In addition,we also consider τ_C, an estimate proposed by Stuart [17] that is akin to τ_S except that an approximation is used for its denominator.

The fact that the output of BIRCH yields the same number of ties on both axes, simplifies some formulas. In this context, other estimates such as τ_a and τ_b (see e.g. [8]) are equivalent to τ_S.

With our notation the estimates considered are thus:

$$\begin{aligned} \rho_W &= 1 - \frac{\sum_{i=1}^m N_i\left(N_i^2 - 1\right) - 6\sum_{i=1}^m N_i(R_i - S_i)^2}{n\left(n^2-1\right)} \\ \rho_S &= 1 - \frac{6\sum_{i=1}^m N_i(R_i - S_i)^2}{n\left(n^2-1\right) - \sum_{i=1}^m N_i\left(N_i^2 - 1\right)} \\ \tau_W &= \frac{2C - \left\{\binom{n}{2} - \sum_{i=1}^m \binom{N_i}{2}\right\}}{\binom{n}{2}} \\ \tau_S &= \frac{2C - \left\{\binom{n}{2} - \sum_{i=1}^m \binom{N_i}{2}\right\}}{\left\{\binom{n}{2} - \sum_{i=1}^m \binom{N_i}{2}\right\}} \\ \tau_C &= \frac{2m\left[2C - \left\{\binom{n}{2} - \sum_{i=1}^m \binom{N_i}{2}\right\}\right]}{n^2\left(m-1\right)} \end{aligned}$$

where

$$C = \sum_{1\le i<j\le m} N_i N_j \mathbb{1}\{\left(\bar{X}_{i1} - \bar{X}_{j1}\right)\left(\bar{X}_{i2} - \bar{X}_{j2}\right) > 0\}$$

to take the ties into consideration.

The next section of this article propose a numerical study of the properties of these estimates when calculated from a BIRCH output on a large dataset. We limit the scope of our numerical endeavor to the estimation of ρ and τ, but further work could also look at the behavior of the ranks in other estimation problems. For instance, one could use mid-ranks in the pseudo-likelihood of Genest et al. [6] to fit a family of copulas to a massive dataset.

As pointed out by Harrington and Salibián-Barrera [9], clusters from a BIRCH output can overlap. As a consequence, a list of the data sorted according to the individual values of X may not result in an increasing sequence of ranks because of some inversions. The results of the next sections show that this behavior does not hinder the performance of the estimates.

3 Numerical results

We perform an extensive Monte Carlo experiment to assess the properties of the BIRCH-based estimates of ρ and τ.

The population values of ρ and τ are typically expressed as a function of the copula underlying the data. A copula is a cumulative distribution function whose margins are uniform on $[0, 1]$. All continuous distributions have a unique underlying copula that contains all information about their dependence structure. It is thus not surprising that copulas arise naturally in the population values of ρ and τ. For an introduction to copulas, see e.g. [15] or [7].

Because of their central role in dependence modeling, we use families of copulas in many of the simulations. Specifically, we use the Normal, Clayton and Gumbel-Hougaard copulas. Technical details about these families are described by Nelsen [13]. For numerical operations, including the generation of pseudo-random observations, we use the “copula” R package [12].

The three families of copulas considered have one parameter in bijective relation with Spearman’s correlation and admit the whole positive range of correlations. To make simulations comparable, the parameter of the copula is set to the value that yields a specific theoretical value of Spearman’s ρ. Therefore, the parameters are always expressed in terms of ρ. The values are obtained using the calibration function available in the “copula” R package. The function returns a moment estimate of the parameter using numerical approximation techniques when there are no closed-form expressions for Kendall’s τ or Spearman’s ρ.

The creation of new clusters in BIRCH is controlled by tuning parameters. We use the same approach as Harrington and Salibián-Barrera [9] for the closeness and compactness criteria of the BIRCH algorithm. A single tuning parameter, the radius, is used to trigger the creation of a new cluster in the algorithm. In our context, this parameter can be thought of as a bandwidth: smaller values of the radius yield a larger number of clusters in the output, hence a better resolution. More details on the choice of the radius and its impact on the results are given in the next section.

When facing a massive dataset too large to be held in the memory, the typical solution consists in drawing a simple random sample from the dataset hence discarding a (possibly large) part of the data. BIRCH allows to use information from all the data points in the analysis. Although allocating the same ranks to the data within each cluster of BIRCH is expected to cause some bias, the fact that more data contribute to this approximation is likely to reduce the variance. As a consequence, mean squared error (MSE) will be the preferred measure of performance since it takes into consideration both bias and variance.

3.1 Global assessment

We first consider scenarios with different correlations and different radii to describe the general behavior of the BIRCH-based estimates of ρ and τ.

Datasets of size 10000 are generated, a small size in the context of BIRCH which is nonetheless sufficient to study the properties of the BIRCH-based estimates. Choosing a larger number would multiply the running time of the simulations without providing a sizable advantage.

Data are simulated from three families of copulas, the Normal, Clayton and Gumbel-Hougaard families, and their parameter was set to yield a Speaman's correlation of 0, 0.3, 0.6 and 0.9.

The number of clusters that BIRCH yields depends on its tuning parameter, the radius. The "birch" R package [2] uses the squared radius as an argument. We thus consider eight different values for the squared radius ranging from 10^{-5} to 0.1 and report the average number of clusters produced by BIRCH for each simulated scenario.

Each choice of radius, copula and correlation forms a scenario. Each scenario is simulated 1000 times. The five estimates of ρ and τ defined in the previous section are calculated, as well as the usual estimates of ρ and τ on the whole sample.

Table 1 displays the results of the simulation for the Normal copula. The column $\bar{m}$ shows the average number of clusters produced by BIRCH for each scenario. The usual estimates of ρ and τ are asymptotically unbiased and we verified numerically that their MSE is due almost exclusively to the variance of the estimate, even for sample sizes as small as 100 data points.

The values in Table 1 are relative values of the MSE corresponding to 100 times the MSE of the usual estimate of ρ (or τ) on the whole sample divided by the MSE of the proposed estimate evaluated on the BIRCH output. We will refer to these numbers as relative efficiencies (RE).

Table 2 and 3 display the same results as Table 1 for the Clayton and Gumbel-Hougaard copulas respectively. Note that the case $\rho = 0$ corresponds to the independent copula which belongs to the three families considered. We arbitrarily chose to include these results in Table 1 only.

Table 1: MSE of the different estimates expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample. Samples of size 10000 are simulated from a Normal copula under different scenarios. Each figure corresponds to an average over 1000 repetitions.

	Radius^2	$\bar{m}$	ρ_W	ρ_S	τ_W	τ_S	τ_C
$\rho = 0$	0.00001	8910	100.0	100.0	100.0	100.0	100.0
	0.0001	4490	100.0	100.0	100.0	100.0	100.0
	0.0005	1420	100.0	100.0	99.9	99.8	99.8
	0.001	812	98.9	98.9	99.3	99.0	99.1
	0.005	147	85.0	85.0	81.1	79.8	80.0
	0.01	66	47.7	47.6	42.4	41.1	41.1
	0.05	15	1.9	1.9	1.7	1.4	1.4
	0.1	8	0.4	0.4	0.4	0.3	0.3
$\rho = 0.3$	0.00001	8820	100.0	100.0	100.0	100.0	100.0
	0.0001	4350	100.0	100.0	99.9	99.8	99.8
	0.0005	1410	99.3	99.3	99.1	98.3	98.6
	0.001	764	97.2	97.2	96.3	94.1	94.8
	0.005	137	57.0	56.9	49.1	37.6	39.3
	0.01	67	20.4	20.3	16.7	11.6	12.1
	0.05	16	0.7	0.7	0.7	0.4	0.4
	0.1	9	0.2	0.2	0.2	0.1	0.1
$\rho = 0.6$	0.00001	8520	100.0	100.0	100.0	100.0	100.0
	0.0001	3970	99.9	99.9	99.8	99.4	99.6
	0.0005	1310	98.4	98.4	96.1	91.5	93.8
	0.001	726	93.7	93.7	86.9	74.5	80.1
	0.005	130	31.9	31.5	18.6	10.0	12.0
	0.01	67	10.6	10.3	6.2	3.1	3.7
	0.05	17	0.6	0.5	0.5	0.2	0.2
	0.1	9	.02	0.2	0.2	0.1	0.1
$\rho = 0.9$	0.00001	7490	100.0	100.0	100.0	99.7	99.9
	0.0001	2820	99.8	99.8	97.8	91.0	95.5
	0.0005	886	89.2	88.8	74.0	39.1	56.9
	0.001	461	66.0	64.9	45.8	15.7	27.2
	0.005	94	9.6	8.6	7.2	1.4	2.5
	0.01	55	3.8	3.1	3.4	0.6	1.0
	0.05	15	0.6	0.3	0.8	0.1	0.1
	0.1	8	0.3	0.1	0.4	0.02	0.04

Table 2: MSE of the different estimates expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample. Samples of size 10000 are simulated from a Clayton copula under different scenarios. Each figure corresponds to an average over 1000 repetitions.

	Radius^2	$\bar{m}$	ρ_W	ρ_S	τ_W	τ_S	τ_C
$\rho = 0.3$	0.00001	8770	100.0	100.0	100.0	100.0	100.0
	0.0001	4320	100.0	100.0	99.9	99.8	99.9
	0.0005	1400	100.0	100.0	98.9	98.1	98.5
	0.001	787	99.4	99.4	96.5	94.4	95.2
	0.005	139	73.0	72.8	57.1	44.7	47.0
	0.01	67	31.5	31.3	23.1	16.0	16.8
	0.05	16	1.4	1.4	1.3	0.8	0.8
	0.1	9	0.3	0.3	0.3	0.2	0.2
$\rho = 0.6$	0.00001	8370	100.0	100.0	100.0	100.0	100.0
	0.0001	3860	99.8	99.8	99.7	99.3	99.6
	0.0005	1280	98.7	98.6	98.0	93.7	96.4
	0.001	721	95.5	95.4	93.7	82.6	89.0
	0.005	131	55.5	54.6	46.2	21.5	28.4
	0.01	67	23.3	22.5	18.2	7.3	9.6
	0.05	17	1.4	1.3	1.3	0.5	0.6
	0.1	10	0.4	0.3	0.4	0.1	0.2
$\rho = 0.9$	0.00001	7150	100.0	100.0	100.0	99.7	99.9
	0.0001	2650	99.6	99.6	98.9	91.9	97.3
	0.0005	844	92.8	92.3	87.8	45.8	72.3
	0.001	498	78.7	77.3	67.4	21.5	44.5
	0.005	93	15.3	13.2	12.8	1.9	4.2
	0.01	54	5.6	4.4	5.3	0.7	1.6
	0.05	16	0.7	0.4	1.1	0.1	0.2
	0.1	9	0.6	0.2	1.4	0.04	0.1

Table 3: MSE of the different estimates expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample. Samples of size 10000 are simulated from a Gumbel-Hougaard copula under different scenarios. Each figure corresponds to an average over 1000 repetitions.

	Radius^2	$\bar{m}$	ρ_W	ρ_S	τ_W	τ_S	τ_C
$\rho = 0.3$	0.00001	8800	100.0	100.0	100.0	100.0	100.0
	0.0001	4380	100.0	100.0	99.9	99.7	99.8
	0.0005	1400	101.0	101.0	98.8	97.9	98.3
	0.001	790	102.0	102.0	96.2	93.9	94.7
	0.005	139	79.0	78.9	53.0	41.2	43.2
	0.01	67	32.1	31.9	21.8	14.9	15.6
	0.05	16	1.2	1.1	1.0	0.6	0.7
	0.1	9	0.3	0.2	0.3	0.2	0.2
$\rho = 0.6$	0.00001	8420	100.0	100.0	100.0	99.9	100.0
	0.0001	3940	99.5	99.5	99.6	99.1	99.4
	0.0005	1300	97.0	96.9	96.9	92.4	95.0
	0.001	752	91.7	91.6	89.9	78.1	84.4
	0.005	133	38.8	38.2	29.5	14.7	18.6
	0.01	68	14.7	14.3	11.0	4.8	6.0
	0.05	17	0.9	0.8	0.8	0.3	0.4
	0.1	10	0.3	0.2	0.3	0.1	0.1
$\rho = 0.9$	0.00001	7310	100.0	100.0	100.0	99.8	100.0
	0.0001	2770	99.0	99.0	98.4	92.2	96.8
	0.0005	905	87.8	87.4	78.2	42.4	63.6
	0.001	530	69.2	68.1	54.4	18.6	35.3
	0.005	104	12.3	10.9	9.2	1.7	3.5
	0.01	59	5.0	4.1	4.5	0.7	1.4
	0.05	18	0.8	0.4	1.1	0.1	0.2
	0.1	10	0.5	0.2	1.0	0.04	0.1

It appears that the strength in the structure of the data has an effect on the performance of the estimates. Whether we compare similar radii or similar numbers of clusters, the performance of the estimates degrades as the theoretical correlation increases. For the three families of copulas, the best performances are achieved under independence.

The relative efficiency of each estimate stays very high even for small number of clusters. The relative efficiency of the estimates of Spearman's ρ remains above 50%e even when the average number of clusters ($\bar{m}$) reaches values as small as about 1.5% of the total number of observations. For the estimates of Kendall's τ, the relative efficiency degrades faster but with as few as 500 clusters, a relative efficiency above 50% is observed in almost all scenarios for at least one estimate.

Relative efficiency is often interpreted as the ratio of sample sizes that are required to achieve comparable performances with the two estimates. Comparing the same estimate on samples of size 5000 and 10000 should thus yield about 50%. We numerically verified this interpretation for Spearman's ρ and Kendall's τ under the different simulated scenarios and obtained values ranging from 40% to 55%.

Based on this interpretation, the simulations show that with a number of clusters as little as 1.5% of the sample size (5% for Kendall's τ), the BIRCH-based estimates can achieve performances at least as good as an estimate based on 50% of the dataset. In terms of memory usage optimization, this is a huge gain as the memory may not even be able to accommodate 50% of the dataset, but be large enough to retain the clustering features of the BIRCH clusters.

Looking at Spearman's ρ estimates, we see that ρ_W achieves slightly better performances than ρ_S, and does so systematically. Therefore, ρ_W seems to be a better choice. For Kendall's τ, the estimate τ_W presents the best performance overall while τ_C performs better than τ_S for high correlations.

For the Clayton copula, we observe the same patterns, except that the estimates seem to achieve a better performance in terms of their MSE.

The performance for the Gumbel-Hougaard copula seems to be slightly lower than for the Clayton, but it appears to be better than for the Normal copula.

As mentioned in the previous section, it is expected that tied ranks in each cluster will introduce some bias. Table 4 displays the bias of the considered estimates when the data follow a Normal copula. The bias figures are multiplied by 1000 for improved readability. Again, the column $\bar{m}$ shows the average number of clusters produced by BIRCH for each scenario. Table 5 and Table 6 display the same results for the Clayton and Gumbel-Hougaard copulas respectively.

It appears that the strength of the correlation affects the performance of the estimates. However, no clear pattern between the theoretical value of ρ and the bias of the estimates arises and different behaviors are observed for different radii.

The first section of Table 4 displays results obtained under the independence copula. The bias of each estimator in this context remains small, even with very few clusters. For instance, choosing a radius of 0.1 yields an average of 8 clusters to summarize the 10000 data points. Nonetheless, the bias is about 0.004 and 0.002 respectively for the estimates of Spearman's ρ and Kendall's τ.

The bias of all estimates of Kendall's τ is always positive, meaning that correlation is overestimated in all the simulated scenarios. For Spearman's ρ, exceptions arise for smaller correlations under the Clayton and Gumbel-Hougaard copulas.

Overall ρ_W achieves a smaller bias than ρ_S, but for small correlations or small radii, their performances often tie. The advantage of ρ_W over ρ_S shows more for a large correlation and a small number of clusters.

For Kendall's estimates, τ_W systematically features the lowest bias and τ_S the highest bias, except under independence where the three estimates perform equally well.

In summary, with an appropriate choice of radius, BIRCH-based estimates feature performances comparable to using the whole sample, while using only a fraction of the memory that would otherwise be required. In many cases, the performance does not degrade significantly until the number of clusters falls below 10% of the original sample size.

When comparing the different estimates, ρ_W and τ_W are the clear winners as they uniformly outperform the other contenders in terms of MSE and bias. As a consequence, further simulations will only report results for these two estimates.

Table 4: Bias (×1000) of the different estimates. Samples of size 10000 are simulated from a Normal copula under different scenarios. Each figure corresponds to an average over 1000 repetitions.

	Radius^2	$\bar{m}$	ρ_W	ρ_S	τ_W	τ_S	τ_C
$\rho = 0$	0.00001	8910	0.19	0.19	0.13	0.13	0.13
	0.0001	4490	0.19	0.19	0.13	0.13	0.13
	0.0005	1420	0.19	0.19	0.12	0.13	0.12
	0.001	812	0.21	0.21	0.15	0.15	0.15
	0.005	147	0.28	0.28	0.11	0.11	0.11
	0.01	66	1.3	1.3	0.9	0.91	0.91
	0.05	15	3.5	3.5	2.5	2.7	2.7
	0.1	8	4	4	1.8	2	2
$\rho = 0.3$	0.00001	8820	0.19	0.19	0.15	0.16	0.16
	0.0001	4350	0.24	0.24	0.19	0.24	0.22
	0.0005	1410	0.86	0.86	0.46	0.65	0.59
	0.001	764	1.1	1.1	0.88	1.2	1.1
	0.005	137	6.7	6.7	5.3	7.1	6.8
	0.01	67	15	15	11	15	14
	0.05	16	87	89	61	81	79
	0.1	9	180	190	120	170	160
$\rho = 0.6$	0.00001	8520	0.098	0.098	0.13	0.15	0.14
	0.0001	3970	0.19	0.19	0.24	0.37	0.3
	0.0005	1310	0.77	0.77	0.92	1.5	1.2
	0.001	726	1.6	1.6	1.9	2.9	2.5
	0.005	130	8.9	9	9.9	15	13
	0.01	67	18	18	18	28	25
	0.05	17	80	85	69	110	100
	0.1	9	140	160	110	190	170
$\rho = 0.9$	0.00001	7490	0.0013	0.0014	0.087	0.15	0.11
	0.0001	2820	0.11	0.11	0.39	0.84	0.58
	0.0005	886	0.72	0.73	1.6	3.4	2.4
	0.001	461	1.5	1.5	2.9	6.4	4.5
	0.005	94	6.1	6.5	8.8	23	17
	0.01	55	9.9	11	12	37	26
	0.05	15	25	36	22	100	73
	0.1	8	37	66	28	180	140

3.2 Effect of different margins

Important properties of rank statistics arise from their invariance to monotone transformations of the marginal distributions. The BIRCH algorithm, however, does not share this invariance as a change in the margins will affect the clustering of the data. To evaluate the effect of such transformations, we simulate data with a Normal copula and different marginal distributions.

It is known that changes of scales have an effect on the output of BIRCH. Harrington and Salibián-Barrera [9] suggest to use a first pass of the BIRCH algorithm to determine the standard deviation of each variable. For subsequent passes, data can be normalized appropriately. This strategy is reflected in our choice of distributions since we only consider scenarios where marginal distributions are on similar scales.

The data of the simulations in Section 3.1 are generated from copulas, meaning that their marginal distributions are uniform on $[0, 1]$. We generate data from a multivariate Normal distribution, i.e. with the same dependence structure as the Normal copula of Section 3.1, but with Normal marginals. Ideally, the effect of this change should be mild.

To test the limitations of BIRCH, we also simulate pseudo-observations with Cauchy margins. The Cauchy distribution generates a large number of extreme values. As a consequence, BIRCH will generate many singleton clusters along the axes to capture these extreme values, leaving fewer clusters to represent the core of the data. In such a challenging setting, one should expect a loss of performance, but we want to see if the method holds reasonably well.

To measure the performance of the estimates, MSE and bias are evaluated in the same terms as in Section 3.1. Even if the same nominal values were used for the squared radius, comparisons for the different

Table 5: Bias (×1000) of the different estimates. Samples of size 10000 are simulated from a Clayton copula under different scenarios. Each figure corresponds to an average over 1000 repetitions.

	Radius^2	$\bar{m}$	ρ_W	ρ_S	τ_W	τ_S	τ_C
$\rho = 0.3$	0.00001	8770	-0.51	-0.51	0.12	0.12	0.12
	0.0001	4320	-0.46	-0.46	0.18	0.23	0.2
	0.0005	1400	-0.11	-0.11	0.5	0.71	0.63
	0.001	787	0.37	0.37	0.93	1.3	1.2
	0.005	139	4.6	4.7	4.7	6.5	6.2
	0.01	67	11	11	9.3	13	12
	0.05	16	58	60	39	58	56
	0.1	9	140	140	88	130	130
$\rho = 0.6$	0.00001	8370	0.4	0.4	0.1	0.12	0.11
	0.0001	3860	0.47	0.47	0.22	0.38	0.29
	0.0005	1280	0.89	0.89	0.7	1.4	1
	0.001	721	1.4	1.5	1.3	2.5	1.8
	0.005	131	5.8	5.9	5.1	10	8.4
	0.01	67	11	11	8.9	19	16
	0.05	17	47	52	34	74	63
	0.1	10	95	110	64	140	120
$\rho = 0.9$	0.00001	7150	0.056	0.056	0.038	0.13	0.068
	0.0001	2650	0.21	0.22	0.31	0.92	0.51
	0.0005	844	0.71	0.73	1.2	3.5	2
	0.001	498	1.3	1.4	2.1	6.2	3.6
	0.005	93	5.6	6.2	6.9	23	15
	0.01	54	9.6	11	10	38	25
	0.05	16	28	41	21	110	72
	0.1	9	30	61	8.2	160	100

Table 6: Bias (×1000) of the different estimates. Samples of size 10000 are simulated from a Gumbel-Hougaard copula under different scenarios. Each figure corresponds to an average over 1000 repetitions.

	Radius^2	$\bar{m}$	ρ_W	ρ_S	τ_W	τ_S	τ_C
$\rho = 0.3$	0.00001	8800	-2	-2	0.2	0.21	0.21
	0.0001	4380	-1.9	-1.9	0.25	0.3	0.28
	0.0005	1400	-1.6	-1.6	0.58	0.78	0.71
	0.001	790	-1.1	-1.1	1	1.4	1.3
	0.005	139	3.9	4	5.2	7	6.7
	0.01	67	11	11	10	14	13
	0.05	16	71	73	50	69	67
	0.1	9	160	170	100	150	140
$\rho = 0.6$	0.00001	8420	1	1	0.15	0.17	0.16
	0.0001	3940	1.1	1.1	0.27	0.42	0.34
	0.0005	1300	1.6	1.6	0.89	1.5	1.2
	0.001	752	2.3	2.3	1.7	2.8	2.2
	0.005	133	8.5	8.6	7.8	13	11
	0.01	68	16	16	14	24	21
	0.05	17	70	75	55	95	84
	0.1	10	130	140	88	160	150
$\rho = 0.9$	0.00001	7310	0.2	0.2	0.023	0.097	0.05
	0.0001	2770	0.32	0.32	0.32	0.85	0.51
	0.0005	905	0.92	0.94	1.5	3.6	2.3
	0.001	530	1.6	1.7	2.7	6.5	4.1
	0.005	104	6.3	6.8	8.7	24	16
	0.01	59	10	11	12	37	25
	0.05	18	26	36	22	100	67
	0.1	10	30	56	13	150	97

marginal distributions would not be straightforward because the choice of margins has an effect on the number of clusters. A fair comparison should look at the performance of the estimates when they use a similar amount of memory, i.e. produce the same number of clusters.

An approximate conversion factor can also be found by expressing the radius as a multiple of a measure of dispersion such as the inter-quartile range (IQR). With the three marginals considered, a radius for copula R_{Copula}, a radius for Normal margins R_{Normal} and a radius for Cauchy margins R_{Cauchy} would be approximately equivalent if they respect the equation

$$R = \frac{R_{\text{Copula}}}{1/2} = \frac{R_{\text{Normal}}}{1.34898} = \frac{R_{\text{Cauchy}}}{2}. \tag{1}$$

Table 7 presents the MSE and bias of ρ_W and τ_W for different correlations and different radii when the data are generated from a multivariate Normal distribution. Table 8 presents the same results for a distribution that has a Normal copula and Cauchy margins.

For multivariate Normal data, the average number of clusters produced by the BIRCH algorithm seems more affected for large correlations. The performance of the estimates is similar, but possibly slightly inferior, to that observed when the data come from a Normal copula (with uniform marginal distributions). The methods keep performing well for fairly large radii that lead to a small number of clusters. The performance in terms of bias is also comparable to that for the Normal copula.

The fact that the Cauchy distribution causes numerous extreme values close to the axes means that BIRCH has a harder time capturing the structure of the data. Despite these difficulties, Table 8 shows reasonable results for a large range of radii. While for other marginal distributions, excellent results are achieved with as few as 100 clusters, a breakdown seems to occur shortly after $\bar{m}$ falls below 1500 or so with Cauchy margins. Before that point, however, the performances are excellent. The BIRCH-based estimate therefore manages

Table 7: MSE and bias of ρ_W and τ_W. The MSE is expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample. The bias is multiplied by 1000 for improved readability. Samples of size 10000 are simulated from a Normal distribution under different scenarios. Each figure corresponds to an average over 1000 repetitions.

			RE		1000×Bias	
	Radius^2	$\bar{m}$	ρ_W	τ_W	ρ_W	τ_W
$\rho = 0.3$	0.00001	9900	100.0	100.0	-0.55	-0.36
	0.0001	9120	100.0	100.0	-0.46	-0.3
	0.0005	6870	100.0	100.0	-0.51	-0.35
	0.001	5420	99.9	100.0	0.53	0.4
	0.005	2350	99.9	99.8	-0.077	-0.027
	0.01	1500	99.2	98.8	0.56	0.42
	0.05	428	88.2	83.0	2.5	1.7
	0.1	232	61.2	50.9	4.6	3.4
	0.5	61	7.5	5.0	14	7.6
	1	39	2.7	1.9	15	6.8
	5	12	0.4	0.4	62	34
$\rho = 0.6$	0.00001	9880	100.0	100.0	0.094	0.12
	0.0001	8950	100.0	100.0	0.1	0.1
	0.0005	6480	100.0	100.0	-0.58	-0.4
	0.001	5010	100.0	99.9	-0.14	-0.052
	0.005	2080	99.7	99.4	0.31	0.29
	0.01	13.0	98.1	97.20	0.85	0.7
	0.05	362	65.8	58.1	4	3.4
	0.1	186	30.6	24.7	8.5	7
	0.5	55	3.6	3.0	22	13
	1	35	1.5	1.4	33	19
	5	8	0.1	0.1	180	150
$\rho = 0.9$	0.00001	9780	100.0	100.0	0.14	0.26
	0.0001	8220	100.0	100.0	0.0095	0.089
	0.0005	5130	100.0	99.8	-0.0011	0.088
	0.001	3680	100.0	100.0	-0.03	0.036
	0.005	1310	88.9	85.2	0.78	1.2
	0.01	789	73.3	66.2	1.2	1.9
	0.05	204	12.1	10.4	5.7	7.6
	0.1	119	4.1	4.0	9.4	11
	0.5	36	0.5	0.5	28	33
	1	18	0.1	0.1	58	86
	5	6	4.2	5.0	7.9	-7.3

Table 8: MSE and bias of ρ_W and τ_W. The MSE is expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample. The bias is multiplied by 1000 for improved readability. Samples of size 10000 are simulated from a distribution with a Normal copula and Cauchy margins under different scenarios. Each figure corresponds to an average over 1000 repetitions.

			RE		1000×Bias	
	Radius^2	$\bar{m}$	ρ_W	τ_W	ρ_W	τ_W
$\rho = 0.3$	0.00001	9970	100.0	100.0	0.058	0.057
	0.0001	9720	100.0	100.0	-0.033	0.0051
	0.0005	8840	100.0	100.0	-0.15	-0.075
	0.001	8090	100.0	100.0	0.43	0.34
	0.005	5710	99.9	99.9	0.24	0.19
	0.01	4700	99.9	99.9	0.17	0.16
	0.05	2800	97.4	96.6	1.1	0.81
	0.1	2190	95.5	91.0	1.4	1
	0.5	1200	42.7	28.6	6.3	3.9
	1	912	17.6	10.8	11	6.9
	5	478	0.9	0.8	74	48
$\rho = 0.6$	0.00001	9960	100.0	100.0	-0.39	-0.27
	0.0001	9650	100.0	100.0	0.18	0.19
	0.0005	8610	100.0	100.0	0.11	0.14
	0.001	7770	100.0	100.0	-0.046	-0.011
	0.005	5340	100.0	99.8	-0.42	-0.28
	0.01	4360	99.8	99.8	0.18	0.17
	0.05	2580	95.9	93.9	1.3	1.1
	0.1	2030	82.0	78.1	2.8	2.1
	0.5	1140	17.6	13.7	12	8.5
	1	884	5.7	4.8	22	15
	5	486	0.4	0.4	94	72
$\rho = 0.9$	0.00001	9930	100.0	100.0	-0.15	-0.13
	0.0001	9340	100.0	100.0	-0.16	-0.15
	0.0005	7720	100.0	100.0	-0.029	0.037
	0.001	6680	100.0	100.0	-0.061	0.0014
	0.005	4300	99.3	98.4	0.2	0.36
	0.01	3460	97.0	95.6	0.43	0.65
	0.05	2060	42.3	36.7	2.4	3.3
	0.1	1640	16.4	15.2	4.5	5.6
	0.5	973	1.9	2.0	14	15
	1	770	0.6	0.8	25	28
	5	439	0.2	0.2	47	63

to yield acceptable results even in a situation where BIRCH is seriously challenged. The performance of the method degrades faster with a high correlation, but even then, performances are acceptable for a squared radius as big as 0.01 which corresponds to 5% of the IQR of the marginal.

We do not report squared radii beyond 5 in Table 8 because the radius is then large enough to cover a large part of the distribution (the IQR of the margins equals 2). Simulations that were run show that for exaggeratedly large radii, the RE tends to 0. This is already observed in Table 8 for a radius of 1 or 5. With very large radii, the core of the data is likely to be in a single cluster with individual clusters capturing extreme values along the axes. Essentially, information about the dependence structure of the data is then lost.

Spearman's ρ and Kendall's τ draw properties from their invariance to marginal transformations. The BIRCH-based estimates do not retain these theoretical properties, but the bias shown in Table 7 and 8 is small for all reasonable values of the radius. Even though the estimates are not theoretically invariant to monotone transformations of the data, they feature some robustness to such changes, meaning that they remain good choices for inference from a massive dataset.

3.3 One BIRCH for many correlations

The simulations performed so far focused on a single correlation and the BIRCH algorithm was performed on the two variables of interest. In practice, a single pass of BIRCH may be used on more dimensions. The resulting clustering can be used to estimate the correlation between two of the variables, or the complete correlation matrix.

To link our next simulation to a real-life context, we use the famous Iris dataset from Fisher [4] which was downloaded from the UCI Machine Learning Repository [5]. The variables sepal length, sepal width, petal length and petal width are measured in centimeters for 50 Iris Setosa, 50 Iris Versicolour and 50 Iris Virginica. We arbitrarily chose to focus on the Iris Setosa and used estimates of its mean and covariance as the parameters for our simulation. Note that the marginal distributions of the four characteristics of the Iris Setosa have standard deviations that range from 0.107 to 0.381. Because these values are of relatively similar magnitude, we decided to proceed directly with BIRCH rather than considering a two-stage approach where the data are first normalized.

In this simulation, the BIRCH algorithm is applied to the four dimensions of the data at once. The clusters will have to represent the behavior of the data on all four dimensions. For an equal number of clusters, it would not be surprising to observe reduced performances when comparing to results where BIRCH is applied to the two variables of interest only.

Samples of 10000 four-variate Normal distributions are generated and the BIRCH algorithm is performed for different values of the squared radius that range from 0.001 to 0.1.

Table 9 displays the MSE (expressed as an RE) and average bias of ρ_W and τ_W for estimating the correlation matrix of the data. For all scenarios and both estimates, the RE is quite good with a reasonable choice of radius, which we will discuss later. However, performances are not as good as the previous simulations where the clusters produced by BIRCH summarized the information about the two variables of interest only. With more dimensions, the clusters have to summarize more information and cannot achieve the same level of performance for all pairs of two variables with the same number of clusters.

A specific correlation may be of interest even if BIRCH is applied to a multivariate dataset. Equations 2 and 3 show the MSE and bias of ρ_W and τ_W for estimating each correlation individually using a squared radius of 0.01. Since the correlation matrix is symmetric, the upper triangular half displays the performance of ρ_W and the lower triangular half that of τ_W.

$$\begin{bmatrix} & RE(\rho_W) \\ RE(\tau_W) & \end{bmatrix} = \begin{bmatrix} & 59.1 & 81.4 & 42.6 \\ 55.4 & & 89.6 & 44.1 \\ 80.5 & 88.9 & & 28.8 \\ 41.0 & 42.4 & 27.2 & \end{bmatrix} \tag{2}$$

$$1000 \begin{bmatrix} & Bias(\rho_W) \\ Bias(\tau_W) & \end{bmatrix} = \begin{bmatrix} & 4.2 & 4.2 & 10 \\ 4.0 & & 2.8 & 10 \\ 2.9 & 1.9 & & 14 \\ 7.3 & 7.2 & 10 & \end{bmatrix} \tag{3}$$

We note that the performances of ρ_W and τ_W are quite similar, especially for the RE. Depending on which correlation is of interest, the performance may be excellent or mild, both in terms of MSE and bias.

Table 9: MSE and average bias for estimating the Spearman and Kendall correlation matrices with ρ_W and τ_W. The MSE is expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample. The bias is multiplied by 1000 for improved readability. Samples of size 10000 are simulated from a four-variate multivariate Normal distribution whose parameters correspond to the characteristics of Iris Setosa. Each figure corresponds to an average over 1000 repetitions and 4 variables.

		RE		1000×Bias	
Radius^2	$\bar{m}$	ρ_W	τ_W	ρ_W	τ_W
0.001	9420	99.9	99.9	0.21	0.16
0.005	5210	90.6	89.4	1.7	1.3
0.01	2900	57.6	55.9	5.7	4.2
0.02	1340	28.8	28.3	13	9.4
0.05	386	9.0	8.7	34	24
0.1	135	3.3	3.1	87	63

The RE remains high except for the correlation between petal length and petal width that features the worst performance. With standard deviations of 0.174 and 0.107 respectively, these two variables feature the smallest variances among the four characteristics. Choosing a radius of $\sqrt{0.01} = 0.1$, the resolution of BIRCH is not fine enough to capture all the information about the structure of the data between these two variables.

With a small improvement in the resolution, this behavior disappears. For a squared radius of 0.005, the RE of the correlation between petal length and petal width is 77.6 for ρ_W and 75.4 for τ_W. Even though those are the lowest RE for that radius, their nominal values are very high compared to what could be done with a sample of 50% of the data (which would requiere approximately the same amount of memory than the $\bar{m} = 5210$ clusters used).

Overall, the performance of the BIRCH-based estimates make them a viable option for analyzing a massive multivariate dataset, even if a single multivariate BIRCH clustering is used to estimate many correlations.

3.4 Massive dataset

All simulations so far have used a sample size of 10000. In this section, we vary the sample size up to 2×10^8 to verify the behavior of the BIRCH based estimates on larger datasets. Note that the default functions and settings of R can hardly handle samples of size 10^7 and beyond, but with BIRCH, these sizes pose no problems.

To see how the estimates evolve with an increase of the sample size, we simulate five datasets from a Normal distribution with $\rho = 0.6$. The radius for BIRCH is set to 10% of the IQR of the marginal distributions which seems to be a reasonable choice based on previous simulations, yielding a squared radius of 0.01820. The BIRCH algorithm is suspended at different sample sizes in order to calculate ρ_W and τ_W for all data read so far. Figure 1 shows the evolution of the two estimates. The horizontal line corresponds to the true value of the parameter.

We note that the five lines tend to converge together, showing that the variance of the estimates decreases as the sample size increases. That convergence appears to occur above the horizontal line, meaning that the estimates are biased even for large sample sizes. This is expected since the chosen radius defines a resolution beyond which the details of the data are lost and cannot be recovered from increasing the sample size.

As a consequence of this bias, the MSE of the estimates will have a lower bound beyond which further improvements will not be possible, unless if the radius is replaced with a smaller value. The choice of radius

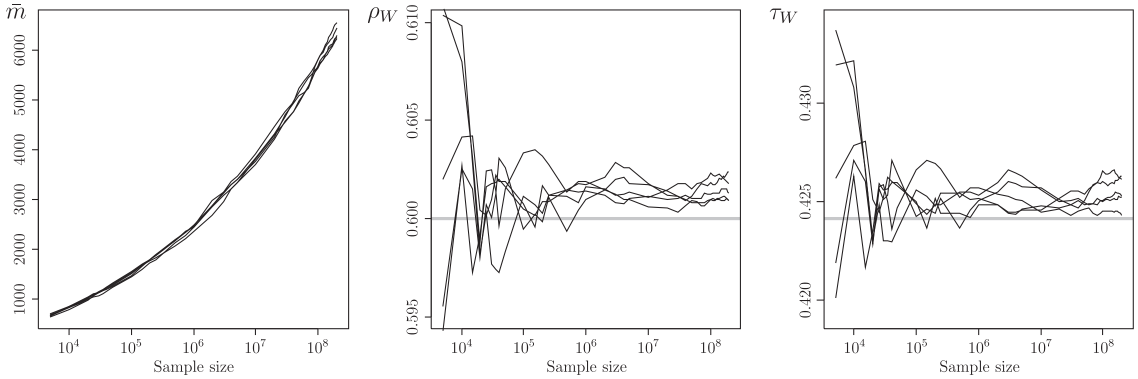

Figure 1: Evolution of ρ_W and τ_W as more data are read from five different datasets, each represented by a line. The data are generated from a Normal distribution with $\rho = 0.6$. The squared radius is set to 0.01820. The horizontal line shows the true value of ρ and τ. The evolution of the number of clusters is shown on the left-most panel. Note the logarithmic scale for sample size.

determines the number of clusters, thus the memory requirements. This means that the memory limitations will constrain our ability to reduce the bias of the estimates to an arbitrarily small value.

Five samples are not enough to provide reliable estimates of the bias and mean squared errors. For sample sizes 10^6 and 10^7 we generate 1000 samples from the same Normal distribution on which we calculate ρ_W and τ_W. The results are summarized in Table 10. We also include in the table what we call an equivalent sample size (ESS). As discussed in Section 3.1, the RE of an estimate can be interpreted as a ratio of the sample sizes required to achieve a similar performance. The ESS uses this interpretation to determine what size of a sample is required to match the performance of the BIRCH-based estimates here.

For both sample sizes, the RE may seem low, but the ESS is very large compared to $\bar{m}$. BIRCH therefore manages to summarize a large part of the information contained in the sample with a small amount of memory.

The bias is fairly low in absolute terms and does not decrease a lot when moving from $n = 10^6$ to $n = 10^7$. This behavior, expected from the observations on Figure 1, may explain why the ESS increases by only about 50% while the sample size is multiplied by 10 in this case. As a matter of fact, a squared bias of 1.23×10^{-6} is observed for ρ_W with $n = 10^7$ where the MSE is 1.50×10^{-6}. The MSE is thus driven by the bias at that stage. We observed on Figure 1 that the bias does not seem to converge to 0. If a bias of 10^{-6} persists asymptotically, the MSE will never fall below 10^{-12}, yielding an ESS that cannot exceed 430000 in this case, even if the sample size goes to infinity.

As samples become massive, they will not fit within main memory. Using BIRCH allows to produce estimates with an ESS that is larger than the maximum sample size that could fit in memory since we observed that $ESS > \bar{m}$, even for the large datasets of Table 10 where the ratio between the two numbers is quite impressive. Within the limitations of the computer, BIRCH can yield a better estimate than taking a random subsample.

Table 10: MSE and bias of ρ_W and τ_W. The MSE is expressed in terms of the relative efficiency (RE) of the BIRCH-based estimates compared to the usual estimates on the whole sample and in terms of an equivalent sample size. The bias is multiplied by 1000 for improved readability. Samples of different sizes n are simulated from a Normal distribution with $\rho = 0.6$. The squared radius was set to 0.01820 for BIRCH. Each figure corresponds to an average over 1000 repetitions.

		MSE as RE		MSE as ESS		1000×Bias	
n	$\bar{m}$	ρ_W	τ_W	ρ_W	τ_W	ρ_W	τ_W
10^6	2491	19.0	16.0	190000	160000	1.2	0.97
10^7	3867	2.8	2.5	280000	250000	1.1	0.85

4 Choice of the radius

Harrington and Salibián-Barrera [9] mention that the choice of radius is scale dependent. Their suggestion is to choose the radius based on the number of clusters produced by BIRCH. We agree with this approach, but note that the marginal distribution of the data has an effect on the link between the number of clusters and the sample size of the database. Figures 2 to 4 help in understanding this link and provide additional guidance to make a wise choice of radius. On massive datasets, some attention must also be paid to the stability of BIRCH, which we discuss first.

4.1 Stability of BIRCH for small radii and massive datasets

Running BIRCH on massive datasets unveiled a peculiar behavior. The BIRCH algorithm uses a tree structure to speeds up calculations. When a certain number of clusters (called leafs in the context of a tree) is attained (the default is $MAXL = 100$ in the birch package [2]), branches are created and the leafs are assigned to them. The next datum is processed through the tree structure, following the closest branch before being compared to the leaves of that branch only. The centroids of the leafs and branches get updated. When the number of branches reaches a predefined limit (the default is $MAXB = 100$ in the birch package [2]), a new level of branches is created and connected to the existing branches that are connected themselves to the leaves. Further levels are possible as well and get created as needed.

Since the branches get updated dynamically, their centroids wobble as data are processed. Even if an appropriate leaf exists for a new datum, the wobbling can cause the wrong branch to be closest to the datum, hence rooting the latter to a branch where a new leaf will unnecessarily be created.

The movements of the branches can be negligible if the radius is large, but when a massive dataset is analyzed with a small radius, we sometimes observe sudden explosions in the number of leafs, which are explained by the behavior described above. None of the examples in this manuscript feature this instability. In particular, such instability was never observed with fewer than 10000 leafs, or with a radius of 0.1IQR or more. Our recommendation is thus to aim for a moderate number of leafs, or to select a somewhat large radius. Otherwise, it would be well advised to monitor the construction of the BIRCH tree and be cautious.

4.2 Number of clusters versus sample size

To trace Figures 2 to 4, the BIRCH algorithm was applied to large datasets and the number of clusters was recorded as the algorithm progressed through the data. The radii are expressed as a fraction of the IQR. They can be recovered from Equation 1 by replacing R with 0.05, 0.075, 0.1 and 0.15. Five datasets were used for each choice of radius. The line in the plots correspond to the average number of clusters for those five datasets.

Since they have an effect of the tree structure, the internal parameters $MAXL$ and $MAXB$ discussed above have an influence on the stability of BIRCH. For a radius equal to 0.05IQR and default values for $MAXL$ and $MAXB$, we observed datasets where the number of clusters exploded after the first few tens of millions of observations. All figures in this section were thus produced with $MAXL = 40$ and $MAXB = 40$ which did not feature instability problems. We also tried $MAXL = 200000$, a large enough value to ensure that all leafs stay within one branch. The plots for the number of clusters obtained were visually identical to those displayed here.

If data are bounded (e.g. copulas yield random numbers on the unit square), it is expected that for a fixed radius, the clusters will eventually cover the whole range of the data and no new clusters will be proposed even if the sample size keeps increasing.

Figure 2 was produced by simulating samples of increasing sizes from a Normal copula with $\rho = 0.6$. Every line corresponds to the average of five datasets to obtain a smoother line. For every radius considered, the number of clusters seems to stall at some point, as expected.

At the opposite of the spectrum, the Cauchy has the potential to produce extreme values that are unbounded. Figure 3 displays the number of clusters as a function of the sample size when the data are generated from a distribution with a Normal copula ($\rho = 0.6$) and margins. Extreme values along the axes are so sparse that they each become a singleton cluster. It is thus not surprising to see the number of clusters increase approximately linearly with the sample size. With Cauchy margins (an extreme case which is unlikely to occur in reality), the main memory of the computer is quickly filled to capacity.

In real-life situations, data are probably not bounded, nor full of extremes. In that sense, Figure 4 is more representative of reality as it presents the number of clusters as a function of the sample size when the data come from a multivariate normal distribution with $\rho = 0.6$. In this case, the number of clusters should not be theoretically bounded as for any given clustering, a large enough sample will make probable the observation of a datum outside the covered area, meaning that the number of clusters can become arbitrarily large.

It is however reinsuring to see on Figure 4 that the increase in the number of clusters slows down rapidly and almost stalls.

The choice of a radius is linked to the available memory. If there are very many extremes (e.g. Cauchy), one has to be more careful to avoid running out of memory. Otherwise, the choice of radius is not strongly

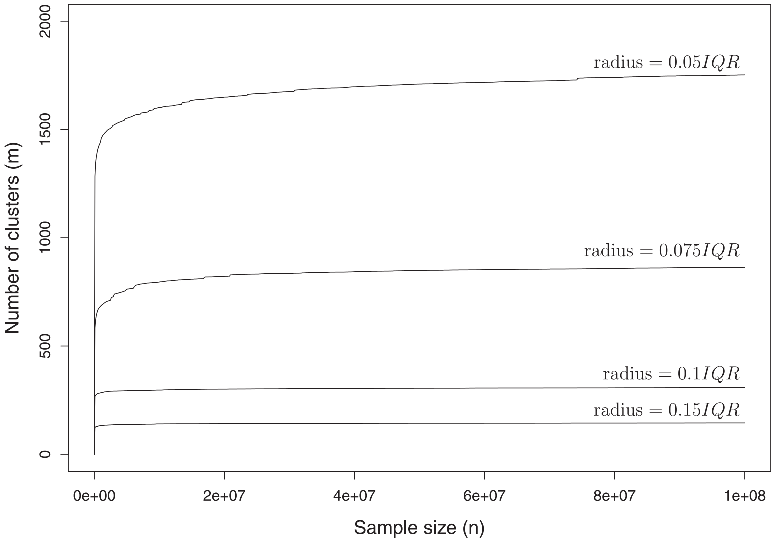

Figure 2: Number of clusters produced by the BIRCH algorithm as a function of the sample size. The data are generated from a Normal copula distribution with $\rho = 0.6$ and different values of the radius are considered. Each line corresponds to an average on five datasets.

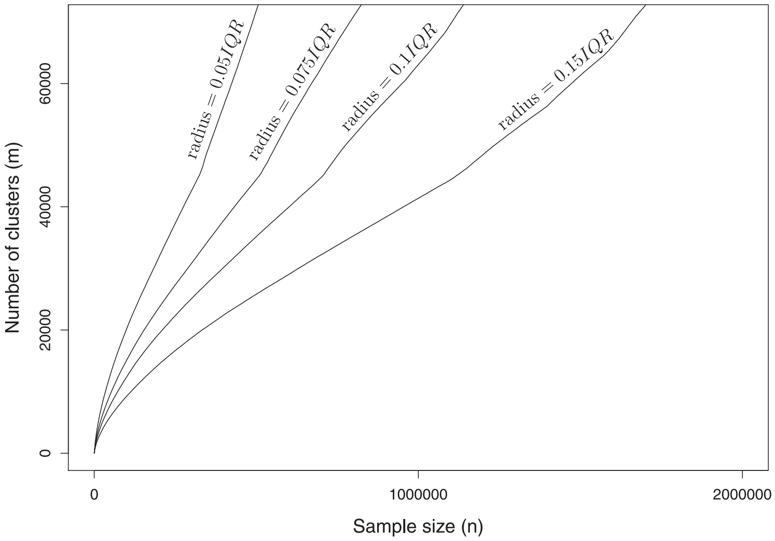

Figure 3: Number of clusters produced by the BIRCH algorithm as a function of the sample size. The data are generated from a distribution with a Normal copula ($\rho = 0.6$) and Cauchy margins. Different values of the radius are considered. Each line corresponds to an average on five datasets.

linked to the size of the database. This fact also opens the door to faster and less memory demanding preprocessing strategies as the estimation of the scale and of the number of clusters could be performed using only a part of the whole sample, even for the evaluation of $\bar{m}$ and the choice of radius.

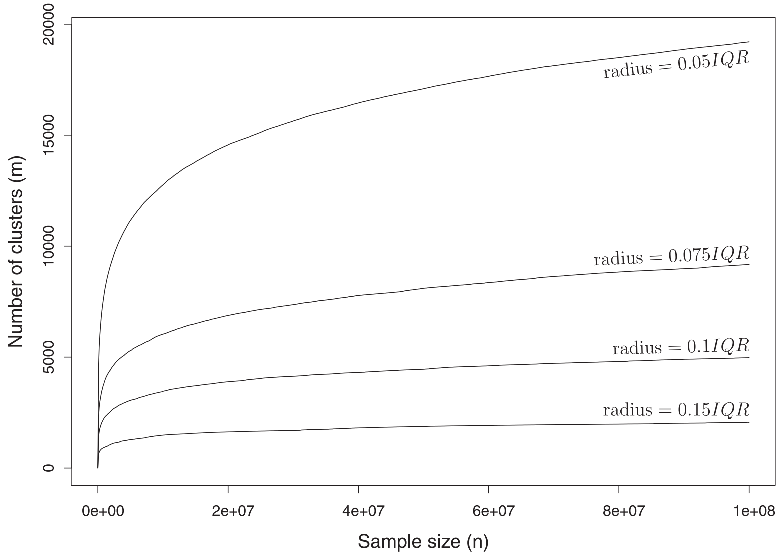

Figure 4: Number of clusters produced by the BIRCH algorithm as a function of the sample size. The data are generated from a multivariate Normal distribution with $\rho = 0.6$ and different values of the radius are considered. Each line corresponds to an average on five datasets.

5 Conclusion

The BIRCH algorithm is designed to handle massive datasets with linear I/O costs by reading every datum only once and summarizing them into clusters. We use the output of BIRCH to calculate approximate rank statistics, specifically estimates of Spearman's ρ and Kendall's τ.

Different estimates that take into consideration the ties produced by BIRCH were considered, but ρ_W and τ_W introduced by Woodbury [19] uniformly performed best when compared to the other contenders.

Rank statistics are popular because of their invariance to marginal transformations of the data. BIRCH does not share this property, but simulations showed that the estimates have some robustness to variations in the margins. The BIRCH-based statistics thus estimate a quantity that is invariant to marginal transformations, but those transformations can affect the performance of the estimation procedure.

In all simulations performed, we expressed the MSE of the estimates in terms of RE, the relative efficiency of the BIRCH-based estimates compared to the classic estimate of ρ or τ evaluated on the whole sample. For all reasonable choices of radii, we observed that RE remains high even with fairly small numbers of clusters, meaning that when the main memory is limited, using BIRCH can lead to a better estimate than using the largest possible subsample that fits into memory. The ratio between ESS and $\bar{m}$ is especially impressive for massive datasets where ESS reached values more than 70 times larger than $\bar{m}$.

The BIRCH algorithm is available as a R package from CRAN [2]. The estimates of ρ and τ considered in this paper are available therein.

This paper focused on the estimation of ρ and τ, but other rank statistics could be approximated by the strategy that we used. The mid-ranks that we defined could be used in the pseudo-likelihood of Genest et al. [6] to infer the parameter of a copula. If the asymptotic properties of BIRCH were developed, an empirical estimate of the copula based on BIRCH could also be defined and used in different testing problems. For instance, the equality of copulas could be tested on massive datasets by adapting the test of Rémillard and Scaillet [14].

References

[1] P. Capéraà and B. Cutsem. *Méthodes et modèles en statistique non paramétrique: exposé fondamental*. Presses de l'Université Laval, 1988. ISBN 9782040165437. URL http://books.google.ca/books?id=LKrA16AwLb0C.

[2] L. Charest, J. Harrington, and M. Salibian-Barrera. *birch: Dealing with very large datasets using BIRCH*, 2011. R package version 1.2-1.

[3] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: Properties and pitfalls. In *Risk Management: Value at risk and beyond*, pages 176–223. Cambridge University Press, 1999.

[4] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7(Part II):179–188, 1936.

[5] A. Frank and A. Asuncion. Uci machine learning repository. irvine, ca: University of california, school ofinformation and computer science., 2010.

[6] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995. doi: 10.1093/biomet/82.3.543. URL http://biomet.oxfordjournals.org/content/82/3/543.abstract.

[7] C. Genest and A.-C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368, 2007.

[8] C. Genest and J. Nešlehová. A primer on copulas for count data. *Astin Bull.*, 37(2):475–515, 2007.

[9] J. Harrington and M. Salibián-Barrera. Finding approximate solutions to combinatorial problems with very large data sets using birch. *Comput. Stat. Data Anal.*, 54:655–667, March 2010. ISSN 0167-9473. doi: 10.1016/j.csda.2008.08.001. URL http://dl.acm.org/citation.cfm?id=1660181.1660741.

[10] M.G. Kendall, The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239. URL http://biomet.oxfordjournals.org/content/33/3/239.short.

[11] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81. URL http://biomet.oxfordjournals.org/content/30/1-2/81.short.

[12] I. Kojadinovic and J. Yan. Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20, 2010. URL http://www.jstatsoft.org/v34/i09/.

[13] R.B. Nelsen. *An introduction to copulas*. Springer series in statistics. Springer, 2006. ISBN 9780387286594. URL http://books.google.ca/books?id=B3ONT5rBv0wC.

[14] B. Rémillard and O. Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100: 377–386, 2009.

[15] B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9:879–885, 1981.

[16] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):pp. 72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159.

[17] A. Stuart. The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 40(1/2): 105 –110, 1953. ISSN 00063444. URL http://www.jstor.org/stable/2333101.

[18] Student. An experimental determination of the probable error of Dr. Spearman's correlation coefficients. *Biometrika*, 13(2-3):263–282, 1921. doi: 10.1093/biomet/13.2-3.263. URL http://biomet.oxfordjournals.org/content/13/2-3/263.short.

[19] M.A. Woodbury. Rank correlation when there are equal variates. *The Annals of Mathematical Statistics*, 11(3): pp. 358–362, 1940. ISSN 00034851. URL http://www.jstor.org/stable/2235684.

[20] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. *SIGMOD Rec.*, 25:103–114, June 1996. ISSN 0163-5808. doi: http://doi.acm.org/10.1145/235968.233324. URL http://doi.acm.org/10.1145/235968.233324.