Web-based Supplementary Materials for Clustering and Classification of Genetic Data Through U-Statistics by G.B. Cybis, M. Valk and S.R.C. Lopes

1 Variance of B_n

This section contains technical details related to the derivations of the homogeneity test of Section 2.2.

First, we shall consider the variance of the B_n -statistic, defined in (7). Assume that under H_0, X_1, \dots, X_n are i.i.d. and we assume $\mu \equiv \mathbb{E}(X_1)$. Additionally we note that B_n is a first-order degenerated U-statistics in the Hoeffding's sense, see [2]. Under this, we can show that the first term in the H-decomposition is null. Thus B_n can be written as a function only of the second term of this decomposition. Let $\phi_E(X_1, X_2) = (X_1 - X_2)'(X_1 - X_2)$ be the Hoeffding's decomposition, where $\phi_E(\cdot, \cdot)$ is the kernel in the expression (2.2) when the Euclidean distance is considered. Its conditional expectation with respect to X_1 is given by

$$\phi_{1}(X_{1}) = \mathbb{E}[(X_{1} - X_{2})'(X_{1} - X_{2})|X_{1}] = \mathbb{E}[X_{1}'X_{1} - 2X_{1}'X_{2} + X_{2}'X_{2}|X_{1}]
= X_{1}'X_{1} - 2X_{1}'\mathbb{E}[X_{2}|X_{1}] + \mathbb{E}[X_{2}'X_{2}|X_{1}] = X_{1}'X_{1} - 2X_{1}'\mu + \mathbb{E}[X_{2}'X_{2}]
= X_{1}'X_{1} - 2X_{1}'\mu + \mu'\mu + \operatorname{tr}(\Sigma) = (X_{1} - \mu)'(X_{1} - \mu) + \operatorname{tr}(\Sigma),$$
(1)

where Σ is the covariance matrix of X and $\operatorname{tr}(\Sigma)$ is the trace of matrix Σ . Similarly, the conditional expectation with respect to X_2 is given by $\phi_1(X_2) = (X_2 - \mu)'(X_2 - \mu) + \operatorname{tr}(\Sigma)$. Also $\phi_0 = \mathbb{E}(\phi_E(X_1, X_2))$ is given by

$$\phi_{0} = \mathbb{E}[(X_{1} - X_{2})'(X_{1} - X_{2})] = \mathbb{E}[X_{1}'X_{1} - 2X_{1}'X_{2} + X_{2}'X_{2}]$$

$$= \mathbb{E}[X_{1}'X_{1}] - 2\mathbb{E}[X_{1}'X_{2}] + \mathbb{E}[X_{2}'X_{2}] = 2\mathrm{tr}(\Sigma) + 2\mu'\mu - 2\left(\sum_{i=1}^{d} \mathrm{Cov}(X_{1i}, X_{2i}) + \mu'\mu\right)$$

$$= 2\mathrm{tr}(\Sigma), \qquad (2)$$

and $\phi_2(X_1, X_2) = \mathbb{E}(\phi_E(X_1, X_2)|X_1, X_2) = (X_1 - X_2)'(X_1 - X_2)$. The second term of the Hoeffding's decomposition of $\phi_E(\cdot, \cdot)$ is given by

$$\psi_{2}(X_{1}, X_{2}) = \phi_{2}(X_{1}, X_{2}) - \phi_{1}(X_{1}) - \phi_{1}(X_{2}) + \phi_{0} = (X_{1} - X_{2})'(X_{1} - X_{2}) -\{(X_{1} - \mu)'(X_{1} - \mu) + \operatorname{tr}(\Sigma)\} - \{(X_{2} - \mu)'(X_{2} - \mu) + \operatorname{tr}(\Sigma)\} + 2\operatorname{tr}(\Sigma) = (X_{1} - X_{2})'(X_{1} - X_{2}) - (X_{1} - \mu)'(X_{1} - \mu) - (X_{2} - \mu)'(X_{2} - \mu) = X_{1}'X_{1} - 2X_{2}'X_{1} + X_{2}'X_{2} - X_{2}'X_{2} + 2X_{2}'\mu - \mu'\mu - X_{1}'X_{1} + 2X_{1}'\mu - \mu'\mu = -2(X_{1} - \mu)'(X_{2} - \mu).$$
(3)

Furthermore, [4] say that $\psi_2(X_i, X_j)$ has the following orthogonality proprieties

$$\mathbb{E}[\psi_2(X_i, X_j)\psi_2(X_i, X_k)] = 0 = \mathbb{E}[\psi_2(X_i, X_j)\psi_2(X_k, X_l)],$$

for all $i, j, k, l \in \{1, 2, \dots, n\}$. We also note that $\mathbb{E}[\psi_2(X_i, X_j)^2] < \infty$, for all $i, j \in \{1, 2, \dots, n\}$. Since the first term of Hoeffding's decomposition is null, we can write B_n as a function of the second term $\psi_2(\cdot, \cdot)$, by the following way

$$B_n = \frac{n_1 n_2}{n(n-1)} \left(\frac{2}{n_1 n_2} \sum_{\substack{i \in S_1 \\ j \in S_2}} \psi_2(X_i, X_j) \right)$$

$$-\frac{2}{n_1(n_1-1)}\sum_{\substack{i,j\in S_1\\i< j}}\psi_2(X_i,X_j)-\frac{2}{n_2(n_2-1)}\sum_{\substack{i,j\in S_2\\i< j}}\psi_2(X_i,X_j)\right).$$
(4)

Thus we write

$$\begin{aligned} \operatorname{Var}(B_n) &= \operatorname{Var} \left\{ \frac{n_1 n_2}{n(n-1)} \left(\frac{2}{n_1 n_2} \sum_{\substack{i \in S_1 \\ j \in S_2}} \psi_2(X_i, X_j) \right) \\ &- \frac{2}{n_1(n_1-1)} \sum_{\substack{i,j \in S_1 \\ i < j}} \psi_2(X_i, X_j) - \frac{2}{n_2(n_2-1)} \sum_{\substack{i,j \in S_2 \\ i < j}} \psi_2(X_i, X_j) \right) \right\} \\ &= \frac{n_1^2 n_2^2}{n^2(n-1)^2} \left(\frac{4}{n_1^2 n_2^2} \sum_{\substack{i \in S_1 \\ j \in S_2}} \operatorname{Var}(\psi_2(X_i, X_j)) \right) \\ &+ \frac{4}{n_1^2(n_1-1)^2} \sum_{\substack{i,j \in S_1 \\ i < j}} \operatorname{Var}(\psi_2(X_i, X_j)) + \frac{4}{n_2^2(n_2-1)^2} \sum_{\substack{i,j \in S_2 \\ i < j}} \operatorname{Var}(\psi_2(X_i, X_j)) \right) \\ &= \sigma^4 \frac{n_1^2 n_2^2}{n^2(n-1)^2} \left(\left(\frac{4}{n_1^2 n_2^2} \right) n_1 n_2 + \left(\frac{4}{n_1^2(n_1-1)^2} \right) \frac{n_1(n_1-1)}{2} \\ &+ \left(\frac{4}{n_2^2(n_2-1)^2} \right) \frac{n_2(n_2-1)}{2} \right) \\ &= \sigma^4 \left(\frac{n_1^2 n_2^2}{n^2(n-1)^2} \right) \left[\frac{4}{n_1 n_2} + \frac{2}{n_1(n_1-1)} + \frac{2}{n_2(n_2-1)} \right] \\ &= \sigma^4 \frac{n_1 n_2}{n^2(n-1)^2} \left[\frac{2n^2 - 6n + 4}{(n_1 - 1)(n_2 - 1)} \right] = C(n, n_1) \sigma^4. \end{aligned}$$
(5)

where $C(n, n_1)$ is given by

$$C(n,n_1) = \frac{n_1 n_2}{n^2 (n-1)^2} \left[\frac{2n^2 - 6n + 4}{(n_1 - 1)(n_2 - 1)} \right]$$
(6)

and σ^4 is given by

$$\sigma^{4} = \operatorname{Var}\left[\psi_{2}(X_{1}, X_{2})\right] = \operatorname{Var}\left[-2(X_{1} - \mu)'(X_{2} - \mu)\right]$$

$$= 4\operatorname{Var}\left(\sum_{k=1}^{L} (X_{1k} - \mu_{k})(X_{2k} - \mu_{k})\right) = 4\sum_{k=1}^{L} \operatorname{Var}\left((X_{1k} - \mu_{k})(X_{2k} - \mu_{k})\right)$$

$$+8\sum_{k < s} \operatorname{Cov}\left((X_{1k} - \mu_{k})(X_{2k} - \mu_{k}), (X_{1s} - \mu_{s})(X_{2s} - \mu_{s})\right)$$

$$= 4\sum_{k=1}^{L} \left(\operatorname{Var}(X_{1k} - \mu_{k})\right)^{2} + 8\sum_{k < s} \operatorname{Cov}\left((X_{1k} - \mu_{k}), (X_{1s} - \mu_{s})\right)^{2} = 4\operatorname{vec}(\Sigma)'\operatorname{vec}(\Sigma).$$
(7)

In Section 2.2 we explore this result to test group homogeneity.

Clustering Algorithm - Euclidean Distance

We used the following algorithm to find the group assignments S_1 and S_2 that minimize the objective function given in expression (8).

Step 1: Initialization

- 1. Randomly choose starting centers for groups G_1 and G_2 from the observations X_1, \dots, X_n .
- 2. Assign to each observation a set of indexes, S_1 or S_2 , based on the smallest Euclidean distance to its center.
- 3. Estimate $\operatorname{Var}(B_n)$ with $n_1 = \lfloor \frac{n+1}{2} \rfloor$ through the bootstrap.
- 4. For each group size in $\{2, \dots, n-2\}$ estimate $Var(B_n)$ through expression (10).

Step 2: Iterate

1. For each observation $i \in \{1, \dots, n\}$, assign X_i to group G_1 if

$$f\left(\{S_1^{-i} \cup i\}, S_2^{-i}\right) < f\left(S_1^{-i}, \{S_2^{-i} \cup i\}\right)$$

and assign X_i to group G_2 otherwise, where S_g^{-i} is the set of all indexes, except i, in group G_g , for $g \in \{1, 2\}$.

2. Repeat while convergence criterion is not met.

Step 3: Convergence

1. Stop when S_1 and S_2 are the same in two consecutive iterations.

Clustering Algorithm - General Dissimilarity

Alternative algorithm to find the group assignments S_1 and S_2 that minimize the objective function given in expression (8), when using a general dissimilarity measure. The dissimilarity measure must satisfy the conditions for asymptotic normality of B_n outlined in [4].

Step 1: Initialization

- 1. Randomly choose starting centers for groups G_1 and G_2 from the observations X_1, \dots, X_n .
- 2. Assign to each observation a set of indexes, S_1 or S_2 , based on the smallest dissimilarity to its center.
- 3. For each group size in $\{2, \dots, \lfloor \frac{n+1}{2} \rfloor\}$ estimate $\operatorname{Var}(Bn)$ though the bootstrap.

Step 2: Iterate

1. For each observation $i \in \{1, \dots, n\}$, assign X_i to group G_1 if

$$f\left(\{S_1^{-i} \cup i\}, S_2^{-i}\right) < f\left(S_1^{-i}, \{S_2^{-i} \cup i\}\right)$$

and assign X_i to group G_2 otherwise, where S_g^{-i} is the set of all indexes, except i, in group G_g , for $g \in \{1, 2\}$.

2. Repeat while convergence criterion is not met.

Step 3: Convergence

1. Stop when S_1 and S_2 are the same in two consecutive iterations.

2 Cluster evaluation

We perform a simulation study to compare our clustering algorithm Bn to k-means clustering (km). Two distinct configurations are considered, Normal distribution vs. Normal distribution and Chi-squared distribution vs. Chi-squared distribution.

For the Normal setup, we simulate $n = \{10, 25, 50, 500\}$ independent samples with dimension $L = \{10, 50, 10, 500\}$. We divide the *n* samples in two groups of size $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$, where $\lfloor x \rfloor$ is the integer part of *x*. The first group is sampled from standard normal distribution and the second group is sampled from a normal distribution with unit variance and mean taking values in $\mu_2 = \{0.5, 1, 2\}$. The results of 100 replications are reported in the Table 1.

For the Chi-squared setup we adopt the same scheme used in the Table 1. The first group is sampled from a chi-squared distribution with 5 degrees of freedom and non-centrality parameter ncp=0. The othes $n - \lfloor n/2 \rfloor$ samples are sampled from chi-squared distribution with 5 degrees of freedom and non-centrality parameter taking values in $ncp = \{1, 2, 4\}$. The results of 100 replications are reported in the Table .

For clustering evaluation we consider the cluster similarity index presented in [1]. Let $C = \{C_1, \dots, C_Q\}$ and $G = \{G_1, \dots, G_k\}, 1 \leq k \leq Q$ be the set of Q true clusters and a k-cluster solution, respectively. The cluster similarity index is defined as

$$Sim(G,C) = \frac{1}{Q} \sum_{i=1}^{Q} \max_{i \le j \le k} Sim(G_j, C_i),$$
(8)

where

$$Sim(G_j, C_i) = 2 \frac{|G_j \cap C_i|}{|G_j| + |C_i|},$$
(9)

and $|\cdot|$ denotes the cardinality of the elements in each set. Note that this similarity measure will return 0 if the two clusterings are completely dissimilar and 1 if they are the same.

			$\mu_2 =$	=0.5			μ_2	=1			μ_2	=2	
$\frac{1}{\Gamma}$	u	10	25	50	100	10	25	50	100	10	25	50	100
	Bn	0.670	0.667	0.667	0.664	0.667	0.668	0.691	0.720	0.996	0.962	0.982	0.99_{2}
TU	km	0.689	0.641	0.623	0.597	0.840	0.864	0.889	0.890	0.998	0.998	0.999	0.999
	Bn	0.673	0.679	0.679	0.689	1.000	0.993	0.976	0.994	1.000	1.000	1.000	1.00(
00	km	0.793	0.830	0.783	0.705	0.997	0.999	0.999	0.999	1.000	1.000	1.000	1.00(
1001	Bn	0.677	0.718	0.670	0.753	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00(
100	km	0.872	0.945	0.959	0.963	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00(
	Bn	0.673	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.00(
nne	km	0.999	1,000	1,000	1,000	1 000	1 000	1,000	1,000	1 000	1 000	1 000	1 000

are preseted as the average cluster similarity	
Comparisons	
distributed data.	
eans for normal	t.
Is Bn and km	nethod's resul
on of method	er and each r
Cluster evaluati	ween true clust
Table 1: (index bet

6

Table 2: Cluster evaluation of methods Bn and $kmeans$ for Chi-squared distributed data. Comparisons are preseted as the average cluster similarity index between true cluster and each method's result.

			ncp)=1			ncp	=2			ncp	=4	
T	u	10	25	50	100	10	25	50	100	10	25	50	100
01	Bn	0.666	0.665	0.667	0.665	0.665	0.666	0.667	0.668	0.672	0.666	0.668	0.669
пт	km	0.649	0.632	0.631	0.622	0.676	0.669	0.656	0.642	0.756	0.744	0.739	0.715
Си И	Bn	0.665	0.664	0.666	0.668	0.668	0.668	0.673	0.675	0.710	0.882	0.943	0.966
ne	km	0.671	0.684	0.688	0.684	0.739	0.809	0.810	0.814	0.928	0.958	0.961	0.966
1001	Bn	0.667	0.666	0.669	0.673	0.665	0.676	0.726	0.792	1.000	1.000	1.000	1.000
00T	km	0.691	0.736	0.759	0.766	0.858	0.905	0.914	0.911	0.991	0.994	0.997	0.995
	Bn	0.670	0.670	0.741	0.824	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
000	km	0.789	0.942	0.965	0.959	0.982	0.999	0.999	0.999	1.000	1.000	1.000	1.000

3 Size and power for Sigclust and U-statistics approach

In this section we present estimates of power and size for testing group homogeneity with the Sigclust (Sg) method proposed by [3] and our method (Bn). Two scenarios are considered. In the first one, we compare empirical power considering the normal distribution for $n = \{10, 15, 20, 25, 30\}$ independent samples with dimension $L = \{500, 1000\}$. The *n* samples are divided into two groups of sample sizes $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$. The first group is sampled from the standard normal distribution and the second group is sampled from a normal distribution with unit variance and mean taking values in $\mu_2 = \{0.0, 0.2, 0.3, 0.4, 0.5\}$. The proportion of p-values lower than 0.05 in a 100 replications are reported in the Table 3. The column $\mu_2 = 0.0$ provides the empirical size estimates.

Although the Sigclust method is not suitable for non normal data, in the second scenario, we simulate data from Chi-squared distribution. The *n* samples are divided in two groups of sample sizes $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$ where *n* takes values $n = \{10, 15, 20, 25, 30\}$ and the dimensions of each sample are $L = \{500, 1000\}$. The first group is sampled from a central Chi-squared distribution with 5 degrees of freedom. The other $n - \lfloor n/2 \rfloor$ samples are sampled from a Chi-squared distribution with 5 degrees of freedom and non-centrality parameter taking values in ncp= $\{0.0, 0.25, 0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0\}$.

Table 3:	Proportion	of p-values	lower	than	$0.05 \ \mathrm{fc}$	r group	homogeneity	r tests	using	Sigclust	and
Bn for 1	normal data.										

				L = 500]	L=1000)	
n μ_2 n		0.0	0.2	0.3	0.4	0.5	0.0	0.2	0.3	0.4	0.5
	Bn	0.01	0.07	0.22	0.84	1.00	0.03	0.05	0.65	0.98	1.00
10	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Bn	0.06	0.11	0.59	1.00	1.00	0.03	0.13	0.95	1.00	1.00
15	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Bn	0.07	0.20	0.87	1.00	1.00	0.02	0.37	0.99	1.00	1.00
20	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	Bn	0.06	0.16	0.98	1.00	1.00	0.04	0.47	1.00	1.00	1.00
25	Sg	0.00	0.00	0.00	0.00	0.47	0.00	0.00	0.00	0.00	0.82
	Sgn	0.00	0.00	0.00	0.00	0.62	0.00	0.00	0.00	0.00	0.92
	Bn	0.09	0.40	1.00	1.00	1.00	0.02	0.65	1.00	1.00	1.00
30	Sg	0.00	0.00	0.00	0.03	1.00	0.00	0.00	0.00	0.01	1.00
	Sgn	0.00	0.00	0.00	0.03	1.00	0.00	0.00	0.00	0.01	1.00

References

- Sonia Pértega Díaz and José A. Vilar. Comparing several parametric and nonparametric approaches to time series clustering: A simulation study. *Journal of Classification*, 27(3):333–362, Nov 2010.
- [2] Justin Lee. U-statistics: Theory and Practice. New York: Marcel Dekker, 1990.
- [3] Yufeng Liu, David Neil Hayes, Andrew Nobel, and JS Marron. Statistical significance of

							ncp				
L	n μ_2 n		0.0	0.25	0.5	1.0	1.5	2.0	3.0	4.0	5.0
	10	Bn	0.07	0.02	0.07	0.17	0.93	1.00	1.00	1.00	1.00
	10	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.90
		Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	1.00
	15	Bn	0.02	0.06	0.06	0.60	0.99	1.00	1.00	1.00	1.00
	10	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.05	1.00	1.00
500		Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.22	1.00	1.00
500	20	Bn	0.03	0.04	0.07	0.81	1.00	1.00	1.00	1.00	1.00
	20	Sg	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
	25	Bn	0.06	0.10	0.12	0.99	1.00	1.00	1.00	1.00	1.00
	20	Sg	0.00	0.00	0.00	0.00	0.00	0.02	1.00	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	0.13	1.00	1.00	1.00
	30	Bn	0.09	0.11	0.20	1.00	1.00	1.00	1.00	1.00	1.00
	50	Sg	0.00	0.00	0.00	0.00	0.00	0.91	1.00	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	0.98	1.00	1.00	1.00
	10	Bn	0.03	0.07	0.01	0.46	1.00	1.00	1.00	1.00	1.00
	10	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97
		Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	1.00
1000	15	Bn	0.04	0.04	0.03	0.94	1.00	1.00	1.00	1.00	1.00
	10	Sg	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	0.00	0.27	1.00	1.00
1000	20	Bn	0.06	0.05	0.06	1.00	1.00	1.00	1.00	1.00	1.00
	20	Sg	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
	25	Bn	0.03	0.07	0.08	1.00	1.00	1.00	1.00	1.00	1.00
	20	Sg	0.00	0.00	0.00	0.00	0.00	0.04	1.00	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	0.23	1.00	1.00	1.00
	30	Bn	0.08	0.03	0.14	1.00	1.00	1.00	1.00	1.00	1.00
		Sg	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
		Sgn	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00

Table 4: Proportion of p-values lower than 0.05 for testing group homogeneity using Sigclust (Sg and normalized Sgn) and Bn with Chi-squared data.

clustering for high-dimension, low-sample size data. Journal of the American Statistical Association, 103(483):1281–1293, 2008.

[4] Aluísio Pinheiro, Pranab K. Sen, and Hildete P. Pinheiro. Decomposability of highdimensional diversity measures: Quasi-statistics, martingales and nonstandard asymptotics. Journal of Multivariate Analysis, 100(8):1645 – 1656, 2009.