Débora Zava Bello

Inferência em agrupamento considerando múltiplos grupos

Porto Alegre - RS, Brasil

30 de março de 2021

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

Débora Zava Bello

Inferência em agrupamento considerando múltiplos grupos

Dissertação de mestrado

Dissertação apresentada para obtenção do grau de Mestre em Estatística pela Universidade Federal do Rio Grande do Sul.

Orientador: Prof. Dr. Marcio Valk

Co-Orientador(a): Profa. Dra. Gabriela Bettella Cybis

Porto Alegre - RS, Brasil

30 de março de 2021

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

Débora Zava Bello

Inferência em agrupamento considerando múltiplos grupos

Dissertação sob o título *"Inferência em agrupamento considerando múltiplos grupos"*, defendida por Débora Zava Belloe aprovada em 30 de março de 2021, na cidade de Porto Alegre, no Estado do Rio Grande do Sul, pela banca examinadora constituída pelos professores:

Prof. Dr. Marcio Valk
Instituto de Matemática – UFRGS

Prof. Dr. Danilo Marcondes Filho
Instituto de Matemática – UFRGS

Profa. Dra. Sílvia Regina Costa Lopes
Instituto de Matemática – UFRGS

Prof. Dr. Rodrigo Lambert
FAMAT – UFU

Porto Alegre, 30 de março de 2021

CIP - Catalogação na Publicação

Zava Bello, Débora
Inferência em agrupamento considerando múltiplos grupos / Débora Zava Bello. -- 2021.
65 f.
Orientador: Marcio Valk.

Coorientadora: Gabriela Bettela Cybis.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística, Programa de Pós-Graduação em Estatística, Porto Alegre, BR-RS, 2021.

1. clustering. 2. múltiplos grupos. 3. agrupamento. I. Valk, Marcio, orient. II. Bettela Cybis, Gabriela, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Resumo

Métodos de agrupamento são ferramentas úteis na identificação de padrões em conjuntos de dados. No contexto de alta dimensionalidade e tamanho amostral pequeno, o desafio de decidir se o agrupamento encontrado é estatisticamente significativo é ainda maior. Entre os métodos de agrupamento adequados à esse contexto, poucos possuem inferência e muitas vezes são específicos para dois grupos. Estamos propondo um método para agrupar de forma ótima em mais conjuntos, nesse caso três. Além de uma abordagem para clusterização dos elementos em três grupos, propomos um teste de homogeneidade para verificar a sua significância. Apresentamos a estatística de teste, suas propriedades assintóticas e, através de simulações, estudamos propriedades como tamanho e poder do teste proposto. Comparações com outras metodologias binárias indicam que nossa proposta é mais adequada para situações em que os dados têm uma estrutura inerente de três grupos.

Palavras-chave: Cluster. U-estatística. Inferência. Múltiplos grupos.

Dedicatória

Dedico esta e todas as conquistas à minha família.

Agradecimentos

Agradeço primeiro à minha família: mãe, avô, tias, tio, irmãos e primos. Por todo apoio seja financeiro ou emocional. Vocês são minha inspiração, não só na vida acadêmica.

Ao meu orientador Dr Marcio Valk, pelos ensinamentos, paciência, apoio, incentivo e momentos de vídeo-chamada com seus familiares.

A todos os professores que me acompanharam durante a minha jornada até aqui, em especial à Dra Gabriella Cybis pela ensinamentos durante a coorientação. Agradeço também aos professores que aceitaram fazer parte da minha banca por suas considerações quanto ao trabalho.

Aos meus colegas de mestrado por compartilharem os momentos de desespero presencialmente ou remotamente.

Aos Mestres Daiane Oliveira e Ruan Vianna por todo o apoio não somente durante a pós graduação, mas desde o planejamento de ingressar no programa.

Aos amigos que me apoiaram durante a pandemia. Em especial: Paulo Zava pelas conversas aleatórias enquanto eu tentava focar na dissertação; Vanessa Medeiros pela disposição em ajudar; Byanka Cardoso pelas palavras de incentivo e apoio; e Carolina Medeiros pelas conversas amigáveis e discussões.

A Joe e 23 pessoas que estão 12 horas de distância, por serem as distrações mais difíceis durante o projeto.

A todos vocês, muito obrigada.

Sumário

1 Introdução

Métodos de agrupamento são ferramentas estatísticas que consistem em separar dados baseadas em características que estes possuem. O presente trabalho tem o objetivo de apresentar nova abordagem para clusterização de dados, baseada em U-estatística, para três grupos enquanto verifica se esta separação é estatisticamente significativa. Tal método é recomendado para casos de alta dimensão dos dados e tamanho pequeno de amostra (HDLSS). A metodologia apresentada pode ser adaptada para diferentes medidas de distância, além de ser não paramétrica. Como o método é uma extensão do agrupamento baseado em U-estatística, *uclust* (VALK; CYBIS, 2020), a nova abordagem busca ser mais poderosa ao separar os dados em três grupos, quando essa estrutura específica de grupos é inerente ao conjunto de dados.

U-estatística vem sendo utilizada para construção de métodos de agrupamento e se mostrou uma ótima ferramenta para obtenção de propriedades estatísticas desses métodos. (PINHEIRO; SEN; PINHEIRO, 2009) apresentou como adequar a teoria do agrupamento baseado em U-estatística em MANOVA ou modelos de alta dimensão. Para dados de séries temporais a metodologia baseada em U-estatística, mostrou-se adequada para encontrar agrupamentos estatisticamente significativos, como mostrado em (VALK; PINHEIRO, 2012).

Essa abordagem começou ainda a ser aplicada para estudar inferência em agrupamentos. O trabalho de (CYBIS; VALK; LOPES, 2018) apresentou teste de homogeneidade em um conjunto de dados. Também foi proposto como obter significância em classificação de um elemento em um determinado grupo.

Por fim, essa teoria foi generalizada por (VALK; CYBIS, 2020), abrangendo o caso em que o conjunto de dados possui um outlier e apresentando método hierárquico para agrupamento dos dados com a verificação da significância estatística desse agrupamento.

O presente trabalho é organizado da forma que primeiro uma noção básica de U-estatística e suas propriedades são expostas. Na sequência, o artigo completo e seu ma-

terial suplementar encontram-se anexados. O referido artigo consta de ampla introdução ao assunto, descreve a metodologia utilizada e apresenta a proposição do método para o qual as propriedades assintóticas são demonstradas. Além disso, estudos de simulação e aplicação em dados reais foram abordados ainda no artigo.

2 U-estatísticas

Para melhor entender a discussão teórica desse trabalho, um resumo da teoria de U-estatística é abordada nessa seção. A motivação por trás da classe, além de resultados teóricos importantes para o desenvolvimento do método proposto são explicados aqui.

A classe de U-estatísticas foi apresentada por (HALMOS, 1946) e (HOEFFDING, 1948), em 1946 e 1948 respectivamente. Esta é motivada como (LEE, 1990) apresenta.

Seja $\theta = \theta(F)$ um funcional definido no conjunto de funções de distribuição $F \in \mathcal{F}$, $\mathcal{F}$ em $\mathbb{R}$. Tem-se então o interesse em estimar $\theta(F)$ na base de uma amostra de variáveis aleatórias (v.a.'s), $X_1, \cdots, X_n$ independentes e identicamente distribuídas (i.i.d.) com função de distribuição acumulada F.

O interesse é saber se existe estimador para θ independente de qual seja F. E, se existe, se este é não viesado. Caso exista mais de um estimador não viesado para θ, busca-se verificar ainda qual o melhor. Com base nessas curiosidades, o trabalho de (LEE, 1990) apresenta teoremas e lemas discutindo esses desejos e por fim define U-estatística com suas descobertas.

Seja $\mathcal{F}$ um subconjunto do conjunto de funções de distribuição em $\mathbb{R}$ e $\theta(F)$ um funcional definido em $\mathcal{F}$. Suponha então que para cada n inteiro suficientemente grande, existe uma função $f_n(X_1, \cdots, X_n)$ de n variáveis, tais que

$$E\left[f_n(X_1, \cdots, X_n)\right] = \theta(F) \tag{2.1}$$

para todo F em $\mathcal{F}$, onde $X_1, \cdots, X_n$ é sequência de v.a.'s com distribuição F. Então, $\theta(\cdot)$ admite estimador não viesado.

De maneira mais formal, o livro de (LEE, 1990) apresenta o seguinte teorema.

Teorema 2.1 *Um funcional θ definido em um conjunto $\mathcal{F}$ de funções de distribuição*

admite estimador não viesado se e somente se existe uma função ψ de k variáveis tal que

$$\theta(F) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(x_1, x_2, \cdots, x_k) dF(x_1) \cdots dF(x_k) \tag{2.2}$$

para todo F em $\mathcal{F}$.

Se θ é definido como (2.2), então um estimador não viesado é da forma

$$f_n(X_1, \cdots, X_n) = \psi(X_1, \cdots, X_k). \tag{2.3}$$

Um funcional que satisfaz (2.2) para alguma função ψ é chamado de **estatística funcional regular de grau** k. A função ψ é chamada de **kernel** do funcional.

Para a investigação quanto a unicidade do estimador, primeiro define-se aqui estimadores idênticos. Considera-se estimadores como idênticos se estes concordam com o mesmo conjunto de Borel E. A escolha de E depende do conjunto $\mathcal{F}$ abordado. Lee cita como exemplos de escolhas para E: se $\mathcal{F}$ é todas as distribuições no conjunto $\{0, 1\}$, então uma escolha válida para E é $\{0, 1\}$; ou ainda, se $\mathcal{F}$ é todas as distribuições onde a média existe, então uma escolha para E é $\mathbb{R}$.

Além disso, se $\mathcal{F}$ é suficientemente grande (até a igualdade em E), então há somente um estimador simétrico não viesado. Com "$\mathcal{F}$ suficientemente grande"considera-se que $\mathcal{F}$ é grande o suficiente para incluir todas as distribuições com suporte finito em E.

Teorema 2.2 *Seja $\mathcal{F}$ que contem todos as distribuições com suporte finito em* E *e seja θ funcional regular que satisfaz* (2.2)*. Então, para a igualdade em* E*, há somente um único estimador simétrico não viesado para θ.*

Quanto aos interesses iniciais, é preciso verificar ainda como caracterizar o estimador. Seja

$$\psi^{[n]}(x_1, \cdots, x_n) = \frac{(n-k)!}{n!} \sum \psi(x_{1_i}, \cdots, x_{i_k}), \tag{2.4}$$

onde a soma compreende todas as $\frac{(n-k)!}{n!}$ permutações $(i_1, \cdots, i_k)$ de inteiros distintos escolhidos entre $\{1, 2, \cdots, n\}$. É possível mostrar que $\psi^{[n]}$ também é não viesado. Então,

quando há mais de um estimador não viesado, o livro de (LEE, 1990) apresenta resultados para encontrar qual o melhor estimador.

Lema 2.1 *Seja $\mathcal{F}$ conjunto que contem todas as distribuições com suporte finito em* E *e seja f uma função simétrica de n variáveis com* $\int \cdots \int_{\mathbb{R}^n} f(x_1, \cdots, x_n) \prod_{i=1}^{n} dF(x_i) = 0$, $\forall f \in \mathcal{F}$. *Então,* $f(x_1, \cdots, x_n) = 0$ *quando* $x_i \in \mathrm{E}$, $i = 1, \cdots, n$.

No caso onde $\mathrm{E} = \mathbb{R}$ o estimador simétrico essencialmente único $\psi^{[n]}$ é também o de variância mínima.

Teorema 2.3 *Seja θ um funcional regular de grau k definido da forma* (2.2) *em um conjunto $\mathcal{F}$ de funções de distribuição contendo todas as distribuições com suporte finito. Seja f estimador não viesado de θ baseado em uma amostra de tamanho n que satisfaça* (2.1). *Então,* $\mathrm{Var}(f) \geqslant \mathrm{Var}(\psi^{[n]})$, $\forall f \in \mathcal{F}$.

Teorema 2.4 *(I) Seja θ estatística regular funcional de grau k com kernel ψ definida em um conjunto $\mathcal{F}$ de funções de distribuição contendo todas as funções de distribuição absolutamente contínuas. Então, $\psi^{[n]}$ é o único estimador simétrico não viesado para θ.*

(II) O estimador $\psi^{[n]}$ tem variância mínima na classe de todos os estimadores não viesados de θ.

Lema 2.2 *Seja θ estatística funcional regular com kernel simétrico ψ de ordem k definida no conjunto $\mathcal{F}$ de funções de distribuição absolutamente contínuas e suponha $\theta(F) = 0$ $\forall F \in \mathcal{F}$. Então, $\psi = 0$ quase certamente em $\mathbb{R}^k$.*

Os resultados apresentados nos lemas anteriores são pertinentes ao conceito de estatística completa. Lembrando que seja $X_1, \cdots, X_n$ uma amostra aleatória de $X \sim \mathcal{D}(\theta)$ e $T(\mathbf{X})$ uma estatística da amostra. Esta é dita **estatística completa** se $E\left[g\left(T(\mathbf{X})\right)\right] = 0$ $\forall\ \theta \Rightarrow \mathrm{P}\left(g(T(\mathbf{X}))\right) = 1\ \forall\ \theta$.

Com todos os interesses iniciais satisfeitos, o próximo passo é a definição de U-estatística como um estimador simétrico não viesado e único.

Definição 2.1 *Para famílias $\mathcal{F}$ contendo todas as distribuições com suporte finito ou todas absolutamente contínuas escolhe-se como estimativa para θ o estimador essencialmente único $\hat{\theta} = \psi^{[n]}$ definido como* (2.4). *Pode-se definir então*

$$\psi^{[k]}(x_1, \cdots, x_k) = \left(\frac{1}{k!}\right) \sum \psi(x_{i_1}, \cdots, x_{i_k}), \tag{2.5}$$

onde a soma é composta por todas as permutações $(i_1, \cdots, i_k)$ *de* $\{1, \cdots, k\}$. *Então, pode-se escrever*

$$\hat{\theta} = \binom{n}{k}^{-1} \sum_{(n,k)} \psi^{[k]}(X_{i_1}, \cdots, X_{i_k}), \tag{2.6}$$

onde a $\sum_{(n,k)}$ *é composta por todos os subgrupos* $1 \leqslant i_1 \leqslant \cdots \leqslant i_k \leqslant n$ *de* $\{1, \cdots, n\}$. *Porém, é possível notar que*

$$\begin{aligned} \int \cdots \int_{\mathbb{R}^k} \psi(x_1, \cdots, x_k) \prod_{i=1}^{k} dF(x_i) &= \\ &= \int \cdots \int_{\mathbb{R}^k} \psi^{[k]}(x_1, \cdots, x_k) \prod_{i=1}^{k} dF(x_i). \end{aligned}$$

Dessa forma, sem perda de generalidade pode-se reescrever o estimador simétrico, único e não viesado da forma

$$\hat{\theta} = \binom{n}{k}^{-1} \sum_{(n,k)} \psi(X_{i_1}, \cdots, X_{i_k}). \tag{2.7}$$

Este estimador é chamado de ***U-estatística****.*

Uma outra notação para U-estatística é U_n, para a continuação do trabalho esta será abordada.

$$U_n = \binom{n}{k}^{-1} \sum_{(n,k)} \psi(X_{i_1}, \cdots, X_{i_k}). \tag{2.8}$$

A literatura apresenta ampla discussão quanto as escolhas dos conjuntos E e $\mathcal{F}$. Para os métodos desenvolvidos nesse trabalho esse questionamento não se fez necessário. Porém, caso o leitor esteja interessado em outros universos para os conjuntos, este é

convidado a ler os trabalhos de (BELL et al., 1960) e (YAMATO; MAESONO, 1989).

2.1 Variância de U-estatística

Ainda com o foco nas discussões futuras do presente trabalho, uma análise da teoria de U-estatística é essencial. Dessa vez, será abordada a variância de U-estatística para amostras finitas. Propriedades destas são amplamente discutidas em (LEE, 1990). Nessa seção um resumo das propriedades mais importantes é apresentado.

A variância de uma U-estatística baseada em variáveis aleatórias independentes e identicamente distribuídas pode ser expressa em termos de esperanças condicionais. Defina então para $c = 1, 2, \cdots, k$ as esperanças condicionais:

$$\psi_c(x_1, \cdots, x_c) = \mathrm{E}[\psi(x_1, \cdots, x_c, X_{c+1}, \cdots, X_k)]$$

e a variância

$$\sigma_c^2 = \mathrm{Var}[\psi_c(X_1, \cdots, X_c)].$$

Além disso, defina

$$\sigma_0^2 = 0.$$

A partir dessas definições, se seguem algumas propriedades das esperanças condicionais.

Teorema 2.5 *As funções ψ_c definidas acima têm as seguintes propriedades:*

(I) $\psi_c(x_1, \cdots, x_c) = \mathrm{E}[\psi_d(x_1, \cdots, x_c, X_{c+1}, \cdots, X_d)]$, *para* $1 \leqslant c < d \leqslant k$

(II) $\mathrm{E}[(\psi_c(X_1, \cdots, X_c)] = \mathrm{E}_\psi(X_1, \cdots, X_k)$

A variância σ_c^2 das esperanças condicionais tem uma interpretação de covariância, para isso (LEE, 1990) apresenta o seguinte teorema.

Teorema 2.6 *Sejam S_1 e S_2 dois k-subconjuntos de $\{1, \cdots, n\}$ com c elementos em comum. Então, uma forma alternativa de σ_c^2 é*

$$\sigma_c^2 = \mathrm{Cov}\left(\psi(S_1), \psi(S_2)\right).$$

Com base dessas esperanças condicionais é possível então estabelecer uma expressão para a variância de uma U-estatística de amostra finita, esta se encontra no teorema apresentado por (LEE, 1990) a seguir.

Teorema 2.7 *Seja U_n U-estatística com kernel ψ de grau k. Então,*

$$\mathrm{Var}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\sigma_c^2.$$

Os dois próximos resultados têm papel importante na demonstração do Teorema Central do Limite (TCL) de U-estatísticas.

Teorema 2.8 *Para $0 \leqslant c \leqslant d \leqslant k$*

$$\frac{\sigma_c^2}{c} \leqslant \frac{\sigma_d^2}{d}.$$

Teorema 2.9 *A função $\mathrm{Var}(nU_n)$ é decrescente em relação a n.*

As demonstrações dos teoremas e lemas se encontram em (LEE, 1990). O livro citado também aborda covariância de duas U-estatísticas além de momentos maiores de U-estatísticas. Novamente, como o objetivo dessa seção é fornecer uma base para entender os cálculos utilizados nos métodos desenvolvidos pelo trabalho, estes foram omitidos do presente texto.

2.2 Decomposição de Hoeffding

Essa seção tem o objetivo de apresentar a decomposição de Hoeffding, resultado muito utilizado na teoria de U-estatística e nos métodos propostos pelo projeto. Essa decomposição permite reescrever a U-estatística de grau k em termos de somas de U-estatísticas não correlacionadas de grau $1, 2, \cdots, k$.

O primeiro passo para falar da decomposição é apresentar os kernels $h^{(1)}, h^{(2)}, \cdots, h^{(k)}$ de graus $1, 2, \cdots, k$, definidos recursivamente pelas equações:

$$h^{(1)}(x_1) = \psi_1(x_1) - \theta \tag{2.9}$$

e

$$h^{(j)}(x_1, \cdots, x_j) = \psi_j(x_1, \cdots, x_j) - \sum_{k=1}^{j-1} \sum_{(j,k)} h^{(k)}(x_{i_1}, \cdots, x_{i_k}) - \theta \tag{2.10}$$

para $j = 2, 3, \cdots, k$. Em posse destes, é possível reescrever os subconjuntos de $\{1, \cdots, n\}$.

Seja $S_j(i_j, \cdots, i_k)$ a soma $\sum h^{(j)}(x_{\nu_1}, \cdots, x_{\nu_j})$ de todos os j subconjuntos $\{\nu_1, \cdots, \nu_j\}$ de $\{i_1, \cdots, i_k\}$.

Pode-se então utilizar a relação

$$\sum_{(n,k)} S_j(i_1, \cdots, i_k) = \binom{n-j}{k-j} \sum_{(n,j)} h^{(j)}(x_{\nu_1}, \cdots, x_{\nu_j}), \tag{2.11}$$

$$\binom{n}{k}^{-1} \binom{n-j}{k-j} = \binom{k}{j} \binom{n}{j}^{-1} \tag{2.12}$$

e a relação (2.10) quando $j = k$ para reescrever U_n da forma

$$\begin{aligned} U_n &= \binom{n}{k}^{-1} \sum_{(n,k)} \psi(x_{i_1}, \cdots, x_{i_k}) \\ &= \binom{n}{k}^{-1} \sum_{(n,k)} \sum_{j=1}^{k} S_j(i_1, \cdots, i_k) + \theta \\ &= \theta + \binom{n}{k}^{-1} \sum_{j=1}^{k} \binom{n-j}{k-j} \sum_{(n,j)} h^{(j)}(x_{\nu_1}, \cdots, x_{\nu_j}) \\ &= \theta + \sum_{j=1}^{k} \binom{k}{j} H_n^{(j)}, \end{aligned}$$

onde $H_n^{(j)}$ é a U-estatística de grau j com kernel $h^{(j)}$.

De maneira mais formal, o livro de (LEE, 1990) descreve o seguinte teorema definindo a decomposição de Hoeffding de uma U-estatística.

Teorema 2.10 *Para $j = 1, \cdots, k$, seja $H_n^{(j)}$ a U-estatística de kernel $h^{(j)}$ definido por* (2.10)*. Então*

$$U_n = \theta + \sum_{j=1}^{k} \binom{k}{j} H_n^{(j)}. \tag{2.13}$$

A decomposição (2.13) é chamada de **decomposição de Hoeffding**. As funções $H_n^{(j)}$ são não-correlacionadas, com variância de ordem decrescente em n. Para demonstrações dessas afirmações, consultar (LEE, 1990).

Essa decomposição é utilizada para verificar as propriedades da estatística proposta pelo presente trabalho.

2.3 Teoria assintótica

Para a elaboração dos métodos propostos nesse projeto a teoria assintótica utilizada foi a normalidade assintótica de U-estatísticas. Novamente, para mais detalhes e demonstrações é indicado a leitura do terceiro capítulo do trabalho do (LEE, 1990).

Teorema 2.11 *Seja $\sigma_1^2 > 0$. Então, $n^{1/2}(U_n - \theta)$ é assintoticamente normal com média zero e variância assintótica $k^2\sigma_1^2$.*

Esse resultado é consequência do Teorema Central do Limite para variáveis aleatórias independentes.

Até o momento foram apresentados resultados, teoremas e lemas envolvendo U-estatísticas de grau k, porém para o desenvolvimento do método proposto pelo projeto serão utilizadas U-estatísticas de grau $k = 2$. Com a leitura disposta até aqui é esperado que o leitor acompanhe sem dificuldade as discussões abordadas pelo artigo no capítulo a seguir.

Referências

BELL, C. et al. On the completeness of order statistics. *Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 31, n. 3, p. 794–797, 1960.

CYBIS, G. B.; VALK, M.; LOPES, S. R. C. Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*, 2018.

HALMOS, P. R. The theory of unbiased estimation. *Annals of Math Stat*, 1946.

HOEFFDING, W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistic*, 1948.

LEE, A. J. *U-statistics theory and practice*. [S.l.]: Marcel Dekker, 1990.

PINHEIRO, A.; SEN, P. K.; PINHEIRO, H. P. Decomposability of high-dimensional diversity measures: Quasi-u-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, 2009.

VALK, M.; CYBIS, G. B. U-statistical inference for hierarchical clustering. *Journal of Computational and Graphical Statistics*, 2020.

VALK, M.; PINHEIRO, A. Time-series clustering via quasi u-statistics. *Journal of Time Series Analysis*, 2012.

YAMATO, H.; MAESONO, Y. Deficiencies of u-statistics of degree 2 under symmetric distributions. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 18, n. 1, p. 53–66, 1989.

3 Artigo Bello, Valk e Cybis (2021)

Autores: Débora Zava Bello, Marcio Valk e Gabriela Cybis

Título: Clustering inference in multiple groups

Ano: 2021.

Clustering inference in multiple groups

Debora Zava Bello, Marcio Valk and Gabriela Cybis

March 2021

Abstract

Inference in clustering is paramount to uncovering inherent group structure in the data. Clustering methods which assess statistical significance have recently drawn attention owing to their importance for the identification of patterns in high dimensional data with applications in many scientific fields. We present here a U-statistics based approach, specially tailored for high-dimensional data, that clusters the data into *three* groups while assessing the significance of such partitions. Because our approach stands on the U-statistics based clustering framework of *uclust*, it inherits its characteristics being a non-parametric method relying on very few assumptions about the data, and thus can be applied to a wide range of dataset. Furthermore our method aims to be a more powerful tool to find the best partitions of the data into three groups when that particular structure is present. In order to do so, we first propose an extension of the test U-statistic and develop its asymptotic theory. Additionally we propose a ternary non-nested significance clustering method. Our approach is tested through multiple simulations and found to have more statistical power than competing alternatives in all scenarios considered. An application to image recognition shows that our proposal presents a superior performance for this special case.

1 Introduction

In clusters analysis the aim is to divide data into groups of similar items and there are different ways to accomplish this task. A large number of algorithms based on different measures have been proposed and each different measure may lead to potentially different results ([Euan et al., 2019]). Clusters can be inherently present in the data like phylogenetic analysis ([Rosenberg et al., 2002, Chen et al., 2015]) or they can be built when clustering should take place regardless of whether innate cluster structure is present as in customer segmentation ([Motlagh et al., 2019, Hennig, 2015]). In order to evaluate clustering methods, it is necessary to consider the context, the objectives of clustering and to have a suitable measure of dissimilarity ([Von Luxburg et al., 2012]). A critical issue is how to discover inherent cluster structure in data, in other words, whether the clusters represent in fact an important feature or are simply the result of sample variation. This becomes even more challenging when considering the context

of high dimensional data. We present here a U-statistics based approach that clusters the data in three groups while assessing the significance of such partitions. Our method is specially tailored for high-dimensional data and adaptable to different distance measures.

In a typical application of inference in clustering when the groups are already defined and there is no need for an algorithm or method to find them, the null hypothesis is that all groups are random samples from the same population (overall sample homogeneity). In the multivariate analysis of variance (M)ANOVA procedure, when presented in terms of a linear model, the homogeneity of groups stands for equality of means between all groups. Assumptions of independence and normality of the data, homoscedasticity of variance and homogeneity in group are required for exact (finite sample) inference. In addition, a large sample size, depending on the dimension of the data is generally necessary. For the context where there is no information about the existence of groups and the objective is to know if they exist and what they are, some approaches have been proposed for addressing the problem of assessing significance of partitions, or determining which clustering layers represent actual population structure and which are simple consequence of spurious random effects. To avoid resorting to heuristic criteria or the researcher's judgement to define which partition levels should be assigned meaning these approaches proposes to assess statistical significance. However the success of these methods depends on the underlying cluster structure ([Adolfsson et al., 2019]).

Several approaches have been proposed to assess statistical significance in clustering, for example the procedure presented in [McLachlan and Peel, 2004] which considers mixture models of distributions such as the Gaussian. A maximum likelihood approach is used by [Demidenko, 2018] to test no-clusters hypothesis. However, when the data are high dimensional and have small sample sizes the problem becomes increasingly challenging, since it involves complete parametric estimation, usually requiring costly matrix inversions. The works of [McShane et al., 2002, Helgeson et al., 2020] address this issue by using reduction of dimensionality of the data matrix and sparse covariance estimation. An approach inspired on the bootstrap strategy is proposed by [Shimodaira et al., 2004] which is implemented in the R package *pvclust* ([Suzuki and Shimodaira, 2006]) and used in phylogenetics to assess confidence in hierarchical clustering. [Liu et al., 2008] proposes a statistical test to assess the significance of clustering the data into K groups, specifically tailored to the high dimension low sample size (HDLSS) scenario, that has been implemented in the R package *sigclust*. However, the implementation and applications consider only two groups. Additionally, [Kimes et al., 2017] extend the method to assess significance in hierarchical clustering. However, this approach requires that the data comes from a single multivariate normal distribution, which can be an issue since rejection of the no cluster hypothesis may be a simple consequence of non-normal data.

Our work focuses specifically on the HDLSS setting and extends the works of [Cybis et al., 2018, Valk and Cybis, 2020] making it possible to simultaneously test the homogeneity of *three* groups, one of which may have size *one*. The test

statistic to compare *three* groups, where one of them may be an outlier, is a extension of the test statistic B_n proposed by [Pinheiro et al., 2009]. Here the hypotheses are similar to those of (M)ANOVA where the null is that the elements in the *three* groups come from the same distribution (homogeneity, no-clusters) versus the alternative hypothesis that the data distribution (not necessarily normal) of at least one of the groups is different from the others. Asymptotic normality of the extended B_n is obtained using U-statistics theory. An estimator for the variance of the extended B_n is proposed. In addition, we have developed an algorithm (*uclust*3) that finds the best significant separation in *three* groups. Simulation studies show that our proposal presents coherent results, such as control of Type I Error and the increased Power to identify clusters as they become more separated. Furthermore, our comparative simulation study with other methods shows that in the case where there are exactly *three* groups, the approach we are proposing has greater power, that is, greater ability to correctly identify *three* clusters. More accurate results of *uclust*3 are found in an application to real image recognition data, corroborating the better performance of our approach observed in the simulations. Although we are using Euclidean distance and simulating data with normal distribution, these aspects are not essential to the validity of the method properties.

The steps to developing our three groups clustering method are outlined as follows. First, in Section 2.1 we review the U-statistics based theory of the homogeneity test of [Cybis et al., 2018] and present the U-statistics theory for *three* groups. In Section 2.2 we present the extension of the B_n statistics proposed by [Pinheiro et al., 2009] to contemplate *three* groups in which one may have size one, in order to devise a clustering algorithm that can properly identify outlier elements. Additionally an investigation of theoretical properties that show its compatibility with the previous framework and asymptotic theory, is also presented. In Section 2.3 we explore the variance aspects of the extended B_n and propose an approach to estimate this variance. In Section 3 we propose the *uclust*3 method which finds the statistically significant data partition that better separates the sample into three groups. The remainder of the paper focuses on evaluating the methodology through simulation studies, in Section 4, and an applications to real data in Section 5. Finally, in Section 6 we discuss the overall results.

2 Methods

2.1 U-Statistics based test for three group separation

Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be a random sample of n L-dimensional vectors divided in three groups G_1, G_2 and G_3 of sample sizes n_1, n_2 and n_3, respectively, where $n = n_1 + n_2 + n_3$. In the g-th group, for $g \in \{1, 2, 3\}$, observations $\mathbf{X}_1^{(g)}, \ldots, \mathbf{X}_{n_g}^{(g)}$ are assumed to be independent and identically distributed with a L-variate distribution F_g. Here, the distribution F_g admits finite mean vector $\boldsymbol{\mu}_g$ and positive definite dispersion matrix $\boldsymbol{\Sigma}_g$ (not necessarily multi-normal).

Following the approach of [Sen, 2006] and [Pinheiro et al., 2009], we define the functional distance $\theta(F_g, F_{g'})$ as

$$\theta(F_g, F_{g'}) = \int\int \phi(x_1, x_2) dF_g(x_1) dF_{g'}(x_2), \quad x_1, x_2 \in \mathbb{R}^L, \tag{1}$$

where $g, g' \in \{1, 2, 3\}$ and $\phi(\cdot, \cdot)$ is a symmetric kernel of order 2. If we assume that $\theta(\cdot, \cdot)$ is a convex linear function of its marginal components, then we have

$$\theta(F_g, F_{g'}) \geq \frac{1}{2} \{\theta(F_g, F_g) + \theta(F_{g'}, F_{g'})\}, \tag{2}$$

for all distributions F_g and $F_{g'}$, with equality holding whenever $\mu_g = \mu_{g'}$.

Note that the functional $\theta(\cdot, \cdot)$ can be used to define both distance within and between groups. It follows from U-statistics theory that an unbiased estimator of this functional for within group distance $\theta(F_g, F_g)$ is a generalized U-statistic [Hoeffding, 1948], with kernel $\phi(\cdot, \cdot)$, defined as

$$U_{n_g}^{(g)} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g)}), \tag{3}$$

where $g \in \{1, 2, 3\}$. Analogously, the unbiased estimator for the between group functional distance $\theta(F_g, F_{g'})$ is defined by

$$U_{n_g, n_{g'}}^{(g,g')} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(\mathbf{X}_i^{(g)}, \mathbf{X}_j^{(g')}), \tag{4}$$

where $g, g' \in \{1, 2, 3\}$ and $g \neq g'$.

The combined sample U-statistic is usually decomposed as

$$\begin{aligned} U_n &= \sum_{g=1}^{3} \frac{n_g}{n} U_{n_g}^{(g)} + \sum_{1 \leq g < g' \leq 3} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g, n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\} \\ &= W_n + B_n. \end{aligned} \tag{5}$$

Decomposition (5) leads to the statistic B_n, which provides the focal point of our methodology,

$$B_n = \sum_{1 \leq g < g' \leq 3} \frac{n_g n_{g'}}{n(n-1)} \left\{ 2U_{n_g, n_{g'}}^{(g,g')} - U_{n_g}^{(g)} - U_{n_{g'}}^{(g')} \right\}. \tag{6}$$

Here $U_{n_g}^{(g)}$ for $g \in \{1, 2, 3\}$ are U-statistics associated to within group distances, as defined in (3), and $U_{n_g n_{g'}}^{(g,g')}$, $g \neq g' \in \{1, 2, 3\}$, are the U-statistics associated to between group distances as defined in (4). Note that the definition of $U_{n_g}^{(g)}$ require a minimum of 2 elements in the group. This imposes minimum group sizes $n_g \geq 2$, for $g \in \{1, 2, 3\}$ for proper definition of B_n.

The methodology proposed in [Cybis et al., 2018] and [Valk and Cybis, 2020] considers a group homogeneity test which verifies whether two groups in fact

constitute separated groups, or if they stem from the same distribution. In this work, for data arranged in three groups G_1, G_2 and G_3, the interest is in verifying whether the data are homogeneous or if there is at least one group statistically separated. Thus, the null hypothesis H_0 states that $F_1 = F_2 = F_3$, while the alternative H_1 states that there are $i \neq j$, $\in \{1, 2, 3\}$ where $F_i \neq F_j$. In cases where groups G_1, G_2 and G_3 have more than two elements, the asymptotic properties of B_n are addressed in [Pinheiro et al., 2009]. The statistics B_n is in the class of degenerate U-statistics for which asymptotic normality prevails and the convergence rates are L and/or $\sqrt{n}$. Additionally, under the null, we have $\mathbb{E}(B_n) = 0$ and under the alternative, $\mathbb{E}(B_n) > 0$. The null hypothesis is rejected for large values of standardized B_n, where the variance of B_n, under H_0, is obtained by a resampling procedure [Sen, 2006].

2.2 The extension of test U-statistics for tree groups

The homogeneity test proposed in [Cybis et al., 2018] presents an essential concept for our clustering algorithm. However, the group size restriction required by the definition of the U-statistic B_n in (6) constrains this method to cases where all subgroups have sizes $n_i \geq 2$, $i = 1, 2, 3$, and consequently clustering methods will fail in cases where the data has an outlier. In order to build a clustering algorithm that admits groups of size 1 we propose an extension of B_n. We can assume, without loss of generality, that only the group G_1 may have one element, and define

$$B_n = \begin{cases} \frac{2n_2}{n(n-1)} \left(U^{(1,2)}_{1,n_2} - U^{(2)}_{n_2} \right) + \frac{2n_3}{n(n-1)} \left(U^{(1,3)}_{1,n_3} - U^{(3)}_{n_3} \right) \\ + \frac{n_2 n_3}{n(n-1)} \left(2U^{(2,3)}_{n_2,n_3} - U^{(2)}_{n_2} - U^{(3)}_{n_3} \right), \qquad \text{if } n_1 = 1, \text{ and } n_2, n_3 > 1 \\ \sum_{1 \leq i < j \leq 3} \frac{n_i n_j}{n(n-1)} \left(2U^{(i,j)}_{n_i,n_j} - U^{(i)}_{n_i} - U^{(j)}_{n_j} \right), \qquad \text{if } n_1, n_2, n_3 > 1. \end{cases} \tag{7}$$

where $U^{(g,g')}_{n_g,n_{g'}}$ and $U^{(g)}_{n_g}$ are defined, respectively, in (4) and (3).

This is a natural extension of B_n considering data separation in three groups, when allowing for clusters of size 1. This extension coincides with that of expression (6) for group of sizes $n_1, n_2, n_3 > 1$, and thus all properties mentioned above are still valid for the new definition in that case. We ascertain the validity of these asymptotic properties or analogous alternatives in the case of $n_1 = 1$.

Note that, when G_1 has size one, we can rewrite B_n as

$$\begin{aligned} B_n \quad = \quad & \frac{2n_2}{n(n-1)} U^{(1,2)}_{1,n_2} + \frac{2n_3}{n(n-1)} U^{(1,3)}_{1,n_3} + \frac{2n_2 n_3}{n(n-1)} U^{(2,3)}_{n_2,n_3} \\ & - \frac{n_2(2+n_3)}{n(n-1)} U^{(2)}_{n_2} - \frac{n_3(2+n_2)}{n(n-1)} U^{(3)}_{n_3} \end{aligned}$$

where $U_{1,g}^{(1,g)}$ and $U_{n_g}^{(g)}$, $g = 2, 3$ are as defined in (4) and (3). If we consider the extension of B_n in (7), then we can write the combined sample U-statistics as

$$U_n \quad = \quad B_n + W_n^*.$$

where W_n^* is an appropriate modification the term W_n. Thus, B_n still arises from the decomposition of the combined sample U-statistics into B_n and a modified term W_n. This extended definition allows us to build a U-test when a group has size 1. We conveniently labeled the data in order to arrange the groups as follows. Let $G_1 = \{\mathbf{X}_1\}$, $G_2 = \{\mathbf{X}_2, \ldots, \mathbf{X}_{n_2+1}\}$ and $G_3 = \{\mathbf{X}_{n_2+2}, \ldots, \mathbf{X}_n\}$, $n = 1 + n_2 + n_3$. We still have $\mathbb{E}[B_n] = 0$, under the null hypothesis of overall group homogeneity. Additionally, if we make the assumption that

$$\theta_{gg'} > \theta_g, \tag{8}$$

for $g \neq g' \in \{1, 2, 3\}$ where $\theta_g = \mathrm{E}\left[\phi(X_g, X_g)\right]$ and $\theta_{gg'} = \mathrm{E}\left[\phi(X_g, X_{g'})\right]$, then under alternative we have that $\mathbb{E}[B_n] > 0$. Note that this assumption is usual and when (8) is valid then equation (2) is always satisfied.

Asymptotic theory for the B_n statistic for group sizes greater than 2 is developed in the work of [Pinheiro et al., 2009], where it is established that B_n is a degenerate U-statistic and asymptotic normality is provided. The following theorems demonstrate that the extended B_n is a non degenerated U-statistics and establish the asymptotic distribution of the extended B_n under H_0 for increasing dimension L and sample size n, requiring regularity conditions akin to those of the $n_1, n_2, n_3 > 1$ case.

Theorem 1 Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. $L \times 1$ random vectors. Let $\phi(\cdot, \cdot)$ be a kernel of degree 2 satisfying $\mathbb{E}[\phi(\mathbf{X}_1, \mathbf{X}_2)^2] < \infty$ and $\mathrm{Var}[\mathbb{E}(\phi(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1)] = \sigma_1^2 > 0$. Consider definition (7) for B_n when $n_1 = 1$ and let $V_n = \mathrm{Var}(B_n)$, $\tau_n = (n/2)V_n^{1/2}$ and $W = J_1 + J_2 - J_3 - J_4$, where $\frac{\psi_1(X_1)}{\tau_n} \xrightarrow{D} J_1$, and J_2, J_3 and J_4 are random variables with normal distribution. Then

$$\frac{(n/2)B_n}{\tau_n} \xrightarrow{D} W \text{ as n } \to \infty. \tag{9}$$

Proof: See Supplementary Material.

This result shows that the test statistic asymptotically converges in n to a non-degenerate random variable whose limit distribution depends on the choice of kernel $\phi(\cdot, \cdot)$.

Theorem 2 Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. $L \times 1$ random vectors. Let $\phi(\cdot, \cdot)$ be a kernel of degree 2 such that

$$\phi(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{L}\sum_{l=1}^{L} \phi^*(X_{il}, X_{jl}) \tag{10}$$

for some kernel $\phi^*(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$, where X_{il} is the l-th entry of $\mathbf{X}_i$. Define $\phi_1^*(x_{il}) = \mathbb{E}[\phi^*(X_{il}, X_{jl})|X_{il} = x_{il}]$ and suppose $\mathrm{Var}(\phi_1^*(X_{il})) > 0$ and $\mathrm{Var}(\phi^*(X_{il}, X_{jl})) < \infty$. Let B_n be defined by (7) for the case where $n_1 = 1$, and assume that all conditions in Theorem 1 hold. Suppose also that

$$\sum_{1\le l<m\le n} \mathbb{E}[\phi^*(X_{il}, X_{jl})\phi^*(X_{im}, X_{jm})] = O(L) \tag{11}$$

and

$$\sum_{1\le l<m\le L} \mathbb{E}[\phi_1^*(X_{il})\phi_1^*(X_{jm})] = O(L). \tag{12}$$

Then

$$\frac{B_n}{\sqrt{\mathrm{Var}(B_n)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad L \to \infty. \tag{13}$$

Proof: See Supplementary Material.

This result is fundamental to our inference procedure for clustering in the HDLSS context.

2.3 Variance of B_n

In the *utest* the estimation of B_n's variance under H_0 plays an essential role in hypothesis testing (see [Cybis et al., 2018]). As shown below, even under H_0, the variance of B_n depends on the particular group configuration under consideration. For the homogeneity test of Section 3, we must evaluate this variance for the many group configurations visited in an optimization algorithm. This variance estimation is performed through a resampling procedure, however it becomes computationally expensive to perform one resampling procedure for each individual group size configuration. To circumvent this issue, [Cybis et al., 2018] propose a reweighting scheme taking advantage of analytic calculations for the variance for the case $K = 2$ groups. They are able to compute all variances from a single resampling procedure. In this section we extend their argument to the case of $K = 3$ groups.

In this Section we provide an estimator for the variance of B_n under H_0 based on U-statistics properties of B_n. For cases where all groups have more than two elements, the Hoeffding decomposition of B_n can be found in [Pinheiro et al., 2009] which is given by

$$B_n \;=\; \left(\frac{2}{n(n-1)}\right) \sum_{1\le i<j\le n} \eta_{nij}\psi_2(X_i, X_j), \tag{14}$$

where $\psi_2(\dot{;})$ is the second order term of the Hoeffding decomposition of B_n and

$$\eta_{nij} \;=\; \begin{cases} 1, & \text{if } i \text{ and } j \text{ are from different groups} \\ -\frac{(n-n_g)}{n_g-1}, & \text{if } i \text{ and } j \text{ are from the same group } g. \end{cases} \tag{15}$$

Thereby,

$$\mathrm{Var}(B_n) = \left(\frac{2}{n(n-1)}\right)^2 \tau_2^2 \sum_{1\leq i<j\leq n} \eta_{nij}^2. \tag{16}$$

where $\tau_2^2 = \mathrm{Var}(\psi_2(X_1, X_2))$. From [Pinheiro et al., 2009] we also know that

$$\sum_{1\leq i<j\leq n} \eta_{nij}^2 = \binom{n}{2}(G-1)\left\{1 + \frac{1}{n}\sum_{g=1}^{G} \frac{n-n_g}{(n_g-1)(G-1)}\right\}. \tag{17}$$

For the case in which we have three groups, G_1, G_2 and G_3, with sizes n_1, n_2 and n_3, respectively, where $n_1 + n_2 + n_3 = n$, it can be rewritten as

$$C_n(n_1, n_2) = \sum_{1\leq i<j\leq n} \eta_{nij}^2 = 2\binom{n}{2}\left\{1 + \frac{1}{n}\sum_{g=1}^{3} \frac{n-n_g}{2(n_g-1)}\right\}, \tag{18}$$

and therefore

$$\mathrm{Var}(B_n) \quad = \quad \left(\frac{2}{n(n-1)}\right)^2 \tau_2^2 C_n(n_1, n_2) = V_{n_1,n_2}. \tag{19}$$

Note that only τ_2^2 depends on the probability distribution of the data. Given three groups of sizes n_1, n_2 and n_3, the variance of B_n for this configuration is estimated through a resampling procedure. For optimization purposes, it is not interesting to perform a resampling procedure for each group configuration, so the idea is to use (the relation) expression (19) to estimate B_n's variance for any group configuration from a single resampling procedure. Let G_1^*, G_2^* and G_3^*, with sizes n_1^*, n_2^* and n_3^*, respectively, where $n_1^* + n_2^* + n_3^* = n$, be an other group configuration for the same data set. From (19) it follows that

$$V_{n_1^*,n_2^*} \quad = \quad \frac{C_n(n_1^*, n_2^*)}{C_n(n_1, n_2)} V_{n_1,n_2}. \tag{20}$$

Thus estimating V_{n_1,n_2} through a resampling procedure is sufficient to estimate the variance of B_n for any other group configuration. Although the variance of B_n is estimated under H_0, we note that the choice of n_1 and n_2 may be important to reduce the bias of the variance estimator. To understand the $C_n(\cdot,\cdot)$ function's behavior we plot (18) assuming that $n_1, n_2, n_3 \geq 2$ and $n = n_1 + n_2 + n_3$. As τ_2^2 does not depend on group sizes, the behavior of $C_n(\cdot,\cdot)$ governs the behavior of B_n's variance and Figure 1 shows that smaller values are obtained when groups have balanced sizes, while larger values of $C_n(\cdot,\cdot)$ are obtained when group sizes are unbalanced.

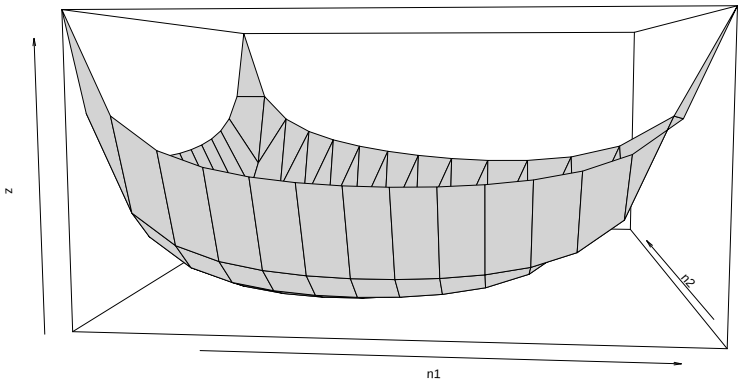

Figure 1: $C_n(\cdot,\cdot)$ function behavior for $n_1, n_2, n_3 \geq 2$ and $n = n_1 + n_2 + n_3$.

2.3.1 Variance of the extended B_n

We propose an extended statistic B_n in (7) to accommodate cases in which the data set is divided into three groups, one of which has size one. For inference purposes it is essential establish a strategy to estimate the variance of the extended B_n. Through the Hoeffding decomposition of (7) (see Suplementary Material) we have that the variance of the extended B_n is

$$\mathrm{Var}(B_n) = \zeta_1(n)\tau_1^2 + \zeta_2(n, n_2)\tau_2^2, \tag{21}$$

where $\tau_1^2 = \mathrm{Var}(\psi_1(X_1))$ and $\tau_2^2 = \mathrm{Var}(\psi_2(X_1, X_2))$ are, respectively, the variance of the first and the second order terms of the Hoeffding decomposition,

$$\begin{aligned} \zeta_1(n) &= \frac{4}{n(n-1)}, \\ \zeta_2(n, n_2) &= \frac{4}{n^2(n-1)} + \frac{4n_2n_3}{n^2(n-1)^2} + \frac{2n_2(2+n_3)^2}{n^2(n_2-1)(n-1)^2} \\ &\quad + \frac{2n_3(2+n_2)^2}{n^2(n_3-1)(n-1)^2}, \end{aligned} \tag{22}$$

$n_1 = 1$, and $n_3 = n - n_2 - 1$. Note that in expression (21) the terms τ_1^2 and τ_2^2 depend on the probability distribution of the data, $\zeta_1(\cdot)$ depends only on n and $\zeta_2(\cdot,\cdot)$ depends on n and n_2 since $n_3 = n - n_2 - 1$. Thus for another group

configuration keeping one of the groups with size one, the only change occurs at n_2, say n_2^*. For this new group configuration, the extended B_n variance is given by

$$\mathrm{Var}(B_n) = \zeta_1(n)\tau_1^2 + \zeta_2(n, n_2^*)\tau_2^2. \tag{23}$$

Again, the choice of n_2 may affect the variance of the estimator. Denoting (21) by V_{n_2} and (23) by $V_{n_2^*}$, we have from simple algebra that

$$V_{n_2^*} \quad = \quad V_{n_2} + [\zeta_2(n, n_2^*) - \zeta_2(n, n_2)]\tau_2^2. \tag{24}$$

For a given n_2 we can estimate V_{n_2} from a resampling procedure. Additionally, an estimate for τ_2^2 can be obtained from the strategy employed to estimate the variance of B_n without outlier through expression (19) as

$$\widehat{\tau}_2^2 \quad = \quad \frac{\widehat{V}_{n_1,n_2}}{C(n_1, n_2)\left(\frac{2}{n(n-1)}\right)^2}. \tag{25}$$

Thus we have a procedure to estimate the extended B_n´s variance for any group configuration from only two independent resampling procedures, through expression

$$\widehat{V}_{n_2^*} \quad = \quad \widehat{V}_{n_2} + [\zeta_2(n, n_2^*) - \zeta_2(n, n_2)]\widehat{\tau}_2^2, \tag{26}$$

where $\widehat{\tau}_2^2$ is obtained from the resampling employed to estimate the variance of B_n without outlier and $\widehat{V}_{n_2}$ is obtained from an additional resampling specific to $n_1 = 1$ case. Thus, taking into account the resampling procedure performed to estimate the variance of B_n when the groups are larger than two and, with one more resampling procedure for the size one group, we have an estimator for extended B_n's variance.

In Figure 2 we have the behavior of $\zeta_2(n, n_2)$ as a function of n_2.

These results are fundamental for the development of feasible algorithms that find significant clusters which is computationally challenging problem.

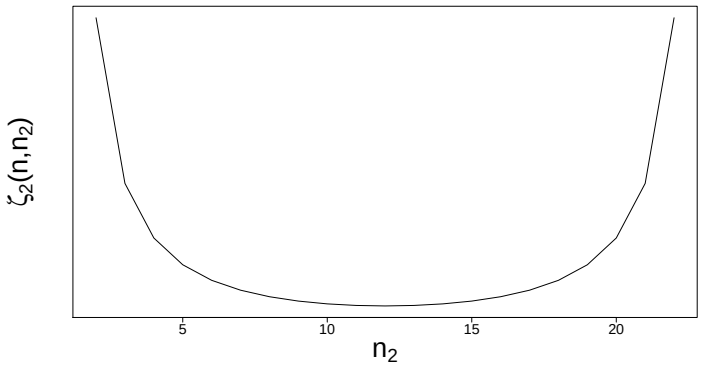

Figure 2: Behavior of function $\zeta_2(n, n_2)$ for a given n, with $n_1 = 1$ and $n = 1 + n_2 + n_3$.

3 Homogeneity test for three groups

Assessment of group homogeneity is a great challenge for standard statistics, especially in the HDLSS context. The *uclust* algorithm presented in [Cybis et al., 2018] and [Valk and Cybis, 2020] is effective to assess overall group homogeneity by verifying whether there exists some significant partition of the data in two groups. Here we are proposing an extension of the *uclust* algorithm for data partitions in three groups G_1, G_2 and G_3. A combinatorial procedure like the one proposed by [Valk and Pinheiro, 2012] in which a *utest* is applied for each possible partition of all group elements into three subgroups has serious computational restrictions due to the exponential increase in the number of tests that need to be performed. We show that (See Section S2 in the Supplementary Material) the number of possible assignments of all n elements in three subgroups is

$$\gamma_3(n) \quad = \quad \frac{233(3^{n-6}) + 1 + n - n^2 - 2^n}{2}, \tag{27}$$

which becomes computationally onerous, especially for large sample size n. To address this issue, we proceed similarly to [Cybis et al., 2018] proposing an optimization procedure to assess group homogeneity by finding the group configuration G_1, G_2 and G_3 that maximizes the objective function

$$f(G_1, G_2, G_3) = \frac{B_n}{\sqrt{\text{Var}(B_n)}}. \tag{28}$$

By maximizing the standardized B_n we must apply only one test. If this three group partition is found significant, then there is at least one subgroup that is significantly different from the others. However, if H_0 is not rejected for this partition, then all other three group partitions will also be non-significant, and the

whole data will be considered homogeneous. While only the group configuration with maximum standardized B_n is tested we have to consider the distribution of B_n's maximum under H_0. Making the untrue, but useful, simplifying assumption that the B_n's are independent for different group configurations, the asymptotic cumulative distribution function of the maximum standardized B_n is given by

$$F_{\max}(x) = \mathbb{P}\left(\max\left(\frac{B_n}{\sqrt{\mathrm{Var}(B_n)}}\right) < x\right) = \Phi(x)^{n^*}, \tag{29}$$

where $n^* = \gamma_3(n)$, for $\gamma_3(n)$ defined in (27) and $\Phi(\cdot)^{n^*}$ is the standard normal cumulative distribution function at the power n^*. For $F_{\max}(x) > 1 - \alpha$, we reject the null hypothesis of overall group homogeneity with α significance level.

The number of tests increases rapidly, even for moderate sample size due to the combinatorial nature of our approach. The maximum distribution in (29) adequately accounts for multiple testing for reasonably small values of n^*. However, this approach has some shortcomings since n^* rapidly increases. Proceeding similarly to [Valk and Cybis, 2020] and considering the simplifying assumption that the B_n's are independent, we use extreme value theory and model it as Gumbel. However, the Gumbel approximation is only valid for very large values of n^*. Thus, for small n we employ the standard max distribution of (29), and when $n^* \geq 2^{28}$ the Gumbel distribution.

3.1 The clustering method *uclust*3

Our homogeneity test in the Section 3 is a method that finds the configuration of *three* subgroups that maximizes the standardized B_n. This is appropriate for the context, since if the homogeneity test accepts the null for this partition, then it would also be accepted for all other partitions. However, the standardized B_n might not be the best criteria to choose between competing partitions when more than one significant group separation exists. This issue is addressed in [Cybis et al., 2018] and arises from the fact that the variance of B_n has different magnitudes depending on subgroup sizes n_1 and n_2 (expression (18) dictates the relationship between variances, which is shown in Figure 1). Consequently, this criteria favours partitions with group sizes of smaller variance, namely $n_1, n_2 \approx n/3$. We note that the magnitude of the variance is quite different when we have a size one group, being much smaller in that case. Again if we use the standardized B_n statistic as a criterion, we will have an effect of choosing groups of size one over the configurations of groups that present greater variance according to the Figure 1.

Considering this issue, we proceed similarly to [Valk and Cybis, 2020] starting by testing overall group homogeneity which is based on maximum of standardized B_n. If the dataset is not homogeneous we adopt instead the maximum B_n as the criteria for finding the configuration that better divides the sample into *three* groups. Thus our significance clustering algorithm *uclust*3 will find the partition with maximum B_n among the universe of all significant partitions

in *three* groups. This is sufficient to ensure that the chosen configuration is statistically significant. However, it is not efficient to find all arrangements of the data in *three* groups that are statistically significant. Furthermore, we cannot simply test the clusters that maximizes B_n since there are non-homogeneous samples for which this maximal partition is not significant.

Based on these characteristics of the B_n we propose a restricted search algorithm, which is based on the behaviors of the B_n's variances (see Figure 1). It starts from the group configuration that maximizes B_n and if that partition is not significant, it searches for partitions whose B_n's variances are smaller than the previous one. This is suitable since only for smaller variances, standardized B_n can be significant. The equation (20) is used to avoid a new resampling procedure to estimate the B_n's variance. As there is a difference in the magnitudes of the B_n's variances (see Figures 1 and 2) this algorithm treats separately the cases when we have a group of size one and the cases with no outlier. The detailed algorithm can be found in Section S3 of the supplementary materials.

4 Simulation Studies

In this section we present simulation studies in order to evaluate some aspects of our proposed methodology. For that we simulate canonical data and use the euclidean distance on our studies, but those are not mandatory for our methods. As presented in Section 2.3, B_n's variance has a behavior that depends on the groups sizes. Moreover when we have a size one group, the order of magnitude of the B_n's variance is quite different when compared to cases in which groups sizes are larger than one. For this reason, our simulations studies typically have a configuration in which a group has size 1 and another configuration in which all groups have more than one element. Figures 1 and 2 show that B_n's variance is smaller at a central group configuration, where the three groups have approximately the same number of elements. Conversely, the variance is greater for extreme group configurations, in which one of the groups has only two elements and the other has $n/2$ elements (or $n-1-n_2$ elements for cases where we have a group of size one). Naturally, the third group's size is defined as $n_3 = n - n_1 - n_2$. These scenarios are explored in our simulation studies.

In the Section 4.1 we evaluate the empirical size and power of the proposed *utest* **for homogeneity of three groups**. Section 4.2 present a simulation study to evaluate the empirical properties of the homogeneity test *uclust*3. The ability to find correct clusters of *uclust*3 and *kmeans* clustering are compared in Section 4.3.

4.1 Simulations for the *utest*

We present here a simulation study to evaluate the empirical performance of the *utest* for three groups. We simulate data from independent normally distributed (i.i.d.) samples divided in three groups G_1, G_2 and G_3. The elements of the L dimensional vectors in G_1 are generated from i.i.d. normal with mean $m_1 = 0$

and standard deviation equal to one. The vectors in G_2 and G_3 have the same properties with mean m_2 and m_3, respectively. In order to allow a graphical representation of the power of the test which is the proportion of rejection considering a significance level α (the power curves), the groups were symmetrically separated and on the x-axis the difference $m_2 - m_1$ is reported. The difference $m_3 - m_2 = m_2 - m_1$. The sample size n takes values in $\{10, 20, 50\}$. Figure 3 presents power curves of the *utest* for *three* groups with separation degree $m_2 - m_1$, where the vectors have dimension $L = 1000$ (gray) and $L = 2000$ (black) and we have 100 replications of each scenario. Furthermore group G_1 has size one and group G_2 was set to have size $n_2 = \lfloor n/3 \rfloor$, where $\lfloor x \rfloor$ means the integer part of x. Naturally the third group's size is defined as $n_3 = n - 1 - n_2$. The significance level used to determine whether the test rejects the null hypothesis that the elements in G_1, G_2 and G_3 have the same distribution was $\alpha = 0.05$.

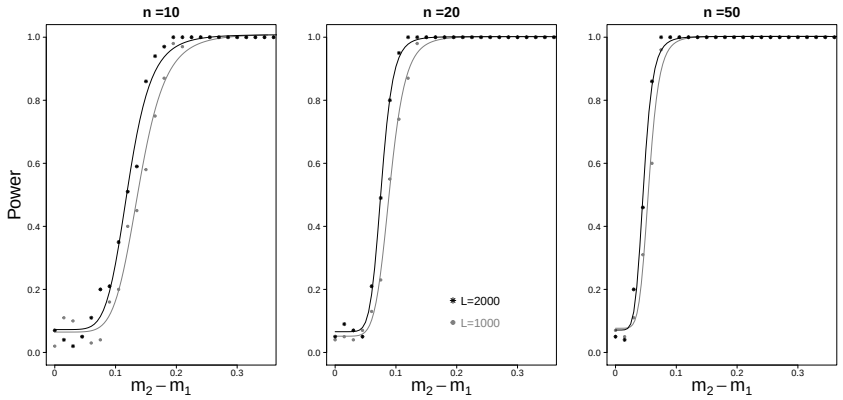

Figure 3: Power curves of *utest* for two dimension $L = 1000$ (gray) and $L = 2000$ (black) for 100 replications of each scenario of $n \in \{10, 20, 50\}$ with $\alpha = 0.05$.

The empirical results obtained in this study reported in Figure 3 corroborate the theoretical properties. As the L increases, the rejection ratio also increases and as the groups become more separated, the power increases. When there is no separation, $m_2 - m_1 = 0$, the rejection ratio is close to the significance level α suggesting control of Type I error. Similar results are found for cases where all groups have more than one element (see Figure S1 in the Supplementary Material).

4.2 Simulations for the homogeneity test in *uclust*3

To evaluate the statistical properties of the homogeneity test *uclust*3 considering the max distribution (29) with the Gumbel correction when appropriate, we simulate data with the same characteristics as the data in Section 4.1. For each sample size n in $\{10, 20, 50\}$, group G_1 has size one and group G_2 was set to have size $n_2 = 2$ and $n_2 = n/2$, and consequently the third group's size was defined as $n_3 = n - 1 - n_2$. Table 1 shows the proportion of rejection of the null hypothesis for significance level $\alpha = 0.05$ considering two scenarios of (m_2, m_3) and the dimension L taking values in $\{1000, 2000\}$.

Table 1: Empirical power of the homogeneity test *uclust*3 with a group of size one

n	$(m_2,\ m_3)$	(n_2)	Dimension L	
			1000	2000
10	(0.25, 0.5)	2	0.27	0.36
		5	0.69	0.89
	(0.5, 1)	2	0.22	0.25
		5	0.98	1
20	(0.25, 0.5)	2	0.93	1
		10	1	1
	(0.5, 1)	2	0.9	0.89
		10	0.92	1
50	(0.25, 0.5)	2	0.68	0.68
		25	1	0.99
	(0.5, 1)	2	0.99	0.96
		25	1	1

We can observe that even in an extreme group configuration, where the group G_1 has size one and the group G_2 has size two, the method presents consistent empirical power to reject the null hypothesis. The power increases as L and/or n and/or the difference between m_2 and m_3 increases, emphasizing the inherent properties of the method.

Supplementary Table S1 presents estimates of type I error rates for *uclust*3. The significance level considered in this simulations was $\alpha = 0.05$ and we can observe that the method presents an adequate control of the Type I Error for cases where $L >> n$ (typically HDLSS scenario). Supplementary Table S2 presents power of the *uclust*3 for group configurations of sizes greater than 1. For small sample size n the test had more difficulty in finding the correct clusters. However, for larger n the method showed an excellent performance.

4.3 Simulations for finding correct clusters

In order to evaluate the accuracy of our clustering method, we present simulation studies comparing *uclust*3 with *kmeans* clustering, one of the most popular clustering algorithms. We refer the reader to the vastly cited work of [Jain, 2010] for a general discussion about *kmeans*. The data were simulated under the same distribution scheme of Section 4.2, with $Re = 100$ replications and the methods were compared in terms of mean Adjusted Rand Index (ARI) which measures the agreement of clustering results with simulation scenarios, adjusting for randomness [Hubert and Arabie, 1985]. An ARI of one indicates perfect matching. No inference is used in this analysis. This is an appropriate comparison as both methods are set to find exactly three groups. Table 2 reports the results for three sample sizes $n \in \{10, 20, 50\}$, two dimension $L \in \{1000, 2000\}$ and three groups of sizes n_1, n_2 and $n_3 = n - n_1 - n_2$. The data vectors in group G_1 have zero mean and the data vectors in G_2 and G_3 have mean m_2 and m_3, respectively. Note that the clustering method *uclust*3, based on the maximization of B_n is comparable to *kmeans* to find the correct clusters, considering this data configuration. However for larger sample sizes, as the clusters become better defined, with greater separation between the means, *uclust*3 outperforms *kmeans*. Table S3 shows that for the case where G_1 has size one, *kmeans* tends to perform slightly better for smaller sample sizes.

Table 2: Comparison of mean ARI and standard deviation (Sd) of the accuracy in clustering of *kmeans* and *uclust*3 methods.

n	$(m_2,\ m_3)$	$(n_1,\ n_2)$	Method	Dimension L			
				1000		2000	
				Mean	Sd	Mean	Sd
10	(0.25, 0.5)	(2, 5)	*kmeans*	0.59	0.05	0.73	0.06
			*uclust*3	0.58	0.03	0.63	0.02
		(3, 3)	*kmeans*	0.56	0.05	0.74	0.08
			*uclust*3	0.52	0.05	0.6	0.05
	(0.5, 1)	(2, 5)	*kmeans*	0.91	0.04	0.94	0.03
			uclust3	0.74	0.01	0.74	0
		(3, 3)	*kmeans*	0.9	0.05	0.87	0.07
			*uclust*3	0.92	0.03	0.96	0.02
20	(0.25, 0.5)	(2, 10)	*kmeans*	0.73	0.02	0.77	0.03
			*uclust*3	0.7	0.02	0.74	0.02
		(6, 6)	*kmeans*	0.74	0.05	0.94	0.03
			*uclust*3	0.68	0.04	0.91	0.02
	(0.5, 1)	(2, 10)	*kmeans*	0.96	0.01	0.94	0.02
			*uclust*3	1	0	1	0
		(6, 6)	*kmeans*	0.81	0.07	0.84	0.07
			*uclust*3	1	0	1	0
50	(0.25, 0.5)	(2, 25)	*kmeans*	0.76	0.01	0.79	0.01
			*uclust*3	0.73	0	0.74	0.01
		(16, 16)	*kmeans*	0.93	0.02	0.89	0.05
			*uclust*3	0.94	0	1	0
	(0.5 , 1)	(2 , 25)	*kmeans*	0.95	0.01	0.95	0.01
			*uclust*3	1	0	1	0
		(16, 16)	*kmeans*	0.8	0.07	0.81	0.07
			*uclust*3	1	0	1	0

4.4 Finding correct clusters and comparing *uclust*3 and *uhclust* in a presence of an outlier

A simulation study similar to Section 4.1 was performed to compare our *uclust*3 with the hierarchical methods *uhclust* from [Valk and Cybis, 2020] and *sigclust* from [Kimes et al., 2017, Kimes, 2019] in terms of the ability to correctly find statistically significant groups. The group G_1 has only one element, the size of G_2 is $n_2 = \lfloor n/3 \rfloor$. For all three methods the same level of significance $\alpha = 0.05$ was considered. The *sigclust* method was not able to find the correct groups in any scenario, with a proportion of correct answers equal to zero and for this reason it was excluded from the analysis. Figures 4 and 5 report curves of proportion times that the algorithms found significant separation and correct groups considering different values of $m_2 - m_1$ varying on the x axis, with sample

size n taking values in $\{10, 20, 50\}$ and dimension $L = 1000$ and $L = 2000$ The results are based on 50 repetitions.

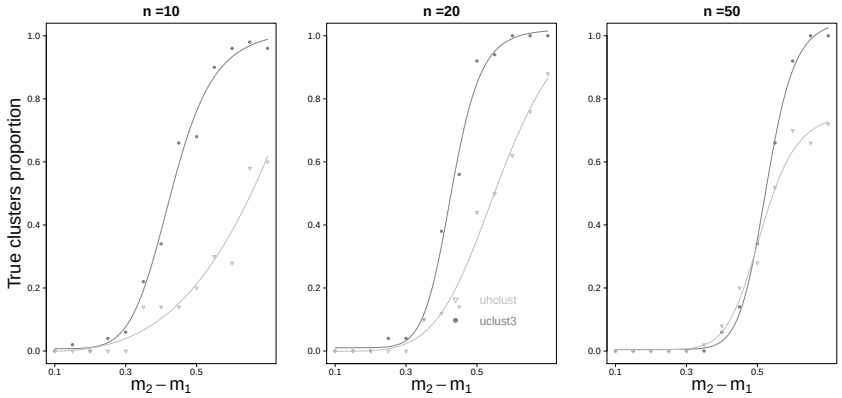

Figure 4: True cluster proportion curves of *uclust*3 (dark gray) and *uhclust* (light gray) for dimension $L = 1000$ with 50 replications of each scenario of n with $\alpha = 0.05$ and one outlier.

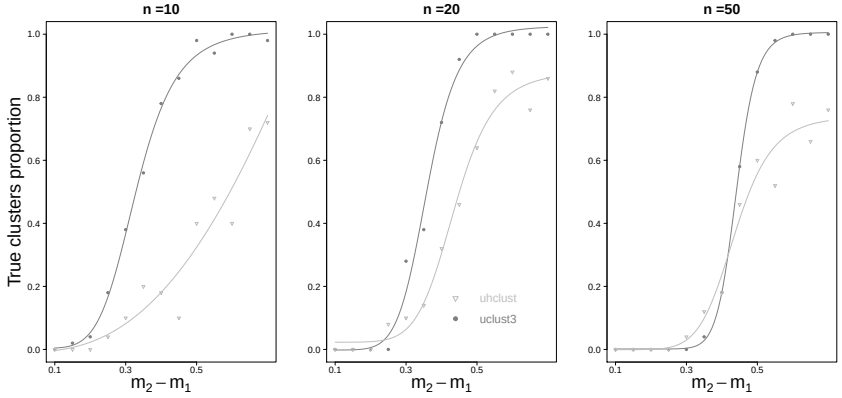

Figure 5: True cluster proportion curves of *uclust*3 (dark gray) and *uhclust* (light gray) for dimension $L = 2000$ with 50 replications of each scenario of n with $\alpha = 0.05$ and one outlier.

The *uclust*3 method (dark grey) outperforms *uhclust* method (light gray) in all scenarios presenting greater ability to find the correct groups for less separation. However, for $n = 50$ these method are more competitive although the method proposed here *uclust*3 still stands out for larger separations. The conclusions do not change with the variation of dimension L. In Section S5 on the supplementary materials we present results of a simulation study for the

cases where there are no outlier. Supplementary Figures S2 and S3 shows the true cluster proportion curves of *uclust*3 and *uhclust* for dimension $L = 1000$ and $L = 2000$. We note that the *uclust*3 method outperforms *uhclust* in all scenarios.

5 Application

We consider a simple example of image recognition to illustrate the applicability of our methodology. The data consists of images from three public figures (Tony Blair, Colin Powell and George W. Bush) which were selected from the Labeled Faces Wild (LFW) dataset ([Huang et al., 2007]). The data were run through OpenFace's convolutional neural network ([Amos et al., 2016]), a procedure that outputs a 128-dimensional representation of the faces which preserves Euclidean distances. In case the reader wants to know more about how the OpenFace works, we recommend reading their website [Amos et al., 2016]. In this illustrative application, we randomly select 10 images from each public figure in the above cited dataset and analysed them using OpenFace. To visualize the data, we provided a heatmap of the transformed images as follows.

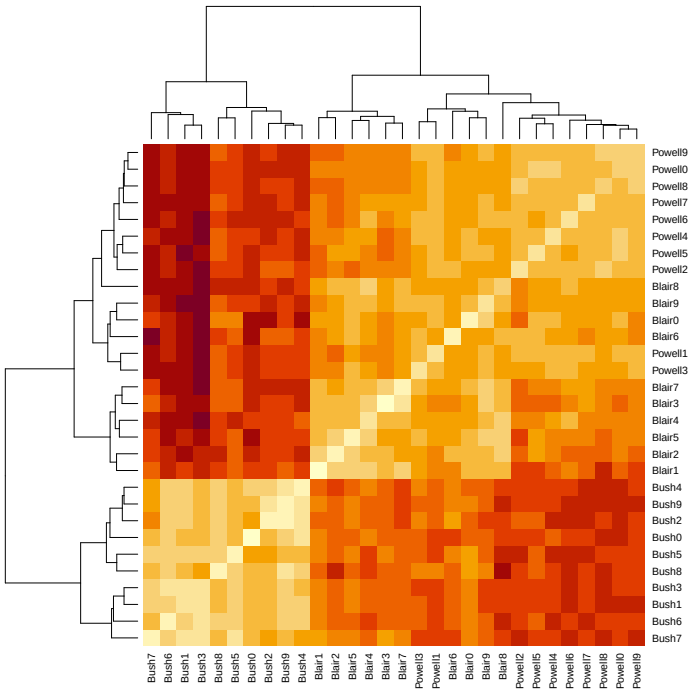

Figure 6: Heatmap of public figures

Then we run *uhclust*, *sigclust* and *uclust*3 with significance level $\alpha = 0.05$. Figure 7 presents the hierarchical clustering dendrogram annotated with p-values for all tests performed in the *uhclust* method. We found 4 homogeneous groups, with a significant division in the Bush image group and an ARI=0.8585. Figure 8 presents the dendrogram with corresponding *sigclust* analysis of the same data which produces six significant clusters, segregating Bush and Powell's images from the reminder and finding one outlier in Blair's group. The ARI for this case was 0.7788. Applying the *uhclust*3 method we found exactly 3 homogeneous groups, each corresponding to one of the public figures with ARI=1.

In the Section S6 in the supplementary materials we consider the same dataset and public figures to carry out an analysis with three groups in which one has size one. Figures S4 and S5 in the supplementary materials present the clustering dendrogram annotated with results of all tests performed in the *uhclust* and *sigclust* methods. None of these methods were able to identify the outlier and both methods achieved ARI of 0.8135593. However, when we applied the *uclust*3 method we found the correct groups with ARI of 1, supporting the best results *uclust*3 in the simulation study.

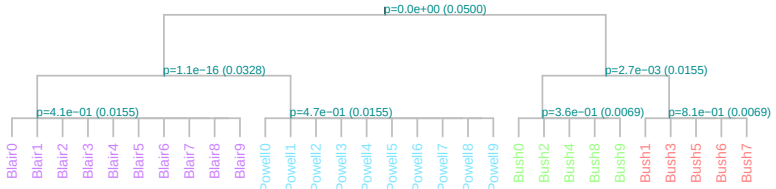

Figure 7: Annotated dendrogram of significance analysis for hierarchical clustering *uhclust* for 30 pictures of 3 public figures. P-values and corrected significance levels α^* are shown for each test performed at the corresponding node.

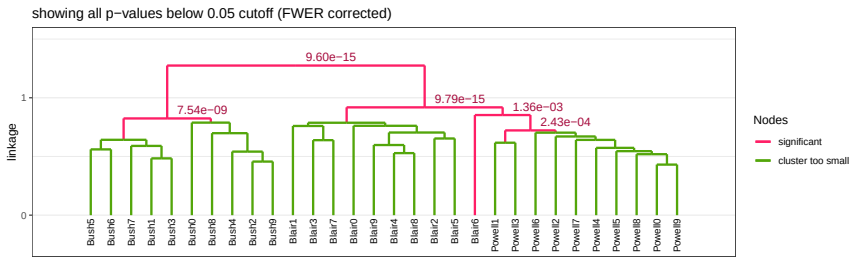

Figure 8: Annotated dendrogram of significance analysis for hierarchical clustering *sigclust* for 30 pictures of 3 public figures. P-values and corrected significance levels α^* are shown for each test performed at the corresponding node.

6 Discussion

We have developed a clustering method that separates a dataset specifically into *three* groups allowing the assessment of significance of this partition. Our methodology is based on the U-statistics clustering framework proposed in [Pinheiro et al., 2009] and is an extension of the approach of [Cybis et al., 2018, Valk and Cybis, 2020]. Considering the B_n statistic of [Pinheiro et al., 2009] that aims to test homogeneity of three predefined groups we propose an extension of the B_n statistic to allow for an outlier, namely one of the groups has only one element ($n_1 = 1$). Additionally we verified statistical properties that ensure the compatibility of this new definition with the overall framework. We then considered group homogeneity testing with this newly defined statistic, and explored empirical properties such as Type I error control and power, showing adequate preformance. Afterwards, we extended this framework to address the issue of partitioning a dataset into three optimal statistically significant clusters, proposing a new clustering criteria that defines the *uclust*3 method. This differs from previous methods for instead of find and testing a two group separation, *uclust*3 finds the best significant partitions in three clusters. This can pave the way for inference in K groups.

This U-statistics based methodology can be applied to a wide range of problems, since they make very few assumptions about the distribution of the data. Although in the simulation study and in the application we have used Euclidean distance, this is not a necessary requirement for theory development. Additionally, even if the data come from a non-normal multivariate distribution, the required asymptotic normality is guaranteed as long as the distances have finite variance and the sum of all distance covariances do not grow too fast ($O(L)$ see Theorem 2). The clustering procedures *uclust*3 proposed here require large L since B_n for $n_1 = 1$ is only asymptotically normal in the dimension L. As verified in previously work of [Valk and Cybis, 2020], for the settings in the simulation studies, in practice our tests achieve good Type I error control having difficulties only when L is smaller than $10n$. This is, by excellence, the HDLSS

setting.

An important step for developing the homogeneity test is to establish the number of possible configurations of n elements separated in three groups. A system of recursive equations was developed to solve this combinatorial problem and the idea may be used to solve an equivalent problem involving $K > 3$ groups.

The significance clustering method *uclust*3 proposed here returns the partition that better separates the data into three statistically significant groups in terms of the B_n statistic. Thus we can compare it with *kmeans*, which is one of the most popular clustering method, regarding the ability of correctly find three groups. A simulation study suggests that *uclust*3 is competitive with *kmeans* when we have a size one group and outperforms *kmeans* in the context in which groups having an underlying cluster structure with more than 2 elements each and large sample sizes.

Since our methodology is a natural extension of the *uclust* method proposed by [Valk and Cybis, 2020] it inherits many helpful properties such as the ability to avoid the hazards of directly estimating the covariance matrix, by obtaining $\text{Var}(B_n)$ through resampling. However, they have different purposes, while *uclust* aims to find the best significant partition in two groups, *uclust*3 aims to find the best significant separation in three groups, so they are not directly comparable. To support the usefulness of the *uclust*3, we carried out a simulation study to compare this method with the hierarchical version of *uclust* (*uhclsut*) and with another hierarchical approach (*sigclust*), which both are able to find a significantly partition into three groups, when this partition exists. We simulated normal data with a three group structure, separating these groups in terms of the means and use the proportion of correct configurations found to compare the methods. In the situations considered, *sigclust* had serious difficulties in finding the proper arrangement, while *uclust*3 performed better than *uhclust* in all scenarios. Additionally, in the application to image recognition data sets we select three public figures and observe that the *uclust*3 method was the only one able to correctly find the three groups of figures.

Finally the conclusion is that our *uclust*3 method is appropriate to separate a high dimensional low sample size dataset into *three* groups, being more powerful than some other methods in the specific situation in which a structure of three groups is present in the dataset.

Supplementary material

Supplementary material: Derivations, supplementary tables and figures (pdf)

Code: R-functions containing all methods developed in this article (will be available in the uclust package at CRAN).

Data: Dataset used in the application and corresponding script (zip).

Acknowledgements

We would like to thanks Pedro Fusieger for the fruitful discussions about the number of possible assignments of all n elements in three subgroups.

Funding

Debora Zava Bello research was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

[Adolfsson et al., 2019] Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26.

[Amos et al., 2016] Amos, B., Ludwiczuk, B., and Satyanarayanan, M. (2016). OpenFace: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

[Chen et al., 2015] Chen, G. K., Chi, E. C., Ranola, J. M. O., and Lange, K. (2015). Convex clustering: an attractive alternative to herarchical clustering. *PLoS Computational Biology*, 11(5):e1004228.

[Cybis et al., 2018] Cybis, G. B., Valk, M., and Lopes, S. R. C. (2018). Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*.

[Demidenko, 2018] Demidenko, E. (2018). The next-generation k-means algorithm. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(4):153–166.

[Euan et al., 2019] Euan, C., Sun, Y., Ombao, H., et al. (2019). Coherence-based time series clustering for statistical inference and visualization of brain connectivity. *Annals of Applied Statistics*, 13(2):990–1015.

[Helgeson et al., 2020] Helgeson, E. S., Vock, D. M., and Bair, E. (2020). Nonparametric cluster significance testing with reference to a unimodal null distribution. *Biometrics*.

[Hennig, 2015] Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, 64:53–62.

[Hoeffding, 1948] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.

[Huang et al., 2007] Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

[Jain, 2010] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

[Kimes, 2019] Kimes, P. (2019). pkimes/sigclust2 documentation. (Accessed on 02/02/2021).

[Kimes et al., 2017] Kimes, P. K., Liu, Y., Neil Hayes, D., and Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821.

[Liu et al., 2008] Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. (2008). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293.

[McLachlan and Peel, 2004] McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[McShane et al., 2002] McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469.

[Motlagh et al., 2019] Motlagh, O., Berry, A., and O'Neil, L. (2019). Clustering of residential electricity customers using load time series. *Applied energy*, 237:11–24.

[Pinheiro et al., 2009] Pinheiro, A., Sen, P. K., and Pinheiro, H. P. (2009). Decomposability of high-dimensional diversity measures: Quasi-u-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*.

[Rosenberg et al., 2002] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

[Sen, 2006] Sen, P. K. (2006). Robust statistical inference for high-dimensional data models with application to genomics. *Austrian journal of statistics*, 35(2&3):197–214.

[Shimodaira et al., 2004] Shimodaira, H. et al. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *The Annals of Statistics*, 32(6):2616–2641.

[Suzuki and Shimodaira, 2006] Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542.

[Valk and Cybis, 2020] Valk, M. and Cybis, G. B. (2020). U-statistical inference for hierarchical clustering. *Journal of Computational and Graphical Statistics*.

[Valk and Pinheiro, 2012] Valk, M. and Pinheiro, A. (2012). Time-series clustering via quasi u-statistics. *Journal of Time Series Analysis*.

[Von Luxburg et al., 2012] Von Luxburg, U., Williamson, R. C., and Guyon, I. (2012). Clustering: Science or art? In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 65–79. JMLR Workshop and Conference Proceedings.

APÊNDICE 1 – Material suplementar

Autores: Débora Zava Bello, Marcio Valk e Gabriela Cybis

Título: Supplementary material: Clustering inference in multiple groups

Ano: 2021.

Supplementary material: Clustering inference in multiple groups

Debora Zava Bello, Marcio Valk and Gabriela Cybis

March 2021

S1 The extended B_n for three groups

In this work we propose an extension of the statistic B_n for three groups allowing for a size one group. This extension, as shown in the Section 2.2 of the main manuscript, was defined as

$$B_n = \begin{cases} \frac{2n_2}{n(n-1)}\left(U^{(1,2)}_{1,n_2} - U^{(2)}_{n_2}\right) + \frac{2n_3}{n(n-1)}\left(U^{(1,3)}_{1,n_3} - U^{(3)}_{n_3}\right) \\ +\frac{n_2n_3}{n(n-1)}\left(2U^{(2,3)}_{n_2,n_3} - U^{(2)}_{n_2} - U^{(3)}_{n_3}\right), \qquad \text{if } n_1 = 1, \text{ and } n_2, n_3 > 1 \\ \sum_{1\le i<j\le 3} \frac{n_i n_j}{n(n-1)}\left(2U^{(i,j)}_{n_i,n_j} - U^{(i)}_{n_i} - U^{(j)}_{n_j}\right), \qquad \text{if } n_1, n_2, n_3 > 1. \end{cases} \tag{S.1}$$

where $U^{(g,g')}_{n_g,n_{g'}}$ and $U^{(g)}_{n_g}$ are defined, respectively, in equations (3) and (4) in the manuscript. As properties of B_n are well described for cases where groups have more than one element we focus on the special case in which one of the groups has size one. Without loss of generality assume that $n_1 = 1$ and $n_2, n_3 > 1$. Thus B_n becomes

$$\begin{aligned} B_n &= \frac{2n_2U^{(1,2)}_{1,n_2}}{n(n-1)} - \frac{2n_2U^{(2)}_{n_2}}{n(n-1)} + \frac{2n_3U^{(1,3)}_{1,n_3}}{n(n-1)} - \frac{2n_3U^{(3)}_{n_3}}{n(n-1)} + \frac{2n_2n_3U^{(2,3)}_{n_2,n_3}}{n(n-1)} \\ &\quad - \frac{n_2n_3U^{(2)}_{n_2}}{n(n-1)} - \frac{n_2n_3U^{(3)}_{n_3}}{n(n-1)} \\ &= \frac{2n_2U^{(1,2)}_{1,n_2}}{n(n-1)} + \frac{2n_3U^{(1,3)}_{1,n_3}}{n(n-1)} + \frac{2n_2n_3U^{(2,3)}_{n_2,n_3}}{n(n-1)} - \frac{n_2(2+n_3)U^{(2)}_{n_2}}{n(n-1)} - \\ &\quad \frac{n_3(2+n_2)U^{(3)}_{n_3}}{n(n-1)}. \end{aligned}$$

where $U_{n_k}^{(k)} = \binom{n_k}{2}^{-1} \sum_{1 \le i < j \le k} \phi(X_{ki}, X_{kj})$ and
$U_{n_g, n_{g'}} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{i=1}^{n_{g'}} \phi(X_{gi}, X_{g'j})$.
The Hoeffding decomposition of B_n is

$$
\begin{aligned}
B_n &= \frac{2n_2}{n(n-1)} \left[\frac{1}{n_2} \sum_{i=1}^{n_2} \phi(X_1, X_{2i}) \right] + \frac{2n_3}{n(n-1)} \left[\frac{1}{n_3} \sum_{j=1}^{n_3} \phi(X_1, X_{3j}) \right] + \\
&\quad + \frac{2n_2 n_3}{n(n-1)} \left[\frac{1}{n_2 n_3} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \phi(X_{2i}, X_{3j}) \right] \\
&\quad - \frac{n_2(2+n_3)}{n(n-1)} \left[\binom{n_2}{2}^{-1} \sum_{1 \le i < j \le n_2} \phi(X_{2i}, X_{2j}) \right] \\
&\quad - \frac{n_3(2+n_2)}{n(n+1)} \left[\binom{n_3}{2}^{-1} \sum_{1 \le i < j \le n_3} \phi(X_{3i}, X_{3j}) \right] \\
&= \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \phi(X_1, X_{2i}) + \frac{2}{n(n-1)} \sum_{j=1}^{n_3} \phi(X_1, X_{3j}) + \\
&\quad + \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \phi(X_{2i}, X_{3j}) \\
&\quad - \frac{n_2(2+n_3)}{n(n-1)} \frac{2}{n_2(n_2-1)} \sum_{1 \le i < j \le n_2} \phi(X_{2i}, X_{2j}) \\
&\quad - \frac{n_3(2+n_2)}{n(n-1)} \frac{2}{n_3(n_3-1)} \sum_{1 \le i < j \le n_3} \phi(X_{3i}, X_{3j}).
\end{aligned}
$$

Is known from the theory of U-statistics (see [Hoeffding, 1948]) that the kernel $\phi(,)$ can be expressed as sum of orthogonal components, $\phi(X_i, X_j) = \psi_1(X_i) + \psi_1(X_j) + \psi_2(X_i, X_j) + \theta$, where $\psi_1(X_i) = \mathbb{E}[\phi(X_i, X_j)|X_i]$, and $\psi_2(X_i, X_j) = \mathbb{E}[\phi(X_i, X_j)|X_i, X_j]$.

Then,

$$
\begin{aligned}
B_n \quad = \quad & \\
& \frac{2}{n(n-1)} \sum_{i=1}^{n_2} [\psi_1(X_1) + \psi_1(X_{2i}) + \psi_2(X_1, X_{2i}) + \theta] + \\
& + \frac{2}{n(n-1)} \sum_{j=1}^{n_3} [\psi_1(X_1) + \psi_1(X_{3j}) + \psi_2(X_1, X_{3j}) + \theta] + \\
& + \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} [\psi_1(X_{2i}) + \psi_1(X_{3j}) + \psi_2(X_{2i}, X_{3j}) + \theta] + \\
& + \left(- \frac{2(2+n_3)}{n(n-1)(n_2-1)} \right) \sum_{1 \le i < j \le n_2} [\psi_1(X_{2i}) + \psi_1(X_{2j}) \\
& + \psi_2(X2i, X_{2j}) + \theta] + \\
& + \left(- \frac{2(2+n_2)}{n(n-1)(n_3-1)} \right) \sum_{1 \le i < j \le n_3} [\psi_1(X_{3i}) + \psi_1(X_{3j}) + \\
& + \psi_2(X_{3i}, X_{3j}) + \theta] \\
= \quad & \theta \left[\frac{2n_2}{n(n-1)} + \frac{2n_3}{n(n-1)} + \frac{2n_2 n_3}{n(n-1)} - \frac{2(2+n_3)}{n(n-1)(n_2-1)} \frac{n_2(n_2-1)}{2} \right. \\
& \left. - \frac{2(2+n_2)}{n(n-1)(n_3-1)} \frac{n_3(n_3-1)}{2} \right] + \psi_1(X_1) \left[\frac{2n_2}{n(n-1)} + \right. \\
& \left. + \frac{2n_3}{n(n-1)} \right] + \sum_{i=1}^{n_2} \psi_1(X_{2i}) \left[\frac{2}{n(n-1)} + \frac{2n_3}{n(n-1)} - \frac{2(2+n_3)}{n(n-1)} \right] + \\
& + \sum_{j=1}^{n_3} \psi_1(X_{3j}) \left[\frac{2}{n(n-1)} + \frac{2n_2}{n(n-1)} - \frac{2(2+n_2)}{n(n-1)} \right] \\
& + \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \psi_2(X_1, X_{2i}) + \frac{2}{n(n-1)} \sum_{j=1}^{n_3} \psi_2(X_1, X_{3j}) + \\
& + \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \psi_2(X_{2i}, X_{3j}) \\
& - \frac{2(2+n_3)}{n(n-1)(n_2-1)} \sum_{1 \le i < j \le n_2} \psi_2(X_{2i}, X_{2j}) \\
& - \frac{2(2+n_3)}{n(n-1)(n_3-1)} \sum_{1 \le i < j \le n_3} \psi_2(X_{3i}, X_{3j}) \\
= \quad & \psi_1(X_1) \left(\frac{2}{n} \right) + \sum_{i=1}^{n_2} \psi_1(X_{2i}) \left(\frac{2 + 2n_3 - 4 - 2n_3}{n(n-1)} \right) + \\
& + \sum_{j=1}^{n_3} \psi_1(X_{3j}) \left[\frac{2 + 2n_2 - 4 - 2n_2}{n(n-1)} \right] + \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \psi_2(X_1, X_{2i}) + \\
& + \frac{2}{n(n-1)} \sum_{j=1}^{n_3} \psi_2(X_1, X_{3j}) + \frac{2}{3} \frac{2}{n(n-1)} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \psi_2(X_{2i}, X_{3j}) \\
& - \frac{2(2+n_3)}{n(n-1)(n_2-1)} \sum_{1 \le i < j \le n_2} \psi_2(X_{2i}, X_{2j}) \\
& - \frac{2(2+n_2)}{n(n-1)(n_3-1)} \sum_{1 \le i < j \le n_3} \psi_2(X_{3i}, X_{3j})
\end{aligned}
$$

Thus, the Hoeffding decomposition of B_n for size one group case is

$$\begin{aligned} B_n &= \frac{2}{n}\Bigg[\psi_1(X_1) - \frac{1}{n-1}\sum_{i=1}^{n_2}\psi_1(X_{2i}) - \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_1(X_{3j}) + \\ &+ \frac{1}{n-1}\sum_{i=1}^{n_2}\psi_2(X_1, X_{2i}) + \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_2(X_1, X_{3j}) + \\ &+ \frac{1}{n-1}\sum_{i=1}^{n_2}\sum_{j=1}^{n_3}\psi_2(X_{2i}, X_{3j}) - \frac{(2+n_3)}{(n-1)(n_2-1)}\sum_{1\le i<j\le n_2}\psi_2(X_{2i}, X_{2j}) \\ &- \frac{(2+n_2)}{(n-1)(n_3-1)}\sum_{1\le i<j\le n_3}\psi_2(X_{3i}, X_{3j})\Bigg] \end{aligned}$$

S1.1 Finite sample properties of B_n

Let $\mathrm{E}\left[\phi(X_g, X_g)\right] = \theta_g$ and $\mathrm{E}\left[\phi(X_g, X_{g'})\right] = \theta_{gg'}$, then

$$\begin{aligned} \mathrm{E}(B_n) &= \frac{2n_2\theta_{12}}{n(n-1)} + \frac{2n_3\theta_{13}}{n(n-1)} + \frac{2n_2n_3\theta_{23}}{n(n-1)} \\ &\quad - \frac{2(2+n_3)}{n(n-1)(n_2-1)}\frac{n_2(n_2-1)}{2}\theta_2 \\ &\quad - \frac{2(2+n_2)}{n(n-1)(n_3-1)}\frac{n_3(n_3-1)}{2}\theta_3 \\ &= \frac{1}{n(n-1)}\left[2n_2\theta_{12} + 2n_3\theta_{13} + 2n_2n_3\theta_{23} - n_2(2+n_3)\theta_2\right. \\ &\quad \left. -n_3(2+n_2)\theta_3\right] \\ &= \frac{1}{n(n-1)}\left[n_2(2\theta_{12} - 2\theta_2) + n_3(2\theta_{13} - 2\theta_3) + \right. \\ &\quad \left. +n_2n_3(\theta_{23} - \theta_2) + n_2n_3(\theta_{23} - \theta_3)\right] \end{aligned}$$

Under the null hypothesis H_0, $\theta_g = \theta_{gg'}$ and clearly $E(B_n) = 0$. Under the alternative H_1, $\mathbb{E}(B_n) > 0$ since we have $\theta_{gg'} > \theta_g$, for all $g \neq g' \in \{1, 2, 3\}$. This condition was already required in the work of [Valk and Cybis, 2020].

For accessing the B_n's variance we handle with Hoeffding decomposition of $\frac{n}{2}B_n$ and otain $\mathrm{Var}\left(\frac{n}{2}B_n\right)$. It follows that

$$
\begin{aligned}
\frac{n}{2}B_n &= \psi_1(X_1) - \frac{1}{n-1}\sum_{i=1}^{n_2}\psi_1(X_{2i}) - \frac{1}{n-1}\psi_1(X_{3j}) + \\
&+\frac{1}{n-1}\sum_{i=1}^{n_2}\psi_2(X_1, X_{2i}) + \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_2(X_1, X_{3j}) + \\
&+\frac{1}{n-1}\sum_{i=1}^{n_2}\sum_{j=1}^{n_3}\psi_2(X_{2i}, X_{3j}) \\
&-\frac{(2+n_3)}{(n-1)(n_2-1)}\sum_{1\le i<j\le n_2}\psi_2(X_{2i}, X_{3j}) \\
&-\frac{(2+n_2)}{(n-1)(n_3-1)}\sum_{1\le i<j\le n_3}\psi_2(X_{3i}, X_{3j})
\end{aligned}
$$

Define $\tau_1^2 = \mathrm{Var}\left[\psi_1(X_1)\right]$ and $\tau_2^2 = \mathrm{Var}\left[\psi_2(X_1, X_2)\right]$. Then, under H_0 when we have a size one group

$$
\begin{aligned}
\mathrm{Var}\left(\frac{n}{2}B_n\right) &= \tau_1^2 + \left(\frac{1}{n-1}\right)^2\sum_{j=1}^{n_2}\tau_1^2 + \left(\frac{1}{n-1}\right)^2\sum_{j=1}^{n_3}\tau_1^2 + \\
&+\left(\frac{1}{n-1}\right)^2\sum_{i=1}^{n_2}\tau_2^2 + \left(\frac{1}{n-1}\right)^2\sum_{j=1}^{n_3}\tau_2^2 + \\
&+\left(\frac{1}{n-1}\right)^2\sum_{i=1}^{n_2}\sum_{j=1}^{n_3}\tau_2^2 + \\
&+\left[\frac{(2+n_3)}{(n-1)(n_2-1)}\right]^2\sum_{1\le i<j\le n_2}\tau_2^2 + \\
&+\left[\frac{(2+n_2)}{(n-1_(n_3-1)}\right]^2\sum_{1\le i<j\le n_3}\tau_2^2 \\
&= \tau_1^2 + \frac{1}{n-1}\tau_1^2 + \frac{1}{n-1}\tau_1^2 + \\
&+\frac{n_2n_3}{(n-1)^2}\tau_2^2 + \frac{n_2(2+n_3)^2}{2(n_2-1)(n-1)^2}\tau_2^2 + \frac{n_3(2+n_2)^2}{2(n_3-1)(n-1)^2}\tau_2^2 \\
&= \tau_1^2\frac{n}{n-1} + \tau_2^2\left[\frac{1}{n-1} + \frac{n_2n_3}{(n-1)^2} + \right. \\
&\left.+\frac{n_2(2+n_3)^2}{2(n_2-1)(n-1)^2} + \frac{n_3(2+n_2)^2}{2(n_3-1)(n-1)^2}\right]
\end{aligned}
$$

Therefore

$$\begin{aligned}\mathrm{Var}(B_n) &= \tau_1^2\left[\frac{4}{n(n-1)}\right]+\tau_2^2\left[\frac{4}{n^2(n-1)}+\frac{4n_2n_3}{n^2(n-1)^2}+\right.\\ &\left.+\frac{2n_2(2+n_3)^2}{n^2(n_2-1)(n-1)^2}+\frac{2n_3(2+n_2)^2}{n^2(n_3-1)(n-1)^2}\right] \qquad \text{(S.2)}\end{aligned}$$

Note that $n_3 = n-1-n_2$, then we can rewrite $\mathrm{Var}(B_n)$ as

$$\mathrm{Var}(B_n) = \eta_1(n)\tau_1^2+\eta_2(n;n_2)\tau_2^2 \qquad \text{(S.3)}$$

S1.2 Asymptotic properties B_n's variance

We show that

$$\begin{aligned}\mathrm{Var}\left(\frac{n}{2}B_n\right) &= \tau_1^2\frac{n}{n-1}+\tau_2^2\left[\frac{1}{n-1}+\frac{n_2n_3}{(n-1)^2}+\frac{n_2(2+n_3)^2}{2(n_2-1)(n-1)^2}+\right.\\ &\left.+\frac{n_3(2+n_2)^2}{2(n_3-1)(n-1)^2}\right].\end{aligned}$$

Note that

$$\begin{aligned}\mathrm{Var}\left(\frac{n}{2}B_n\right) &= \tau_1^2O(1)+\tau_2^2\left[O(n^{-1})+O(1)+O(1)+O(1)\right]\\ &= O(1).\end{aligned}$$

Let $\tau_n = \frac{n}{2}\sqrt{\mathrm{Var}(B_n)}$.

A simple consequence is that $\mathrm{Var}(\frac{n}{2}B_n) = \frac{n^2}{4}\mathrm{Var}(B_n) = O(1)$. Thus, it follows that

$$\tau_n = \frac{n}{2}\sqrt{\mathrm{Var}(B_n)} = O(1).$$

S1.3 Asymptotic properties of B_n

At the Section 2.2 of the main manuscript we presented two Theorems about the asymptotic properties of B_n, in this present section the proof of those Theorems are conveyed. The following Lemma is an important result required to demonstrate the asymptotic convergence of the test statistic.

Lemma S1.1 *Let* $\frac{X}{\delta_n} \xrightarrow{D} \mathrm{N}(0,1)$, $\delta_n = \mathrm{O}(1)$ *and* $\delta_n^* = \mathrm{O}(1)$. *Then,* $\frac{X}{\delta_n^*} \xrightarrow{D} \mathrm{N}(0,M)$ *where* $M = \lim\left(\frac{\delta_n^2}{\delta_n^{*2}}\right)$.

Proof: Note that

$$\frac{X}{\delta_n^*}\frac{\delta_n}{\delta_n} = \frac{\delta_n}{\delta_n^*}\frac{X}{\delta_n} \xrightarrow{D} \mathrm{N}(0,\gamma),$$

where

$$\gamma = \text{Var}\left(\frac{\delta_n}{\delta_n^*}\frac{X}{\delta_n}\right) \to \lim_{n\to\infty}\left(\frac{\delta_n}{\delta_n^*}\right)^2 = M.$$

Theorem 1 Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. $L \times 1$ random vectors. Let $\phi(\cdot,\cdot)$ be a kernel of degree 2 satisfying $\mathbb{E}[\phi(\mathbf{X}_1, \mathbf{X}_2)^2] < \infty$ and $\text{var}[\mathbb{E}(\phi(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{X}_1)] = \sigma_1^2 > 0$. Consider definition (S.1) for B_n when $n_1 = 1$ and let $V_n = \text{Var}(B_n)$, $\tau_n = (n/2)V_n^{1/2}$ and $W = J_1 + J_2 - J_3 - J_4$, where $\frac{\psi_1(X_1)}{\tau_n} \xrightarrow{D} J_1$, and J_2, J_3 and J_4 are random variables with normal distribution. Then

$$\frac{(n/2)B_n}{\tau_n} \xrightarrow{D} W \text{ as n } \to \infty. \tag{S.4}$$

Proof:

We are interested in the distribution of B_n with fixed L and $n \to \infty$. We just saw that $\tau_n = O(1)$.

From the Hoeffding decomposition of B_n we have:

$$\begin{aligned}
\frac{n}{2}B_n &= \psi_1(X_1) - \frac{1}{n-1}\sum_{i=1}^{n_2}\psi_1(X_{2i}) - \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_1(X_{3j}) + \\
&+\frac{1}{n-1}\sum_{i=1}^{n_2}\psi_2(X_1, X_{2i}) + \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_2(X_1, X_{3j}) + \\
&+\frac{1}{n-1}\sum_{i=1}^{n_2}\sum_{j=1}^{n_3}\psi_2(X_{2i}, X_{3j}) \\
&-\frac{(2+n_3)}{(n-1)(n_2-1)}\sum_{1\le i<j\le n_2}\psi_2(X_{2i}, X_{2j}) \\
&-\frac{(2+n_2)}{(n-1)(n_3-1)}\sum_{1\le i<j\le n_3}\psi_2(X_{3i}, X_{3j})
\end{aligned} \tag{S.5}$$

Observe now that $\frac{n}{2}B_n = W_1 + W_2 - W_3 - W_4$ and the asymptotic distribution of B_n can be obtained from the distribution of each of these variables, where

$$
\begin{aligned}
W_1 &= \psi_1(X_1) - \frac{1}{n-1}\sum_{i=1}^{n_2}\psi_1(X_{2i}) - \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_!(X_{3j}) + \\
&\quad + \frac{1}{n-1}\sum_{i=1}^{n_2}\psi_2(X_1, X_{2i}) + \frac{1}{n-1}\sum_{j=1}^{n_3}\psi_2(X_1, X_{3j}) \qquad \text{(S.6)}\\
W_2 &= \frac{1}{n-1}\sum_{i=1}^{n_2}\sum_{j=1}^{n_3}\psi_2(X_{2i}, X_{3j}) \qquad \text{(S.7)}\\
W_3 &= \frac{2+n_3}{(n-1)(n_2-1)}\sum_{1\le i<j\le n_2}\psi_2(X_{2i}, X_{2j}) \qquad \text{(S.8)}\\
W_4 &= \frac{2+n2}{(n-1)(n_3-1)}\sum_{1\le i<j\le n_3}\psi_2(X_{3i}, X_{3j}) \qquad \text{(S.9)}
\end{aligned}
$$

Under the null hypothesis X_1, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are identically distributed, thus W_1 can be expressed as

$$
W_1 = \psi_1(X_1) - \frac{1}{n-1}\sum_{i=2}^{n}\psi_1(X_i) + \frac{1}{n-1}\sum_{j=2}^{n}\psi_2(X_1, X_j). \qquad \text{(S.10)}
$$

By the Law of Large Numbers (LLN) follows that

$$
\frac{1}{n-1}\sum_{i=2}^{n}\psi_1(X_i) \xrightarrow{P} \mathrm{E}[\psi_1(X_1)] = 0 \qquad \text{(S.11)}
$$

$$
\frac{1}{n-1}\sum_{j=2}^{n}\psi_2(X_1, X_j) \xrightarrow{P} \mathrm{E}[\psi_2(X_1, X_2)] = 0. \qquad \text{(S.12)}
$$

Thereby,

$$
W_1 \xrightarrow{P} \psi_1(X_1).
$$

As $\frac{\psi_1(X_1)}{\tau_n} \xrightarrow{D} J_1$ and $W_1 \xrightarrow{P} \psi_1(X_1)$, then, by Slutsky's theorem, $\frac{W_1}{\tau_n} \xrightarrow{D} J_1$. From the Central Limit Theorem (TCL) we have

$$
\frac{W_2 - E(W_2)}{\sqrt{\mathrm{Var}(W_2)}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n_2}\sum_{j=1}^{n_3}\psi_2(X_{2i}, X_{3j})}{\sqrt{\frac{n_2 n_3}{(n-1)^2}\tau_2^2}} \to N(0,1). \qquad \text{(S.13)}
$$

Observe that $\sqrt{\frac{n_2 n_3}{(n-1)^2}\tau_2^2} = O(1)$ and $\tau_n = O(1)$. Then by Lemma S1.1 follows that

$$\frac{W_2}{\tau_n} \xrightarrow{D} J_2 \sim \mathrm{N}(0, M_2), \text{ where } M_2 = \lim \left(\frac{\frac{n_2 n_3}{(n-1)^2}\tau_2^2}{\tau_n^2} \right). \tag{S.14}$$

Similarly,

$$\frac{W_3 - \mathrm{E}(W_3)}{\sqrt{\mathrm{Var}(W_3)}} = \frac{\frac{(2+n_3)}{(n-1)(n_2-1)} \sum_{1 \le i < j \le n_2} \psi_2(X_{2i}, X_{2j})}{\sqrt{\frac{(2+n_3)^2 n_2 \tau_2^2}{2(n-1)^2(n^2-1)}}} \to N(0,1). \tag{S.15}$$

Other properties are that $\sqrt{\frac{(2+n_3)^2 n_2 \tau_2^2}{2(n-1)^2(n^2-1)}} = O(1)$ and $\tau_n = O(1)$, then by the Lemma S1.1

$$\frac{W_3}{\tau_n} \xrightarrow{D} J_3 \sim \mathrm{N}(0, M_3), \text{ where } M_3 = \lim \left(\frac{\frac{(2+n_3)^2 n_2 \tau_2^2}{2(n-1)^2(n^2-1)}}{\tau_n^2} \right). \tag{S.16}$$

Analogously,

$$\frac{W_4 - \mathrm{E}(W_4)}{\sqrt{\mathrm{Var}}(W_4)} = \frac{\frac{(2+n_2)}{(n-1)(n_3-1)} \sum_{1 \le i < j \le n_3} \psi_2(X_{3i}, X_{3j})}{\sqrt{\frac{(2+n_2)^2 n_3 \tau_2^2}{2(n-1)^2(n_3-1)}}} \to N(0,1).. \tag{S.17}$$

Once more, $\sqrt{\frac{(2+n_2)^2 n_3 \tau_2^2}{2(n-1)^2(n_3-1)}} = O(1)$ and $\tau_n = O(1)$, then

$$\frac{W_4}{\tau_n} \xrightarrow{D} J_4 \sim \mathrm{N}(0, M_4), \text{ where } M_4 = \lim \left(\frac{\frac{(2+n_2)^2 n_3 \tau_2^2}{2(n-1)^2(n_3-1)}}{\tau_n^2} \right). \tag{S.18}$$

Thus, applying Slutsky's theorem we have

$$\begin{aligned} \frac{(n/2)B_n}{\tau_n} &= \frac{(n/2)B_n}{(n/2)V_n^{1/2}} = \frac{B_n}{\sqrt{\mathrm{Var}(B_n)}} \\ &= \frac{W_1 + W_2 - W_3 - W_4}{\tau_n} \xrightarrow{D} J_1 + J_2 - J_3 - J_4 \end{aligned} \tag{S.19}$$

The asymptotic distribution of B_n for a fixed L depends on the choice of the kernel $\phi(,)$. However, it is important to note that B_n converges in distribution to a random variable (not necessarily normal) when n increases. The next theorem addresses the asymptotic distribution of B_n in L for a fixed n.

Theorem 2 Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. $L \times 1$ random vectors. Let $\phi(\cdot,\cdot)$ be a kernel of degree 2 such that

$$\phi(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{L}\sum_{l=1}^{L} \phi^*(X_{il}, X_{jl}) \tag{S.20}$$

for some kernel $\phi^*(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$, where X_{il} is the l-th entry of $\mathbf{X}_i$. Define $\phi_1^*(x_{il}) = \mathbb{E}[\phi^*(X_{il}, X_{jl})|X_{il} = x_{il}]$ and suppose $\text{var}(\phi_1^*(X_{il})) > 0$ and $\text{var}(\phi^*(X_{il}, X_{jl})) < \infty$. Let B_n be defined by (S.1) for the case where $n_1 = 1$, and assume that all conditions in Theorem 1 hold. Suppose also that

$$\sum_{1\le l<m\le n} \mathbb{E}[\phi^*(X_{il}, X_{jl})\phi^*(X_{im}, X_{jm})] = O(L) \tag{S.21}$$

and

$$\sum_{1\le l<m\le n} \mathbb{E}[\phi_1^*(X_{il})\phi_1^*(X_{jm})] = O(L). \tag{S.22}$$

Then

$$\frac{B_n}{\sqrt{\text{Var}(B_n)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad L \to \infty. \tag{S.23}$$

Proof: We start writing $\psi_1(X_i)$ and $\psi_2(X_i, X_j)$ as a function of $\phi_1^*(\cdot)$ and $\phi_2^*(\cdot,\cdot)$. Note that

$$\psi_1(\mathbf{X}_i) = \frac{1}{L}\sum_{l=1}^{L} \psi_1^*(X_{il}) \tag{S.24}$$

$$\begin{aligned}\psi_2(\mathbf{X}_i, \mathbf{X}_j) &= \frac{1}{L}\sum_{l=1}^{L} \phi^*(X_{il},, X_{jl}) - \frac{1}{L}\sum_{l=1}^{L} \psi_1^*(X_{il}) \\ &\quad -\frac{1}{L}\sum_{l=1}^{L} \psi_1^*(X_{jl}) - \theta\end{aligned} \tag{S.25}$$

where

$$\psi_1^*(X_{il}) = \phi_1^*(X_{il}) - \theta \tag{S.26}$$

$$\phi_1^*(x_{il}) = \mathbb{E}[\phi^*(X_{il}, X_{jl}) \mid X_{il} = x_{il}] \tag{S.27}$$

$$\phi_2^*(x_{il}, x_{jl}) = \mathbb{E}[\phi^*(X_{il}, X_{jl}) \mid X_{il} = x_{il}, X_{jl} = x_{jl}]. \tag{S.28}$$

We can write $\psi_1(\cdot)$ as

$$\psi_1(X_i) = \frac{1}{L}\sum_{l=1}^{L} [\phi_1^*(X_{ij}) - \theta], \tag{S.29}$$

or

$$\psi_1(X_i) \quad = \quad \frac{1}{L}\sum_{l=1}^{L}\psi_1^*(X_{ij}). \tag{S.30}$$

Thus the variance of $\psi_1(\cdot)$ is given by

$$\mathrm{Var}(\psi_1(X_i)) \quad = \quad \mathrm{Var}\left[\frac{1}{L}\sum_{l=1}^{L}\psi_1^*(X_{il})\right]. \tag{S.31}$$

By (S.21) we have that

$$\begin{aligned}\mathrm{Var}\left(\psi_1\left(\mathbf{X}_i\right)\right) &= \frac{1}{L^2}\left\{\sum_{l=1}^{L}\mathrm{Var}\left[\psi_1^*\left(X_{il}\right)\right]\right. \\ &\quad \left. +2\sum_{1\le l<m\le L}\mathrm{Cov}\left(\psi_1^*\left(X_{il}\right),\psi_1^*\left(X_{im}\right)\right)\right\} \\ &= O\left(L^{-1}\right)\end{aligned} \tag{S.32}$$

and by (S.22) the variance of $\psi_2(,)$ is

$$\begin{aligned}\mathrm{Var}\left(\psi_2\left(\mathbf{X}_i,\mathbf{X}_j\right)\right) =& \frac{1}{L^2}\left\{\sum_{l=1}^{L}\mathrm{Var}\left(\phi^*\left(X_{il},X_{jl}\right)\right)+\right. \\ &+2\sum_{1\le l<m\le L}\mathrm{Cov}\left(\phi^*\left(X_{il},X_{jl}\right),\phi^*\left(X_{im},X_{jm}\right)\right) \\ &\left.+2\,\mathrm{Var}\left(\frac{1}{L}\sum_{l=1}^{L}\psi_1^*\left(X_{il}\right)\right)\right\} \\ =& O\left(L^{-1}\right).\end{aligned} \tag{S.33}$$

Thus, for fixed n and for $L\to\infty$ it follows that

$$\frac{B_n}{\sqrt{\mathrm{Var}(B_n)}} = V_n^{-1/2}B_n \xrightarrow{D} N(0,1) \tag{S.34}$$

S2 Total of combinations

In order to develop the homogeneity test described in Section 3 of the main manuscript we require the number of different group configurations that can be formed by separating n elements, $x_1, x_2, \ldots, x_n$ into three groups, G_1, G_2 e

G_3. Follows from [Valk and Pinheiro, 2012] that the number of combination of n elements into two groups is

$$p(n) \quad = \quad 2^{n-1} - n - 1. \tag{S.35}$$

Then if we divide n elements into three groups where one of them has size 1, it follows that the number of combinations is

$$\delta_3(n) \quad = \quad (2^{n-2} - n)n. \tag{S.36}$$

Now we focus on the case where all groups have more than one element. We can fix, without loss of generality, x_1 as an element that belongs to the first group, G_1. Thus, we still have $n-1$ elements to be distributed among the three groups. Since we cannot have a unitary group, we need at least one more point for the first group. This group can have up to $n-4$ observations, since the remaining sets must necessarily have two elements each. Thus, we then have the following number of possible first sets

$$\binom{n-1}{1} + \binom{n-1}{2} + \cdots + \binom{n-1}{n-5}.$$

For the remaining elements that need to be divided into two clusters, just divide them into two groups with at least 2 elements in each using the function $p(\cdot)$. Combining these results, we have a number of different configurations of non-unitary groups when we separate n elements into 3 groups given by

$$\begin{aligned} S_3(n) \quad &= \quad \binom{n-1}{1} p(n-2) + \binom{n-1}{2} p(n-3) + \cdots + \binom{n-1}{n-5} p(4) \\ &= \quad \sum_{k=1}^{n-5} \binom{n-1}{k} p(n-k-1). \end{aligned} \tag{S.37}$$

We can still rewrite this equation on a recurring basis. Note that if we already know how many configurations of groups we have with n non-unitary elements, and how many configurations with a unitary group, then it is possible to calculate $S_3(n+1)$ as

$$S_3(n+1) \quad = \quad 3S_3(n) + \delta_3(n). \tag{S.38}$$

With such equations we can rewrite $S_3(n)$ as

$$S_3(n) \quad = \quad \frac{233(3^{n-6}) + 1 + n + n^2 - (2+n)2^{n-1}}{2}. \tag{S.39}$$

Thus, the number of different group configurations where at most one of them has size one is given by

$$\begin{aligned}\gamma_3(n) &= \frac{233(3^{n-6})+1+n+n^2-(2+n)2^{n-1}}{2}+\delta_3(n)\\ &= \frac{233(3^{n-6})+1+n-n^2-2^n}{2}. \end{aligned} \tag{S.40}$$

S3 The clustering method *uclust3*

The algorithm for the clustering method *uclust*3, introduced in the Section 3.1 of the main manuscript, can be described as follows. We apply the homogeneity test on the dataset and if it returns "non homogeneous", we then find the partition $\{G_1^\star, G_2^\star, G_3^\star\}$ that maximizes B_n and set $n_1^\star$ as the smallest subgroup size. Among all possible configurations in which one of the groups has size one, we find the configuration $\{G_1^{\star 1}, G_2^{\star 1}, G_3^{\star 1}\}$ that maximizes B_n and set this B_n value as B_n^1. If $\{G_1^\star, G_2^\star, G_3^\star\}$ is a significant configuration and $B_n > B_n^1$, we have found our optimal partition. If $\{G_1^\star, G_2^\star, G_3^\star\}$ is a significant partition and $B_n < B_n^1$ with B_n^1 significant, then $\{G_1^{\star 1}, G_2^{\star 1}, G_3^{\star 1}\}$ is our optimal partition.

However, if this maximal B_n comes from a non significant partition $\{G_1^\star, G_2^\star, G_3^\star\}$, then there are no other significant partitions in configurations with smaller group size between 2 and $n_1^\star$. The restricted search is done on subgroups with sizes larger than $n_1\star$, until it finds the significant partition and compares with B_n^1, returning the configuration with maximum significant B_n. By exploring this insight, we built the following clustering algorithm based on restricted optimization problems.

*uclust*3 **Algorithm**: Finds the data partition that maximizes B_n in the universe of all significant partitions

Input: Data $\mathbf{X}$
Output: Partition $\{G_1^\star, G_2^\star, G_3^\star\}$
01: Apply homogeneity test to $\mathbf{X}$
02: **if** Accept H_0
03: **Return** $G_1^\star = \emptyset$, $G_2^\star = \emptyset$ and $G_3^\star = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$
04: **else**
05: find $G_1^\star$, $G_2^\star$ and $G_3^\star$ that optimize B_n. Set this results as B_n
06: For $G_1^{\star 1}$ of size one, find $G_1^{\star 1}$, $G_2^{\star 1}$ and $G_3^{\star 1}$ that optimize B_n. Set this results as B_n^1
07: **If** B_n is significant
08: **If** $B_n < B_n^1$ and B_n^1 is significant, $G_1^\star = G_1^{\star 1}$, $G_2^\star = G_2^{\star 1}$ and $G_3^\star = G_3^{\star 1}$
9: **else**
10: Set $G_1^\star$ size ($n_1^\star$) as the smallest size among $G_1^\star$, $G_2^\star$ and $G_3^\star$
11: **while** $\{G_1^\star, G_2^\star, G_3^\star\}$ is not significant partitions
12: **while** $\{G_1^\star, G_2^\star, G_3^\star\}$ is not significant partitions.
13: $n_2 \in \{(n_1^\star + 1), \ldots, (n - 2n_1^\star + 1)\}$, find G_1, G_2 and G_3 that optimize B_n for subgroup size and set $G_1^\star = G_1$, $G_2^\star = G_2$ and $G_3^\star = G_3$
14: $n_1^\star = n_1^\star + 1$
15: **Compare** B_n and B_n^1 and do **08**
16: **Return** $\{G_1^\star, G_2^\star, G_3^\star\}$

The multiple optimization subproblems in the *uclust*3 algorithm are solved through a cyclic coordinate ascent algorithm repeated multiple times with random starting clusters to account for local optima.

S4 Simulations Studies

In this section we present simulation studies in order to evaluate some aspects of our proposed methodology, a complementary material for the simulation studies shown at Section 4 of the main manuscript. At first we evaluate the size and power of the proposed *utest* for homogeneity of three groups.

S4.1 Simulations for the *utest*

We present here a simulation study to evaluate the performance of the *utest* for three groups. The data was simulate as shown at Section 4.1 of the main manuscript, but this time without group of size one. The groups G_1 and G_2 were set with the same size $n_1 = n_2 = \lfloor n/3 \rfloor$.

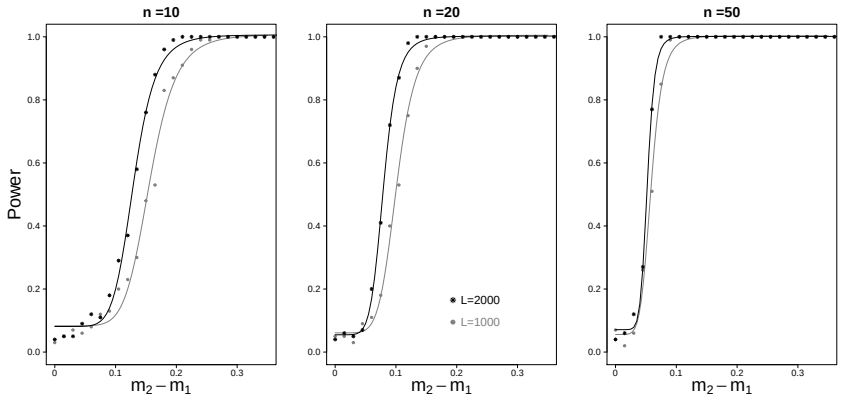

Figure S1: Power curves of *utest* for two dimension $L = 1000$ (gray) and $L = 2000$ (black) for 100 replications of each scenario of n with $n_1 = n_2 = \frac{\lfloor n \rfloor}{3}$ and $n_3 = n - n_1 - n_2$.

S4.2 Simulations for homogeneity test *uclust*3

Similarly to Section 4.2 of the main manuscript we used simulation studies to evaluate the homogeneity test.

S4.2.1 Size of homogeneity test *uclust*3

First the data were simulated following the same distribution. All elements from the $n \in \{10, 20, 30, 40, 50, 100\}$ vectors with dimension $L \in \{1000, 2000\}$ were generated following a Normal distribution with mean 0 and variance 1. The homogeneity test was applied to the dataset and observed if the null hypothesis was rejected or not. This process was replicated 100 times and the size of the test can be seen at the following table.

Table S1: Size of homogeneity test *uclust*3

n	Dimension L	
	1000	2000
10	0.01	0.01
20	0	0
30	0.01	0
40	0.02	0
50	0.03	0.03
100	0.14	0.03

S4.2.2 Power of homogeneity test *uclust*3

In order to evaluate the power of our proposed homogeneity test *uclust*3 we simulate data from independent normally distributed vectors divided in three groups G_1, G_2 and G_3. The L dimensional vectors in G_1 are generated from a independent and identically normal with mean $m_1 = 0$ and variance 1. The elements of the vectors in G_2 and G_3 have the same properties with mean m_2 and m_3, respectively. For each sample size n in $\{10, 20, 50\}$, the G_1 and G_2 group sizes n_1 and n_2 were chosen so that we had a central configuration, in which the groups have approximately the same number of elements and a extremely configuration in which one of the groups has only two elements and the other has $n/2$ elements. Naturally the third group size's is defined as $n_3 = n - n_1 - n_2$.

Table S2: Power of homogeneity test *uclust*3

n	$(m_2,\ m_3)$	$(n_1,\ n_2)$	Dimension L	
			1000	2000
10	(0.25, 0.5)	(2, 5)	0.21	0.31
		(3, 3)	0.06	0.09
	(0.5, 1)	(2, 5)	0.21	0.24
		(3, 3)	0.02	0.02
20	(0.25, 0.5)	(2, 10)	1	1
		(6, 6)	1	1
	(0.5, 1)	(2, 10)	1	1
		(6, 6)	1	1
50	(0.25, 0.5)	(2, 25)	1	1
		(16, 16)	1	1
	(0.5, 1)	(2, 25)	1	1
		(16, 16)	1	1

S4.3 Simulations for finding correct clusters comparing with the *kmeans*

We complement the simulations study in Section 4.3 by performing a comparison between *uclust*3 method and *kmeans* clustering algorithm for the case where we have a size one group.

Table S3: Comparison of mean ARI and standard deviation (Sd) of the accuracy in clustering of *kmeans* and *uclust*3 methods with a size one group.

n	$(m_2,\ m_3)$	(n_2)	Method	Dimension L			
				1000		2000	
				Mean	Sd	Mean	Sd
10	(0.25, 0.5)	2	*kmeans*	0.44	0.03	0.48	0.05
			*uclust*3	0.47	0.03	0.5	0.06
		5	*kmeans*	0.66	0.02	0.73	0.03
			*uclust*3	0.74	0.03	0.79	0.03
	(0.5, 1)	2	*kmeans*	0.86	0.07	0.94	0.04
			*uclust*3	0.75	0.1	0.82	0.08
		5	*kmeans*	0.93	0.03	0.97	0.01
			*uclust*3	0.99	0	1	0
20	(0.25, 0.5)	2	*kmeans*	0.34	0.02	0.33	0.02
			*uclust*3	0.33	0.01	0.36	0.02
		10	*kmeans*	0.73	0	0.77	0.01
			*uclust*3	0.73	0.01	0.74	0.01
	(0.5, 1)	2	*kmeans*	0.62	0.12	0.83	0.09
			*uclust*3	0.67	0.12	0.98	0.01
		10	*kmeans*	0.95	0.01	0.97	0.01
			*uclust*3	0.92	0.02	1	0
50	(0.25, 0.5)	2	*kmeans*	0.17	0.01	0.17	0.01
			*uclust*3	0.15	0	0.15	0
		25	*kmeans*	0.75	0	0.76	0
			*uclust*3	0.74	0	0.74	0
	(0.5, 1)	2	*kmeans*	0.22	0.02	0.35	0.1
			*uclust*3	0.18	0.02	0.41	0.15
		25	*kmeans*	0.9	0.02	0.94	0.01
			*uclust*3	0.82	0.01	0.99	0

Over the 100 replications observing the different scenarios we can conclude that both methods compete, alternating in the presentation of the best results.

S5 Finding correct clusters comparing *uclust*3 and *uhclust* in a presence of an outlier

We complement the simulation study presented in Section 4.4 of the manuscript considering here only groups larger than 2. Figures S2 and S3 report curves of proportion times that the algorithms found significant separation and correct groups considering different values of $m_2 - m_1$ varying on the x axis, sample size n taking values in $\{10, 20, 50\}$ and dimension $L = 1000$ and $L = 2000$. The

results are based on 50 repetitions.

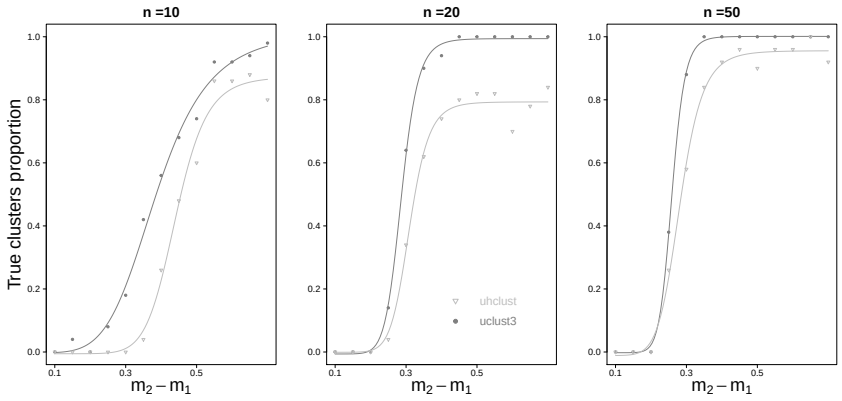

Figure S2: True cluster proportion curves of *uclust*3 and *uhclust* for dimension $L = 1000$ with 50 replications of each scenario of n with $\alpha = 0.05$.

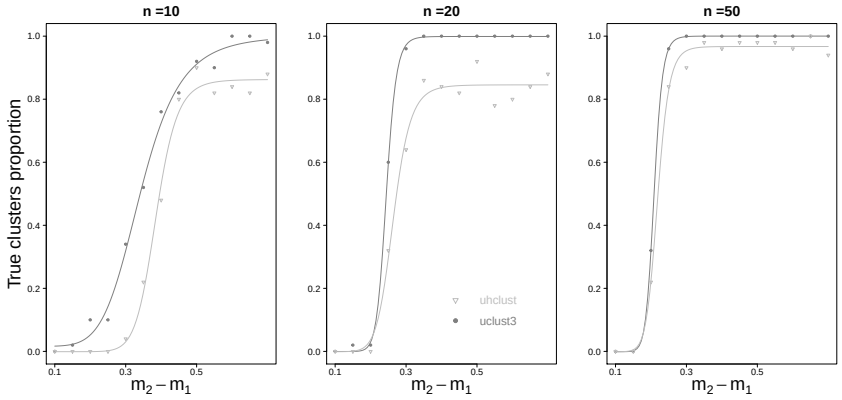

Figure S3: True cluster proportion curves of *uclust*3 and *uhclust* for dimension $L = 2000$ with 50 replications of each scenario of n with $\alpha = 0.05$.

S6 Application

In the interest of evaluating the performance of the proposed method *uclust*3 comparing with *uhclust* and *sigclust* we consider an image group configuration with an outlier. The data are the same as described in Section 5 in the main manuscript. We randomly select 1 image from Tony Blair and 10 images from each other public figure in the above cited dataset and run *uhclust*, *sigclust*

and *uclust*3. Figure S4 presents the dendrogram with *uhclust* groups. Note that *uhclust* finds two significant clusters, with an ARI of 0.8135593. Figure S5 presents the dendrogram with corresponding *sigclust* p-values for the labelled faces dataset. Note that *sigclust* also finds two significant clusters, with an ARI of 0.8135593. None of the methods were able to identify the outlier. However, when applying the *uclust*3 method we find the correct groups with ARI of 1.

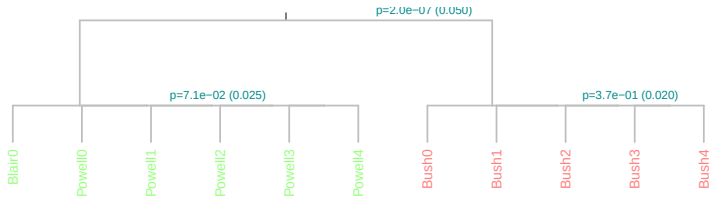

Figure S4: Annotated dendrogram of significance analysis for hierarchical clustering *uhclust* for 11 pictures of 3 public figures. P-values and corrected significance levels α^* are shown for each test performed at the corresponding node.

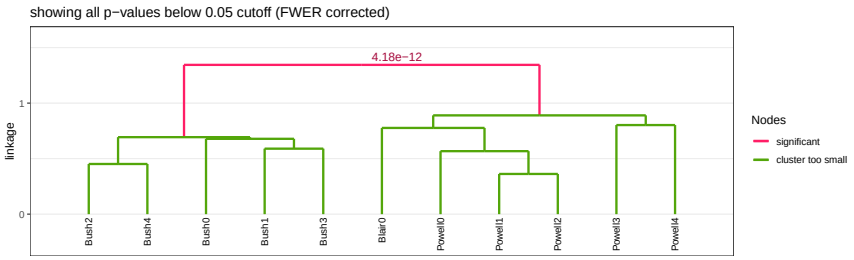

Figure S5: Annotated dendrogram of significance analysis for hierarchical clustering *sigclust* for 11 pictures of 3 public figures. P-values and corrected significance levels α^* are shown for each test performed at the corresponding node.

References

[Hoeffding, 1948] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.

[Valk and Cybis, 2020] Valk, M. and Cybis, G. B. (2020). U-statistical inference for hierarchical clustering. *Journal of Computational and Graphical Statistics*.

[Valk and Pinheiro, 2012] Valk, M. and Pinheiro, A. (2012). Time-series clustering via quasi u-statistics. *Journal of Time Series Analysis*.