

Implementação Computacional de Algoritmos para Agrupamento de Dados HDLSS e HDLLSS

Rafael Lins

Orientador: Prof. Dr. George von Borries

Departamento de Estatística, *Universidade de Brasília*

Julho, 2019



Sumário

1. Introdução
2. Objetivos
3. Metodologia
4. Resultados do Pacote RCLUST
5. Conclusões



Introdução

Introdução

- O avanço tecnológico na coleta de dados \Rightarrow obtenção de informações para um grande número grande de variáveis ao mesmo tempo.



Introdução

- O avanço tecnológico na coleta de dados \Rightarrow obtenção de informações para um grande número grande de variáveis ao mesmo tempo.
- Restrições de tempo e custo \Rightarrow poucas amostras por variável.



Introdução

- O avanço tecnológico na coleta de dados \Rightarrow obtenção de informações para um grande número grande de variáveis ao mesmo tempo.
- Restrições de tempo e custo \Rightarrow poucas amostras por variável.
- Problemas multivariados: n é fixo e $n/d \rightarrow 0$ sendo a dimensão dos dados $d \rightarrow \infty$, podendo ser observado no tempo.



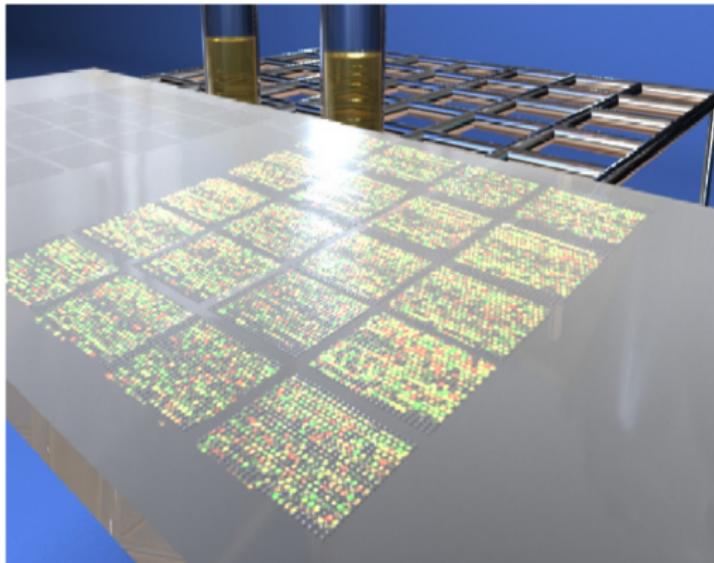
Introdução

- dados **HDLSS** (do inglês, *high dimensional low sample size*);
- dados **HDLLSS** (*high dimensional longitudinal low sample size*).



Introdução

- Exemplos: **análise de microarranjo**, imagens médicas, sinais biopotenciais e assim por diante.



Microarranjo com expressão de milhares de genes.
Por UCN em Chapell Hill e A. Malcolm Campbell ©

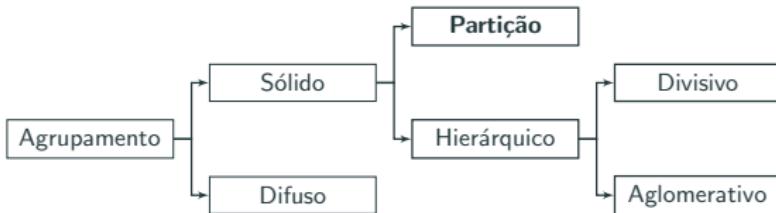


- Algoritmos de agrupamento exercem um papel importante capturando os padrões subjacentes dos dados.
- Os principais desafios para os algoritmos de agrupamento tradicionais:
 - Quantos grupos (ou *clusters*) existem?
 - Tempo de processamento;
 - Qualidade de agrupamento;
 - Sensibilidade a valores discrepantes;
 - Diferentes resultados para transformações monótonas distintas.



von Borries, 2008: PPCLUST

- agrupamento de dados HDLSS;
- algoritmo de partição.



- teste não-paramétrico de ausência de efeito de grupo desenvolvido (Wang & Akritas, 2004);
- p -valor como medida de similaridade.



Estrutura de dados HDLSS

Variável	Observações				Tamanho Amostral
1	X_{11}	X_{12}	...	X_{1n_1}	n_1
2	X_{21}	X_{22}	...	X_{1n_2}	n_2
:	:	:	:	:	:
a	X_{a1}	X_{a2}	...	X_{an_a}	n_a



Teste Presente no PPCLUST

(Teste de ausência de efeito de grupo - Wang & Akritas, 2004.)

Seja $H_0 : F_1(x) = \dots = F_a(x)$ seja satisfeito, com $F_i(x)$ arbitrária. Se $n_i \geq 2$ fixo, assumindo as observações independentes, os seguintes limites existem

$$v_2^2 = \lim_{a \rightarrow \infty} \frac{1}{a} \sum_{i=1}^a \frac{1}{n_i} \sigma_i^2 > 0$$

$$\tau_2 = \lim_{a \rightarrow \infty} \frac{1}{a} \sum_{i=1}^a \frac{2\sigma_i^4}{n_i(n_i - 1)}.$$

Então, como $a \rightarrow \infty$

$$\sqrt{a}(F_R - 1) \xrightarrow{d} N(0, \tau_2/v_2^4)$$



von Borries, 2008: PPCLUSTEL

- agrupamento de dados HDLLSS;
- algoritmo de partição.
- teste de ausência de efeito simples (von Borries, 2008);
- p -valor como medida de similaridade.



Estrutura de dados HDLLSS

Variável	Sujeito	Tempo			
		t_1	t_2	...	t_b
1	1	X_{111}	X_{121}	...	X_{1b1}
	2	X_{112}	X_{122}	...	X_{1b2}
		⋮	⋮	⋮	⋮
	n_1	X_{11n_1}	X_{12n_1}	...	X_{1bn_1}
2	1	X_{211}	X_{221}	...	X_{2b1}
	2	X_{212}	X_{222}	...	X_{2b2}
		⋮	⋮	⋮	⋮
	n_2	X_{21n_2}	X_{22n_2}	...	X_{2bn_2}
		⋮	⋮	⋮	⋮
a	1	X_{a11}	X_{a21}	...	X_{ab1}
	2	X_{a12}	X_{a22}	...	X_{ab2}
		⋮	⋮	⋮	⋮
	n_a	X_{a1n_a}	X_{a2n_a}	...	X_{abn_a}



(Teste de ausência de efeito simples - von Borries, 2008.)

Seja $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, onde μ é a média global; $\alpha_i (i = 1, \dots, a)$ o efeito médio do fator; $\beta_j (j = 1, \dots, b)$ o efeito do ponto no tempo; γ_{ij} o efeito da interação fator-tempo; e ϵ_{ijk} um erro aleatório com distribuição arbitrária F_{ij} , para todo $k = 1, \dots, n_i$. Considere, $H_0(\varphi) : \varphi_{ij} = \alpha_i + \gamma_{ij} = 0$ verdadeira. Se as observações Y_{ijk} têm momento central finito $(2 + \delta)(\delta > 0)$, e o número de repetições é pequeno, com $n_i \geq 2$ e limitado, observado para um número fixo b de pontos no tempo e com $a \rightarrow \infty$,

$$F_\varphi = \sqrt{ab}(MS\varphi - MSE) \xrightarrow{d} N \left(0, \lim_{a \rightarrow \infty} \frac{2}{ab} \sum_{i=1}^a \frac{1}{n_i(n_i - 1)} \sum_{j=1}^b \sum_{j'=1}^a \sigma_{ijj'}^2 \right)$$



Características dos Algoritmos

Vantagens:



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;



Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;



Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;



Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;



Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.
- Implementados no SAS.



Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.
- Implementados no SAS.



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.
- Implementados no SAS.

Desvantagens:



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.
- Implementados no SAS.

Desvantagens:

- Implementados no SAS!



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.
- Implementados no SAS.

Desvantagens:

- Implementados no SAS!



Características dos Algoritmos

Vantagens:

- Invariantes a transformações monótonas;
- Especificação automática do número de grupos;
- Qualidade do agrupamento;
- Sem problema de comparação múltipla;
- Não precisam de redução de dimensão;
- Não precisam de dados balanceados;
- Rápidos e fáceis de utilizar;
- Sem problema de alocação de memória.
- Implementados no SAS.

Desvantagens:

- Implementados no SAS!



Objetivos

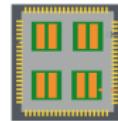
Objetivos

Criação do pacote **RClust** contendo:

- Implementação do PPCLUST e PPCLUSTEL em pacote aberto (gratuito);
- Implementação de um novo algoritmo alternativo adepto de **processamento paralelo**.

PPCLUST → PPCLUST-H

PPCLUSTEL → PPCLUSTEL-H



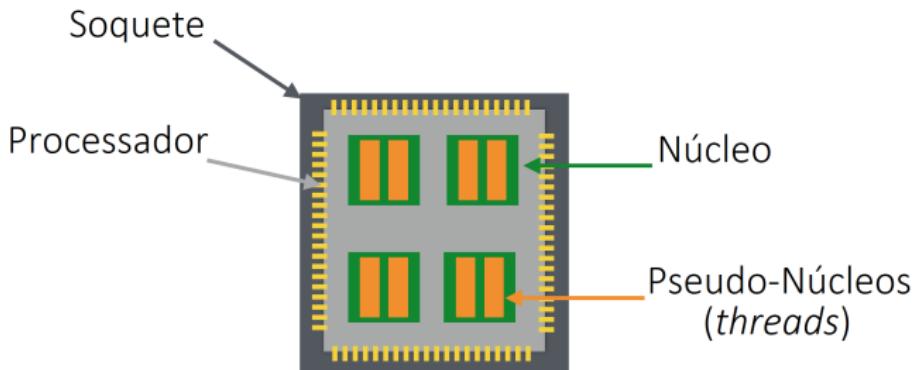
- Comparação entre os procedimentos PPCLUST-H, PPCLUSTEL-H e MCLUST (dados simulados).
- Aplicação em dados de expressão gênica de adenoma e câncer de colón e de leveduras (fungos).



Metodología

Processamento Paralelo no R

- Esquema de um processador *multicore*.



Fonte: github.com/ljdursi

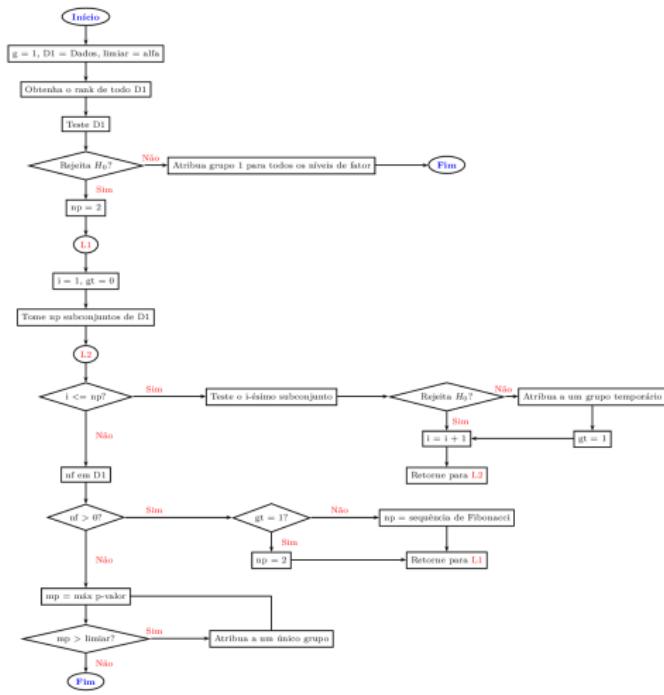


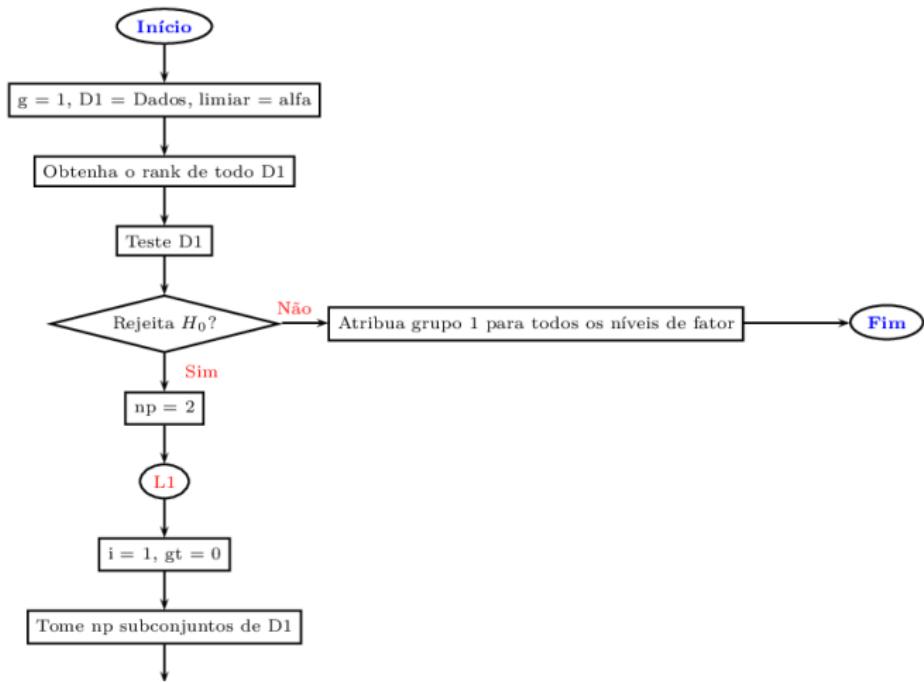
Processamento Paralelo no R

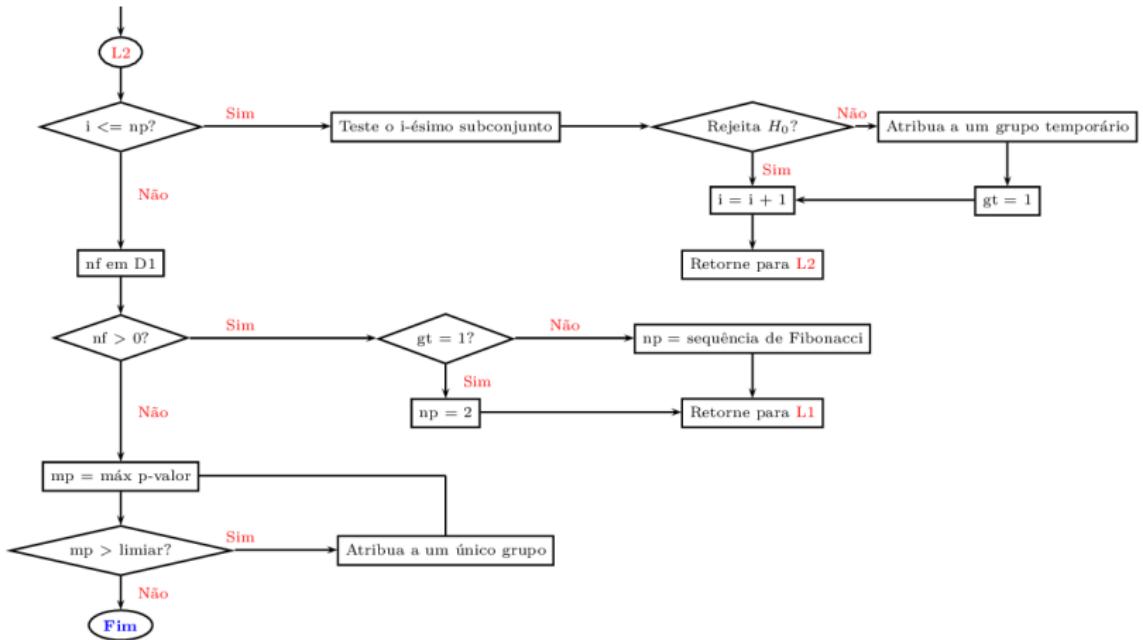
- $\sqrt{a}(F_R - 1) \xrightarrow{d} N(0, \tau_2/v_2^4)$:
 - $F_R = \frac{MST_R}{MSE_R}$
 - $MST_R = \frac{1}{a-1} \sum_{i=1}^a (\bar{R}_{i\cdot} - \tilde{R}_{..})^2$;
 - $MSE_R = \frac{1}{a} \sum_{i=1}^a \frac{1}{n_i} S_{R,i}^2$.
 - $v_2^2 = \lim_{a \rightarrow \infty} \frac{1}{a} \sum_{i=1}^a \frac{1}{n_i} \sigma_i^2 > 0$;
 - $\tau_2 = \lim_{a \rightarrow \infty} \frac{1}{a} \sum_{i=1}^a \frac{2\sigma_i^4}{n_i(n_i - 1)}$



Algoritmo PPCLUST-H







- Pacote R para agrupamento baseado em modelo via mistura de normais, classificação e estimativa de densidade.
- Assume dados com mistura de distribuições normais:

$$f(x) = \sum_{i=1}^G \pi_i f_i(x|\mu_i, \Sigma_i),$$

- x é a matriz de dados;
- π_i é a probabilidade de uma observação pertencer à i -ésima componente ($\pi_i \in [0, 1]$ e $\sum_{i=1}^G \pi_i = 1$);
- f_i é a função densidade de probabilidade gaussiana do i -ésimo grupo com os parâmetros.
- μ_i a média e Σ_i a matriz de covariâncias.
- Técnica de comparação por apresentar melhores resultados em estudos prévios.



Índice de Rand Ajustado (ARI)

- O Índice de Rand Ajustado (Hubert & Arabie, 1985) é uma critério externo para medir a qualidade de agrupamento.
- **ARI = 1**: (agrupamento perfeito) significa que os elementos estão juntos em algum grupo em ambas as partições.
- **ARI = 0**: (agrupamento péssimo) significa que os elementos em cada partição foram alocados aleatoriamente.
- O ARI também é usado como medida de concordância entre dois agrupamentos.



Índice de Rand Ajustado (ARI)

Exemplo:

$$A = \{(1, 2, 3, 4, 5, 6), (7, 8, 9, 10, 11, 12)\} \text{ e}$$

$$B = \{(1, 2, 3, 4), (5, 6, 7, 8), (9, 10, 11, 12)\}.$$

- Tabela de cruzada para o agrupamento.

Grupo	b_1	b_2	b_3	Total
a_1	0	2	4	6
a_2	4	2	0	6
Total	4	4	4	12

$$ARI = 0,36$$



Funções do Pacote Rclust

ppclust(dataset, alpha)

ppclust_h(dataset, alpha, n.cores = 1)

ppclustel(dataset, id, repid, alpha)

ppclustel_h(dataset, id, repid, alpha, n.cores = 1)

Argumentos:

dataset: matriz ou data.frame.

id: o número da coluna que contém o identificador do objeto.

repid: o número da coluna que identifica a repetição.

alpha: limiar.

n.cores: núcleos do processador (n.cores = 1 por padrão).



Resultados do Pacote RCLUST

Análise das Implementações (PPCLUST e PPCLUSTEL)

- Resumo do ARI entre os agrupamentos do SAS e do R.

Algoritmo	ARI	
	Média	D.P.
PPCLUST	0,991	0,003
PPCLUSTEL	0,707	0,079



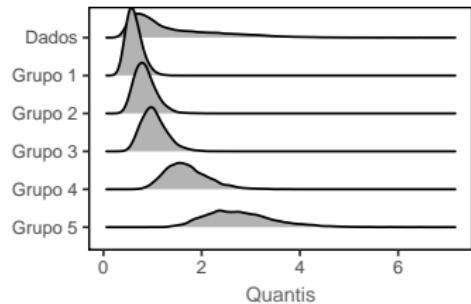
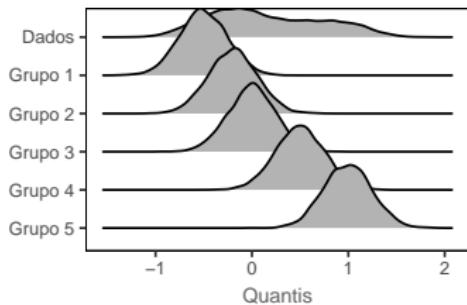
Número de Hipóteses Avaliadas

- Número de testes de hipótese realizados para o agrupamento de um conjunto de dados com 2000 dimensões e 10 observações.
 - PPCLUST: 5079 testes.
 - PPCLUST-H: 50 testes.



PPCLUST-H: Dados Simulados

- 50 conjuntos de dados, com 2000 e 6000 dimensões;
- Observações para cada dimensão: 5, 10 e 15;
- Cada conjunto de dados composto por 5 grupos.



PPCLUST-H: Resultados para os Dados Simulados

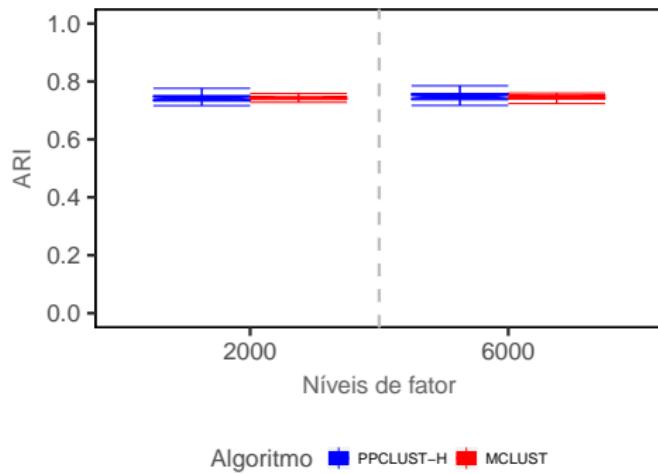
- Resumo do (ARI).

Parâmetros		PPCLUST-H		MCLUST			
		Normal/LogN		Normal		LogN	
a	n_i	Média	D.P.	Média	D.P.	Média	D.P.
2000	5	0,741	0,013	0,746	0,014	0,691	0,014
2000	10	0,861	0,014	0,877	0,009	0,832	0,011
2000	15	0,913	0,016	0,933	0,009	0,896	0,009
6000	5	0,743	0,007	0,745	0,009	0,694	0,009
6000	10	0,863	0,006	0,879	0,004	0,833	0,006
6000	15	0,912	0,013	0,932	0,003	0,896	0,005



PPCLUST-H: Resultados para os Dados Simulados

- ARI para conjuntos de dados gerados com 2000 e 6000 dimensões; 5 observações.



PPCLUST-H: Tempos de Agrupamento

- Resumo dos tempos de processamento.

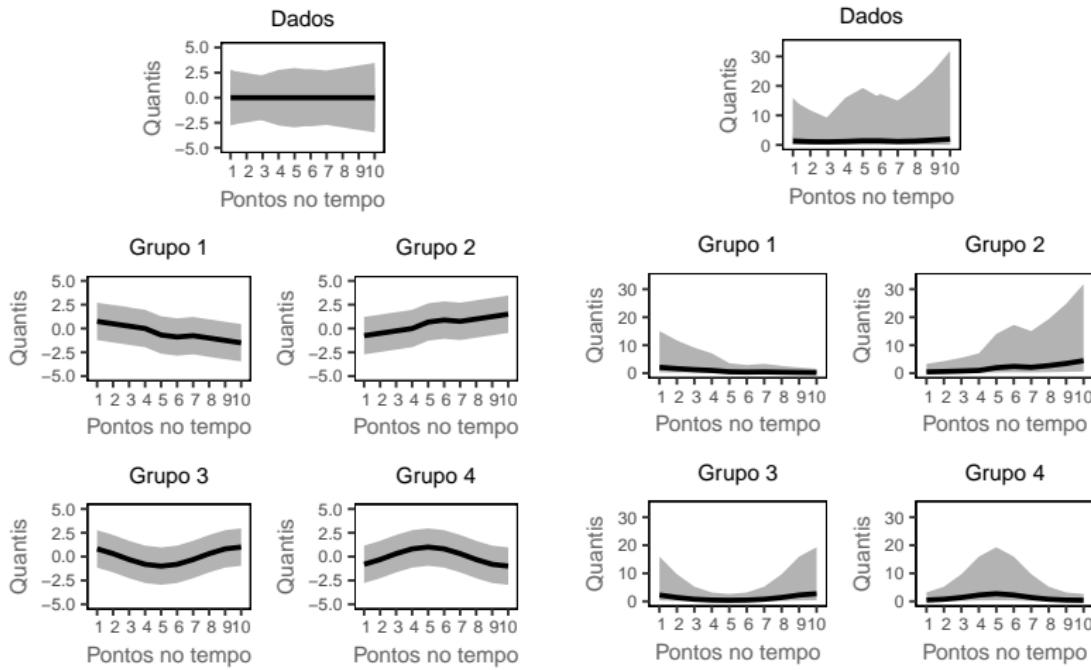
Algoritmo	Dimensões					
	2000		6000		10000	
	Média	D.P.	Média	D.P.	Média	D.P.
PPCLUST-H	0,7s	0,02s	1,5s	0,03s	2,3s	0,05s
PPCLUST-H+	0,8s	0,02s	1,3s	0,02s	1,9s	0,06s
MCLUST	6,4s	0,7s	11s	1,5s	16s	2,7s



- 50 conjuntos de dados, com 2000 e 6000 dimensões;
- 3 observações por dimensão;
- Número de pontos no tempo: 5, 10 e 15;
- Cada conjunto de dados composto por 4 grupos.



PPCLUSTEL-H: Dados Simulados



PPCLUSTEL-H: Dados Simulados

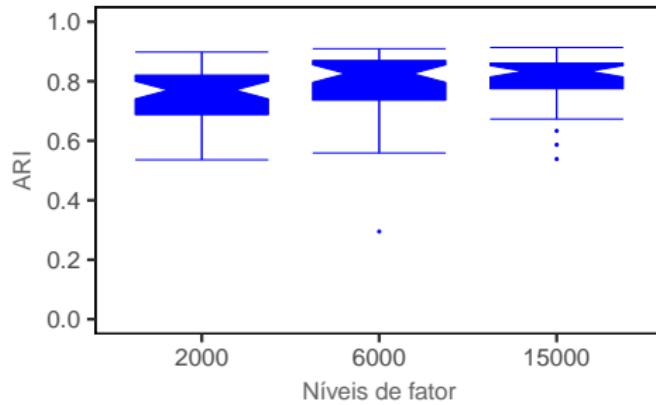
- Resumo do ARI.

Parâmetros		PPCLUSTEL-H		MCLUST	
a	b	Média	D.P.	Média	D.P.
2000	5	0,572	0,056	0,994	0,002
2000	10	0,757	0,092	0,998	0,001
2000	15	0,762	0,130	0,998	0,001
6000	5	0,612	0,062	0,994	0,001
6000	10	0,793	0,109	0,998	0,001
6000	15	0,844	0,086	0,998	0,001



PPCLUSTEL-H: Dados Simulados

- ARI por número de dimensões para o algoritmo PPCLUSTEL-H.



- Resumo do ARI para os algoritmos com o uso dos dados log-normais.

Parâmetros		PPCLUSTEL-H		MCLUST	
<i>a</i>	<i>b</i>	Média	D.P.	Média	D.P.
2000	5	0,593	0,068	0,503	0,068
2000	10	0,739	0,114	0,485	0,051
2000	15	0,747	0,122	0,477	0,055
6000	5	0,651	0,081	0,513	0,064
6000	10	0,805	0,099	0,484	0,054
6000	15	0,807	0,133	0,480	0,053



PPCLUSTEL-H: Dados Simulados

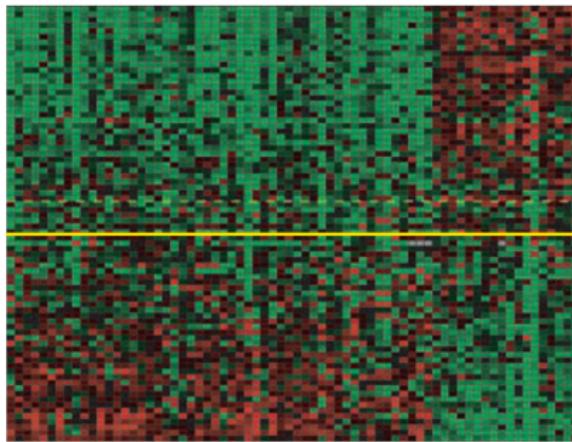
- Resumo dos tempos de processamento.

Algoritmo	Dimensões					
	2000		6000		10000	
	Média	D.P.	Média	D.P.	Média	D.P.
PPCLUSTEL-H	1,9s	0,1s	5s	0,2s	8,3s	0,2s
PPCLUSTEL-H+	1,9s	0,1s	4,5s	0,2s	7,4s	0,7s
MCLUST	3,1s	0,1s	3,7s	0,7s	4,2s	1,5s



Agrupamento de Dados de Expressão Gênica

- Expressão gênica de 70 genes que apresentaram padrões de expressão para a detecção da recorrência de câncer de mama dentro de 5 anos de tratamento (Pierce, 2012).

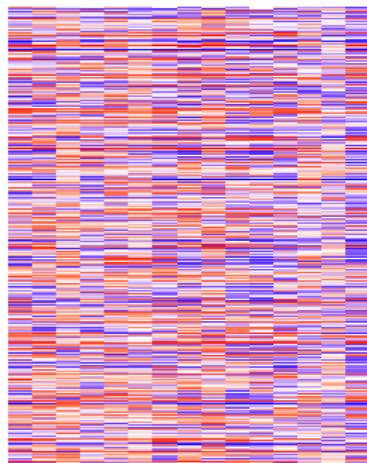
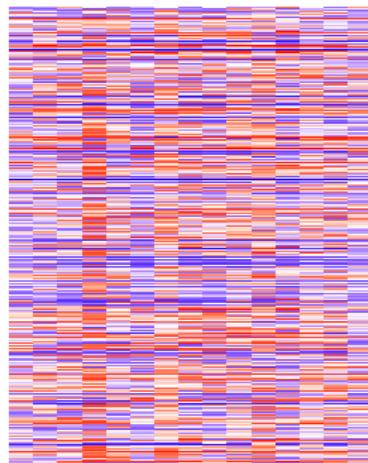


- Dados de expressão gênica de adenoma e câncer de cólon.
 - Conjunto de dados microarrajo pré-processado por Corrada Bravo et al. (2012).
 - Expressão gênica de 5339 genes das biópsias de 15 pacientes com adenoma, 15 pacientes com câncer e de 8 pacientes saudáveis (controles).
 - O PPCLUST-H foi aplicado ($\text{limiar} = 10^{-8}$) para detectar a diferença dos níveis de expressão gênica entre tecidos doentes e tecidos normais.



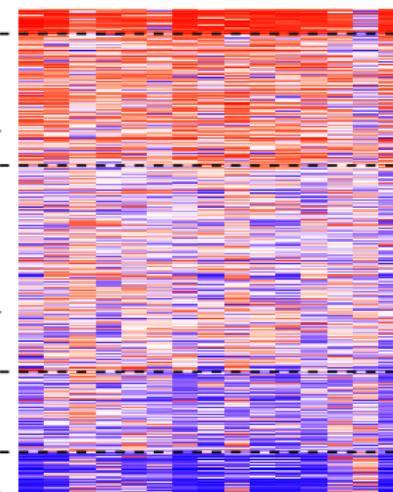
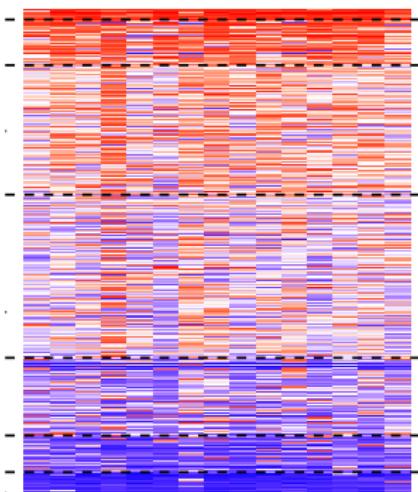
PPCLUST-H: Dados Reais

- Heatmaps para os dados de expressão gênica.



PPCLUST-H: Dados Reais

- *Heatmaps* após o agrupamento realizado pelo PPCLUST-H.



PPCLUST-H: Dados Reais

- Grupos de genes diferentemente expressos em tecidos com adenoma segundo o PPCLUST-H.

Grupo	Qtd. genes	% genes	p-valor
1	108	2	0,003
2	504	9,4	0,049
3	1430	26,8	0,661
4	1802	33,7	0,703
5	858	16	0,521
6	404	7,6	0,999
7	233	4,4	0,999



PPCLUST-H: Dados Reais

- Grupos de genes diferentemente expressos em tecidos com câncer segundo o PPCLUST-H.

Grupo	Qtd. genes	% genes	p-valor
1	265	4,9	$4,9 \times 10^{-6}$
2	1454	27,3	0,999
3	2280	42,7	0,999
4	886	16,6	0,969
5	454	8,5	0,999

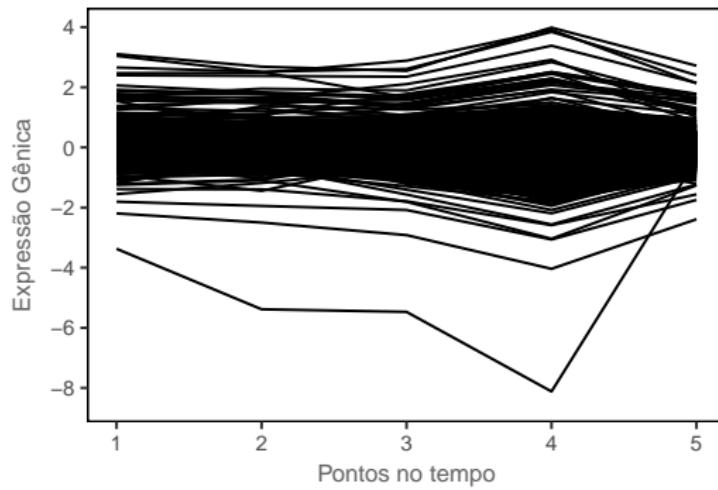


- Dados de expressão gênica de leveduras (fungos).
 - Conjunto de dados microarrajo estudado em Gillespie et al. (2010) e padronizado em Silva (2012).
 - Expressão gênica de 10,928 genes de leveduras:
 - 5 pontos no tempo com 3 repetições.
 - Dois tipos de condições experimentais: cepas de leveduras sem acréscimo de temperatura (controle) e cepas com alteração de temperatura (tratamento).
 - O PPCLUSTEL-H foi aplicado ($\text{limiar} = 10^{-5}$) para detectar a diferença dos níveis de expressão gênica ao longo do tempo.
 - O MCLUST foi utilizado com a informação do número de grupos encontrado pelo PPCLUSTEL-H.



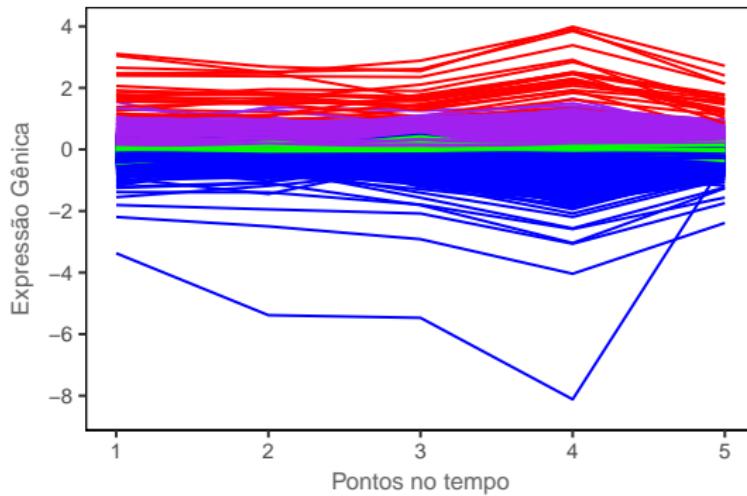
PPCLUSTEL-H: Dados Reais

- Expressão gênica padronizada das leveduras.



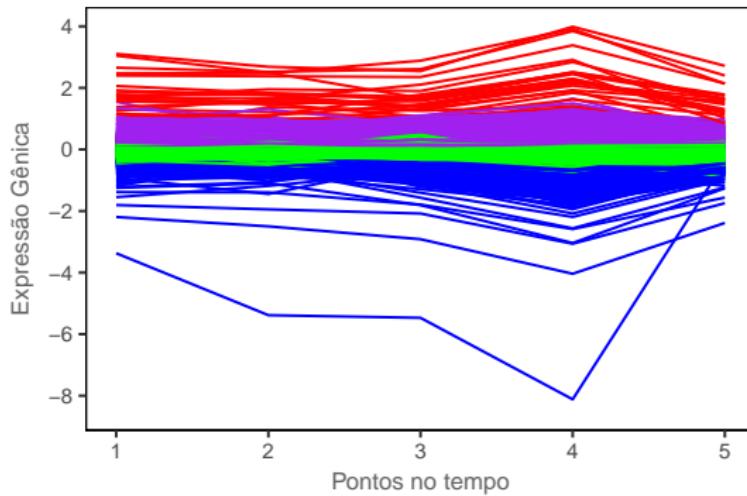
PPCLUSTEL-H: Dados Reais

- Agrupamento obtido pelo algoritmo PPCLUSTEL-H com os dados originais.



PPCLUSTEL-H: Dados Reais

- Agrupamento obtido pelo algoritmo MCLUST com a especificação do número de grupos obtido pelo PPCLUSTEL-H.



- O pacote R está disponível em:

`github.com/rafaelslins/Rclust`



Conclusões

Conclusões

- Os procedimentos PPCLUST-H e PPCLUSTEL-H apresentaram resultados promissores no agrupamento de dados superdimensionados.
- É possível o uso colaborativo dos algoritmos com base em p -valor com o MCLUST.
- O MCLUST tende a superar ambos os algoritmos com base em p -valor apenas quando os dados seguem distribuições simétricas.
- O uso de processamento paralelo no *R* proporcionou resultados positivos.
- É interessante o desenvolvimento de uma metodologia para a escolha do limiar.



Referências Principais i

-  H. Corrada Bravo, V. Pihur, M. McCall, R. A. Irizarry, and J. T. Leek.

Gene expression anti-profiles as a basis for accurate universal cancer signatures.

BMC Bioinformatics, 13(1):272, Oct 2012.

-  L. Kaufman and P. Rousseeuw.

Finding Groups in Data: An Introduction to Cluster Analysis.

09 2009.



-  R Core Team. **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria, 2017.
-  A. E. Raftery, M. Fop, T. B. Murphy, and L. Scrucca. **mclust 5: clustering, classification and density estimation using Gaussian finite mixture models.** *The R Journal*, 8(1):205–233, 2016.
-  A. P. T. Silva. **Implementação, análise e aplicação de algoritmos de agrupamento de dados superdimensionados, longitudinais e com amostras pequenas.**



Dissertação (Mestrado em Estatística), Universidade de Brasília., 2012.

- 📄 G. von Borries.

Partition clustering of High Dimensional Low Sample Size data based on P-Values.

Tese (Doutorado em Estatística), 2008.

- 📄 G. von Borries.

A SAS/JMP Integration for Implementation of a Clustering Algorithm for High Dimensional Low Sample Size Data.

SAS Global Forum. Paper 283, 2009.





H. Wang and M. G. Akritas.

Rank tests for anova with large number of factor levels.

Journal of Nonparametric Statistics, 16(3-4):563–589, 2004.



Obrigado