Additional Results

Analogously to Section 3 in the manuscript, we analyze the performance of Algorithm 1 to select an appropriate number of clusters for a given data set. In this case, we consider the K-means algorithm as the clustering technique[1] As before, given validity measure (**VM**) and a Bregman divergence (**BD**), we consider as baseline running the clustering algorithm (**Cl**) for $K \in \{2, \ldots, 50\}$ and keeping the number of clusters that leads to the best VM[2](**Baseline_Cl**). In addition, we consider running Cl for an increasing number of clusters K and stopping if, after s modifications of the number of clusters, where $s \in \{1, 2, 5, 10\}$, the validity measure, computed via Theorem 1, is not improved (**Cl_s**). Lastly, we also run Cl, for all $K \in \{2, \ldots, 50\}$ and compute VM, as in the previous case, via Theorem 1 (**Cl_all**). In particular, we consider the following Bregman divergences: i) Itakura-Saito divergence (**IS**) and ii) Kullback-Leibler divergence (**KL**)[3]. In terms of the validity measures, we consider i) the Silhouette index (**SH**) and ii) the Davies-Bouldin (**DB**) and iii) Caliński-Harabasz (**CH**) measures, and additionally we compute running times (**RT**).

As discussed in the manuscript, DB, CH and RT are normalized as follows: i) relative $\mathrm{DB}(M) = \frac{\max_{M' \in \mathcal{M}} \mathrm{DB}(M')}{\mathrm{DB}(M)}$, ii) relative $\mathrm{CH}(M) = \frac{\mathrm{CH}(M)}{\min_{M' \in \mathcal{M}} \mathrm{CH}(M')}$ and iii) relative $\mathrm{RT}(M) = \frac{\mathrm{RT}(M)}{\min_{M' \in \mathcal{M}} \mathrm{RT}(M')}$, where M is an algorithm included in the set of considered methods $\mathcal{M}$. In this sense, observe that SH, relative CH and relative DB must be maximized and relative RT should be minimized. In Figs.1-4, we present the obtained results in terms of these measures when maximizing VM following the previosuly mentioned methodologies.

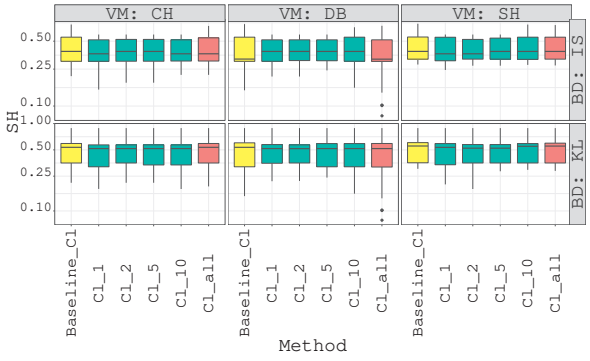

Figure 1: Silhouette index for all data sets and Bregman divergences considered.

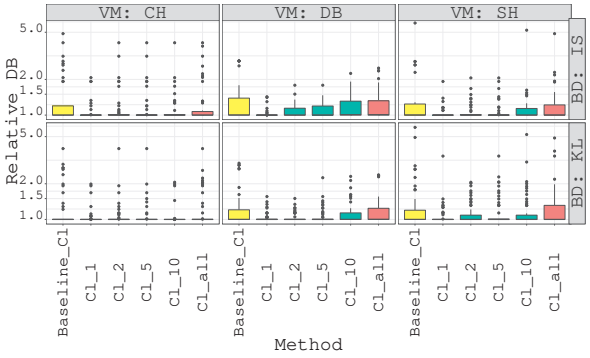

Figure 2: Relative Davies-Bouldin measure for all data sets and Bregman divergences considered.

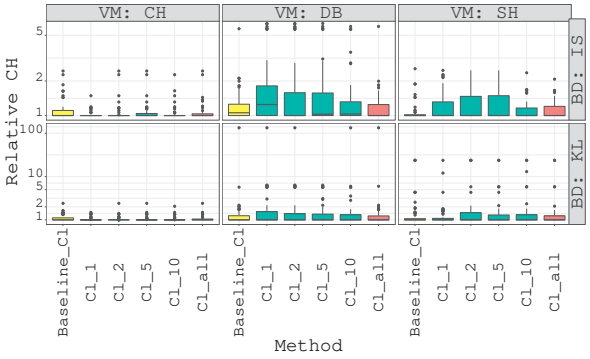

Figure 3: Relative Caliński-Harabasz measure for all data sets and Bregman divergences considered.

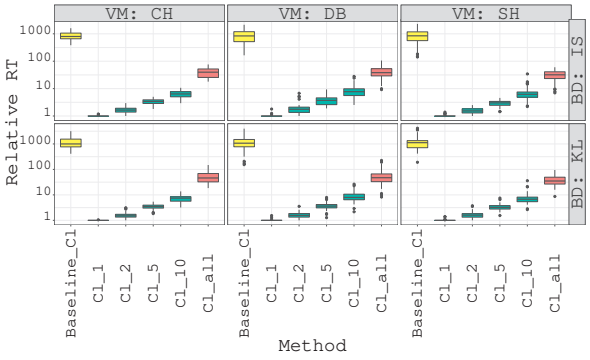

Figure 4: Relative running times for all data sets and Bregman divergences considered.

At first glance, we observe that regardless of the VM index being maximized, the clustering obtained by all Bregman divergences considered have similar SH, DB and CH values. As expected, all clusterig quality measures obtained by Baseline_Cl are of the same order as those obtained by Cl_all, for all experimental settings. More importantly, Cl_1 already achieved fairly similar, and, in some cases, better quality measures than Baseline_Cl, e.g., observe the results obtained when the validity measure optimized is Davies-Bouldin for the Itakura-Saito divergence.

Even though all methods provide fairly similar SH indexes, we observe that, for Cl_s, the obtained quality measure improves as s increases, e.g., for the Kullback-Leibler divergence, the average SH value obtained by Baseline_Cl is 0.52, while for Cl_1 and Cl_10, this value is 0.49 and 0.51, respectively. Finally, for Cl_all, the average SH is 0.52. The same behavior, in general,

[1] At this point it must be highilighted that, as described in Banerjee et al. (2005), the K-means approach can be used, in general, as a centroid-based clustering approach for all Bregman divergences.

[2] In this case, VM is evaluated using the `Scikit-Learn` versions of the given index.

[3] In order to make all data sets, shown in Table 2 of the manuscript, suitable for both IS and KL, we have made the following changes: For IS, if there is an entry of the dataset D that is non-positive, we add to all the entries of the data set $|\min_{\mathbf{x} \in D,\ i \in \{1,\ldots,d\}} x_i| + \epsilon$, where $\varepsilon \ll 1$. For KL, besides the modification we just commented, for each in instance $\mathbf{x} = (x_1, \ldots, x_d) \in D$, we re-define each entry as $x_i = x_i / \sum_{j=1}^{d} x_j$.

is also observed for the other quality measures. For instance, for the Davies-Bouldin measure: for Baseline_Cl the average relative DB is 1.23, while for Cl_1, Cl_10 and Cl_all, we have 1.09, 1.15 and 1.22, respectively.

In terms of the relative RT, we observe the main benefit of our proposal. Regardless of the Bregman divergence considered, we can see a staggering reduction of running time with respect to the baseline. When the Bregman divergence considered is the Itakura-Saito divergence, the average relative RT of Baseline_Cl ascends to 892.24, while, for Cl_1, Cl_10 and Cl_all, these values are just 1.01, 7.61 and 38.62, respectively. On the other time, for the Kullback-Leibler divergence, the relative RT of Baseline_Cl, Cl_1, Cl_10 and Cl_all are 1257.31, 1.01,.8.10 and 52.08, respectively. In other words, Cl_1 was able to be 1243.36 times faster than Baseline_Cl, while converging to solutions of fairly similar clustering quality to Baseline_Cl, measured by SH, DB and CH.

References

Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J., 2005. Clustering with bregman divergences. Journal of machine learning research 6, 1705–1749.