Supplementary Material

This file is divided into two sections. In Section 1, we comment on additional theoretical results to those presented in the manuscript. Furthermore, in Section 2, we present complementary experimental results in which we show the performance of the proposed methodology for other Bregman divergences.

1. Additional Theoretical Results

In this section, we verify that one can reduce the running times of computing other validity measures following similar strategies to those presented in Section 2.1. of the manuscript. In that section, we tackled the most commonly used internal validity measures in the literature (Van Craenendonck and Blockeel, 2015): the Silhouette index, the Davies-Bauldin measure and the Caliński-Harabasz measure.

The Dunn index (Dunn, 1974) is an internal evaluation scheme commonly used for evaluating the performance of a given clustering algorithm. This index is a ratio in which, in the numerator, the minimum dispersion among clusters is measured and, in the denominator, the maximum compactness among clusters is computed. Hence, this metric must be maximized. In the next result, we re-write the Dunn index in terms of the within-cluster sum of errors:

Theorem 1. *Given a clustering $\mathcal{P} = \{P_1, \ldots, P_K\}$, the Dunn index is defined as $di(\mathcal{P}) = \min_{i \neq j} s(\boldsymbol{c}_i, \boldsymbol{c}_j) / \max_{k \in \{1,\ldots,K\}} \triangle_k$, where $\triangle_k = 2 \cdot [|P_k| \cdot (|P_k| - 1)]^{-1} \cdot \sum_{\boldsymbol{x},\boldsymbol{y} \in P_k} s(\boldsymbol{x}, \boldsymbol{y})$. If $s(\cdot, \cdot)$ is symmetric, then $\triangle_k = 2 \cdot [|P_k| - 1]^{-1} \cdot E^{P_k}(\{\boldsymbol{c}_k\})$, where $\boldsymbol{c}_k = \overline{P_k}$.*

Proof. If we assume the Bregman divergence $s(\cdot, \cdot)$ to be symmetric, then the two following equalities hold

$$\sum_{\mathbf{y} \in P_k} s(\mathbf{c}_k, \mathbf{y}) = E^{P_k}(\{\mathbf{c}_k\}), \quad \sum_{\mathbf{x},\mathbf{y} \in P_k,\ \mathbf{x} \neq \mathbf{y}} s(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \cdot \sum_{\mathbf{x} \in P_k} \sum_{\mathbf{y} \in P_k} s(\mathbf{x}, \mathbf{y}) \quad (1)$$

Taking this into account as well as Eq.4 in the manuscript, we have that

$$\begin{aligned}
\triangle_k &= 2 \cdot [|P_k| \cdot (|P_k| - 1)]^{-1} \cdot \sum_{\mathbf{x},\mathbf{y} \in P_k,\ \mathbf{x} \neq \mathbf{y}} s(\mathbf{x}, \mathbf{y}) \\
&= [|P_k| \cdot (|P_k| - 1)]^{-1} \cdot \sum_{\mathbf{y} \in P_k} \sum_{\mathbf{x} \in P_k} s(\mathbf{x}, \mathbf{y}) \\
&= [|P_k| \cdot (|P_k| - 1)]^{-1} \cdot \sum_{\mathbf{y} \in P_k} [E^{P_k}(\{\mathbf{c}_k\}) + |P_k| \cdot s(\mathbf{c}_k, \mathbf{y})] \\
&= [|P_k| - 1]^{-1} \cdot [E^{P_k}(\{\mathbf{c}_k\}) + \sum_{\mathbf{y} \in P_k} s(\mathbf{c}_k, \mathbf{y})] \\
&= 2 \cdot [|P_k| - 1]^{-1} \cdot E^{P_k}(\{\mathbf{c}_k\}) \qquad (2)
\end{aligned}$$

□

Observe that the computational complexity of the Dunn index is, in general, is $O(n^2)$. However, by introducing the symmetry assumption on the divergence $s(\cdot, \cdot)$, one can compute the Dunn index in just $O(K^2)$ time. Note that this assumption is satisfied by some of the most commonly used Bregman divergence, e.g., the squared Euclidean distance and the Mahalanobis distance.

In addition to the previous validity measure index, we have other indexes that are of a similar nature to the Davies-Bouldin and the Caliński-Harabasz measures and that are also commonly used in practice. For instance, we have the well-known WB index (Zhao and Fränti, 2014) and the PBM index (Pakhira et al., 2004):

Definition 1. *Given a clustering $\mathcal{P} = \{P_1, \ldots, P_K\}$ and let $\boldsymbol{c}_k$ be the centroid of P_k, i.e., $\boldsymbol{c}_k = \overline{P_k}$, for all $k \in \{1, \ldots, K\}$, then the WB index is defined as $wb(\mathcal{P}) = \frac{\sum_{k=1}^{K} |P_k| \cdot s(\boldsymbol{c}_k, \overline{X})}{K \cdot E^X(C)}$ and the PBM index is given by $pbm(\mathcal{P}) = \frac{E^X(\{\overline{X}\}) \cdot \max_{i \neq j} s(\boldsymbol{c}_i, \boldsymbol{c}_j)}{K \cdot E^X(C)}$, where $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$.*

Note that both indexes measure the dispersion of the clusters in the numerator and the within-cluster sum of squares in the denominator. Hence, these indexes must be maximized. More importantly, computing both indexes from scratch has a $O(n \cdot K)$ time cost. If we instead, make use of the output of the partitional clustering algorithm and the fact that $\sum_{k=1}^{K} |P_k| \cdot s(\mathbf{c}_k, \overline{X}) = E^X(\{\overline{X}\}) - E^X(C)$, for all Bregman divergences, then the WB index can be computed in $O(n)$ time and, the PBM index, in $O(\max\{n, K^2\})$ time. Similar ideas can be developed for other validity measures and dissimilarity measures by determining the relation between the within-cluster error with respect to the prototype and with respect to any instance in the space.

2. Additional Empirical 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

[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 $x_i = x_i / \sum_{j=1}^{d} x_j$.

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.

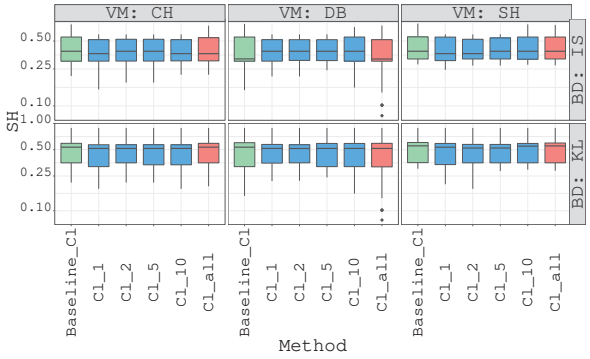

Figure 1: Silhouette index for all data sets and Bregman divergences considered.

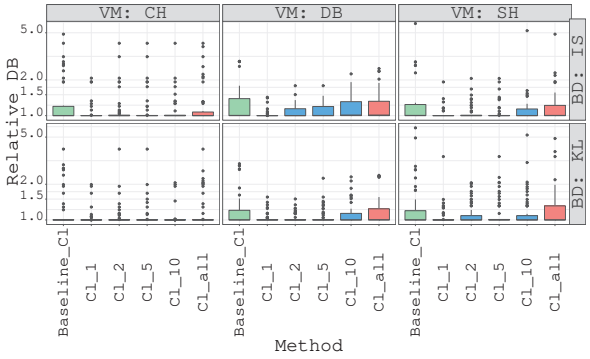

Figure 2: Relative Davies-Bouldin measure for all data sets and Bregman divergences considered.

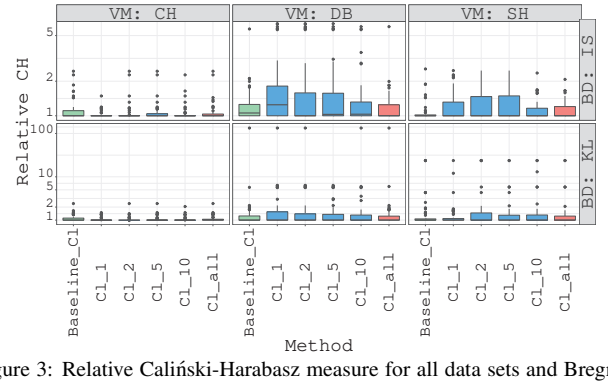

Figure 3: Relative Caliński-Harabasz measure for all data sets and Bregman divergences considered.

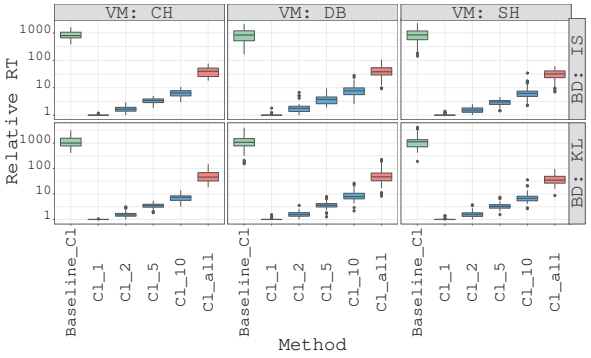

Figure 4: Relative running times for all data sets and Bregman divergences considered.

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, 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.

Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics 4, 95–104.

Pakhira, M.K., Bandyopadhyay, S., Maulik, U., 2004. Validity index for crisp and fuzzy clusters. Pattern recognition 37, 487–501.

Van Craenendonck, T., Blockeel, H., 2015. Using internal validity measures to compare clustering algorithms. Benelearn 2015 Poster presentations (online) , 1–8.

Zhao, Q., Fränti, P., 2014. Wb-index: A sum-of-squares based index for cluster validity. Data & Knowledge Engineering 92, 77–89.