Outcome-Guided Sparse K-Means for Disease Subtype Discovery via Integrating Phenotypic Data with High-Dimensional Transcriptomic Data

Group 7: I Hung, Nathan, Paul

Penn State University

1. Problem Description and Modeling Objective

The challenge in precision medicine is to identify disease subtypes linked to clinical outcomes, as traditional methods often miss complex connections between biological data and clinical traits, resulting in less effective treatments. This study introduces the GuidedSparseK-means algorithm, which integrates clinical data with high-dimensional omics data to address the need for innovative approaches. The algorithm employs a unified objective function that combines weighted K-means clustering, lasso regularization for gene selection, and the inclusion of a phenotypic variable. By iteratively optimizing this function, GuidedSparseK-means aims to yield statistically robust and clinically relevant subtypes, enhancing the potential for personalized treatment strategies.

2. Data Description

The study utilizes high-dimensional gene expression data and clinical phenotypic data from modern epidemiological cohorts, specifically focusing on transcriptomic datasets related to breast cancer and Alzheimer disease.

Gene expression data is used to understand how genes are active in different conditions, such as in healthy versus diseased tissues. By measuring which genes are turned on or off, we can learn about the biological processes happening in a cell. [Emilsson et al., 2008] This data helps identify genes involved in diseases like breast cancer and Alzheimer disease, as used in the study [Meng et al., 2022].

The METABRIC breast cancer gene expression data [Curtis et al., 2012] included samples from tumor banks in the United Kingdom and Canada. It contains gene expression profile of 20,603 genes and 2,501 subjects (see Fig. 1 for details). It also contains various clinical data for every sample, particularly the Nottingham prognostic index (NPI); Estrogen receptor status (ER); HER2 receptor status (HER2); and overall survival (see Fig. 2 for details).

Breast cancer is often tested for three key factors: NPI, ER, and HER2 status, which help guide treatment [Haybittle et al., 1982, Jensen and DeSombre, 1973, Di Fiore et al., 1987]. NPI is a score based on the size of the tumor, the grade of the cancer cells, and how many lymph nodes are affected, helping predict how aggressive the cancer is and how likely it is to spread. ER status checks if the cancer cells have receptors for estrogen, a hormone that can fuel cancer growth. If the cancer is ER-positive, it means estrogen helps the cancer grow, and treatments that block estrogen, like hormone therapy, can be effective. HER2 is a protein that helps cancer cells grow, and if the cancer is HER2-positive, it means the cancer cells have too much of this protein. Targeted therapies like Herceptin can help treat HER2-positive cancers. Together, these tests help doctors determine the best treatment plan for each patient.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Hugo_Sym	Entrez_Ge	MB-0362	MB-0346	MB-0386	MB-0574	MB-0185	MB-0503	MB-0641	MB-0201	MB-0218	MB-0316	MB-0189	MB-0891	MB-0658	MB-0899	MB-0605	MB-0258	MB-0506
2	RERE	473	8.67698	9.65359	9.03359	8.81486	8.73641	9.27426	9.28659	8.43735	8.56997	8.8921	8.92275	9.05397	8.293	8.64369	8.57613	9.14574	8.54905
3	RNF165	494470	6.07533	6.68789	5.91088	5.62874	6.39242	5.9087	6.20673	6.09559	6.38353	5.77369	6.46132	6.09347	6.72509	7.01022	6.05868	6.24806	5.51715
4	PHF7	51533	5.83827	5.60088	6.03072	5.84943	5.54213	5.96466	5.78337	5.73757	5.92393	5.90222	5.53676	5.85508	5.65949	5.88615	6.03453	6.15971	5.53294
5	CIDEA	1149	6.3975	5.24632	10.1118	6.11687	5.1841	7.82817	8.74415	5.48009	5.67158	5.5095	7.19694	9.28632	7.17682	6.90679	8.25508	8.19914	5.20978
6	TENT2	167153	7.90622	8.26726	7.95929	9.20638	8.16284	8.70665	8.51893	7.47841	8.41058	7.77955	8.3859	8.39676	8.17739	7.83995	8.09859	8.30729	7.31616
7	SLC17A3	10786	5.70238	5.52179	5.68953	5.43913	5.46433	5.41748	5.62988	5.68629	5.77027	5.51062	5.59435	5.45831	5.43193	5.59575	5.61916	5.59699	5.75519
8	SDS	10993	6.93074	6.14169	6.52931	6.4301	6.10543	6.68489	5.63275	5.86613	7.4031	6.88175	6.48487	6.11408	7.12992	6.58553	5.89402	7.07497	6.8457
9	ATP6V1C2	245973	5.33286	7.56348	5.48216	5.39867	5.02602	5.26667	5.70135	6.40314	5.50814	5.70904	5.39528	5.48269	5.36261	5.31729	5.65967	5.2985	6.63614
10	F3	2152	5.27568	5.37638	5.46379	5.40976	5.33858	5.49069	5.36327	6.34186	5.61131	5.90692	6.7553	6.36774	5.96516	6.31249	6.70876	5.53585	5.52036
11	FAM71C	196472	5.4439	5.31986	5.25429	5.5123	5.43087	5.36338	5.19161	5.20838	5.6519	5.45213	5.35786	5.46647	5.35046	5.53991	5.484	5.45605	5.65681
12	LIN52	91750	6.65491	6.54614	6.29081	6.27045	6.27157	6.57509	6.33967	6.66956	6.90227	6.07837	6.89618	6.53393	6.53545	6.10532	6.33374	6.70893	6.73589
13	PCOTH	542767	6.11627	5.49612	6.203	5.88299	6.08954	5.7663	5.29765	5.87233	5.98641	5.5258	5.5638	5.93015	5.41094	5.34257	5.41747	5.88525	5.78373
14	GRM1	2911	5.50246	5.3961	5.16625	5.31479	5.26856	5.30381	5.25464	5.16514	5.37929	5.2677	5.21791	5.29465	5.43553	5.58776	5.39449	5.54782	5.4149
15	FXN	2395	6.17085	6.4621	5.88103	6.05984	6.21758	6.42734	6.19281	6.24501	5.99463	6.03673	6.12372	6.02767	6.07558	6.00942	6.04115	5.99431	6.29153
16	SLC9A1	6548	10.0143	9.34853	8.81178	8.95206	9.49188	9.29401	9.07449	9.07348	8.9175	9.20607	9.58331	9.25585	8.64806	8.97137	9.48795	8.87105	9.59564
17	PML	5371	6.22371	5.77166	6.53268	6.11359	6.26953	6.2184	6.69846	6.21369	6.20647	6.93831	6.01376	6.28749	6.48732	6.22253	6.34822	6.63308	6.17536
18	CD164	8763	5.45383	5.8566	5.3812	5.89742	5.86652	5.42492	5.30363	5.29574	5.75063	5.72476	6.14756	6.06548	5.80772	6.44996	5.84967	6.10402	5.70873
19	MOB3A	126308	6.78111	6.9323	6.67118	6.70813	6.34476	6.97223	7.16078	7.0827	6.72793	6.86308	6.77174	6.73445	6.80294	6.7537	7.00117	6.46617	6.59902
20	HGC6.1.1	26236	7.50408	5.47009	5.48708	5.23819	5.11672	5.33222	5.19284	5.19416	5.2079	5.36623	5.22948	5.50549	5.60138	5.39885	5.59709	5.40812	5.46695

Fig. 1: METABRIC breast cancer gene expression dataset

Alzheimer disease (AD) gene expression data [Srinivasan et al., 2020] contains 39,376 transcripts from post-mortem fusiform gyrus tissues of 289 AD subjects (see Fig. 4 for details). It also contained various clinical data for every sample, particularly the Braak stages of AD (see Fig. 5 for details).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Study ID	Patient ID	Sample ID	Age at Di	Type of B	Cancer T	Cancer T	Cellularity	Chemoth	Pam50 +	Cohort	ER status	ER Status	Neoplasm	HER2 sta	HER2 Sta	Tumor Gr	Hormone	Inferred M	Integrativ	Primary T	Lymph no	Mutation	Nottingh	Oncotree	Overall S	Overall S
2	brca_me	MB-0000	MB-0000	75.65	MASTEC	Breast C	Breast In	NA	NO	claudin-l	1	Positive	Positive	3	NEUTRA	Negative	Ductal/IN	YES	Post	4ER+	Right	10	NA	6.044	IDC	140.5	0:LIVING
3	brca_me	MB-0002	MB-0002	43.19	BREAST	Breast C	Breast In	High	NO	LumA	1	Positive	Positive	3	NEUTRA	Negative	Ductal/IN	YES	Pre	4ER+	Right	0	2	4.02	IDC	84.633	0:LIVING
4	brca_me	MB-0005	MB-0005	48.87	MASTEC	Breast C	Breast In	High	YES	LumB	1	Positive	Positive	2	NEUTRA	Negative	Ductal/IN	YES	Pre		Right	1	2	4.03	IDC	163.7	1:DECEA
5	brca_me	MB-0008	MB-0008	47.68	MASTEC	Breast C	Breast M	Moderate	YES	LumB	1	Positive	Positive	2	NEUTRA	Negative	Mixed	YES	Pre		Right	3	1	4.05	MDLC	164.93	0:LIVING
6	brca_me	MB-0008	MB-0008	76.97	MASTEC	Breast C	Breast M	High	YES	LumB	1	Positive	Positive	3	NEUTRA	Negative	Mixed	YES	Post		Right	8	2	6.08	MDLC	41.367	1:DECEA
7	brca_me	MB-0010	MB-0010	78.77	MASTEC	Breast C	Breast In	Moderate	NO	LumB	1	Positive	Positive	3	NEUTRA	Negative	Ductal/IN	YES	Post		Left	0	4	4.062	IDC	7.8	1:DECEA
8	brca_me	MB-0014	MB-0014	56.45	BREAST	Breast C	Breast In	Moderate	YES	LumB	1	Positive	Positive	2	LOSS	Negative	Ductal/IN	YES	Post		Right	1	4	4.02	IDC	164.33	0:LIVING
9	brca_me	MB-0020	MB-0020	70	MASTEC	Breast C	Breast In	High	YES	Normal	1	Negative	Negative	3	NEUTRA	Negative	Lobular	NO	Post	4ER-	Left	NA	NA	6.13	ILC	22.4	1:DECEA
10	brca_me	MB-0022	MB-0022	89.08	BREAST	Breast C	Breast M	Moderate	NO	claudin-l	1	Positive	Positive	2	NEUTRA	Negative	Mixed	YES	Post		Left	1	1	4.058	MDLC	93.533	1:DECEA
11	brca_me	MB-0025	MB-0025	76.24	NA	Breast C	Breast In	NA	NA	NA	1	Positive	Positive	3	NA	NA	Ductal/IN	NA	NA	NA	NA	11	5	6.68	IDC	NA	NA
12	brca_me	MB-0026	MB-0026	86.41	BREAST	Breast C	Breast In	Moderate	NO	LumB	1	Positive	Positive	3	GAIN	Negative	Ductal/IN	YES	Post		Right	1	4	5.032	IDC	36.567	1:DECEA
13	brca_me	MB-0035	MB-0035	84.22	MASTEC	Breast C	Breast In	High	NO	Her2	1	Negative	Positive	2	LOSS	Negative	Lobular	NO	Post		Left	0	5	3.056	ILC	36.267	1:DECEA
14	brca_me	MB-0036	MB-0036	85.49	MASTEC	Breast C	Breast In	Moderate	NO	LumA	1	Positive	Positive	2	NEUTRA	Negative	Ductal/IN	YES	Post		Left	0	1	3.044	IDC	132.03	1:DECEA
15	brca_me	MB-0039	MB-0039	70.91	BREAST	Breast C	Breast In	High	NO	LumB	1	Positive	Positive	1	GAIN	Negative	Ductal/IN	YES	Post	4ER+	Left	0	3	2.042	IDC	163.53	0:LIVING
16	brca_me	MB-0045	MB-0045	45.27	MASTEC	Breast C	Breast In	High	YES	claudin-l	1	Negative	Negative	3	NEUTRA	Negative	Ductal/IN	NO	Pre	4ER-	Right	3	NA	5.038	IDC	164.9	0:LIVING
17	brca_me	MB-0046	MB-0046	83.02	MASTEC	Breast C	Breast In	High	NO	LumA	1	Positive	Positive	3	GAIN	Positive	Ductal/IN	YES	Post		Left	24	2	6.072	IDC	14.133	1:DECEA
18	brca_me	MB-0048	MB-0048	51.46	BREAST	Breast In	Breast In	Low	YES	claudin-l	1	Positive	Positive	2	GAIN	Positive	Ductal/IN	YES	Post	4ER+	Left	1	NA	4.05	IDC	103.83	0:LIVING
19	brca_me	MB-0050	MB-0050	44.64	BREAST	Breast C	Breast M	Moderate	YES	Normal	1	Positive	Positive	2	NEUTRA	Negative	Mixed	YES	Pre		Right	3	NA	4.066	MDLC	75.333	0:LIVING
20	brca_me	MB-0052	MB-0052	70.02	BREAST	Breast C	Breast In	High	NO	LumB	1	Positive	Positive	2	NEUTRA	Negative	Ductal/IN	YES	Post		Right	0	NA	3.046	IDC	161.07	0:LIVING

Fig. 2: METABRIC breast cancer clinical data

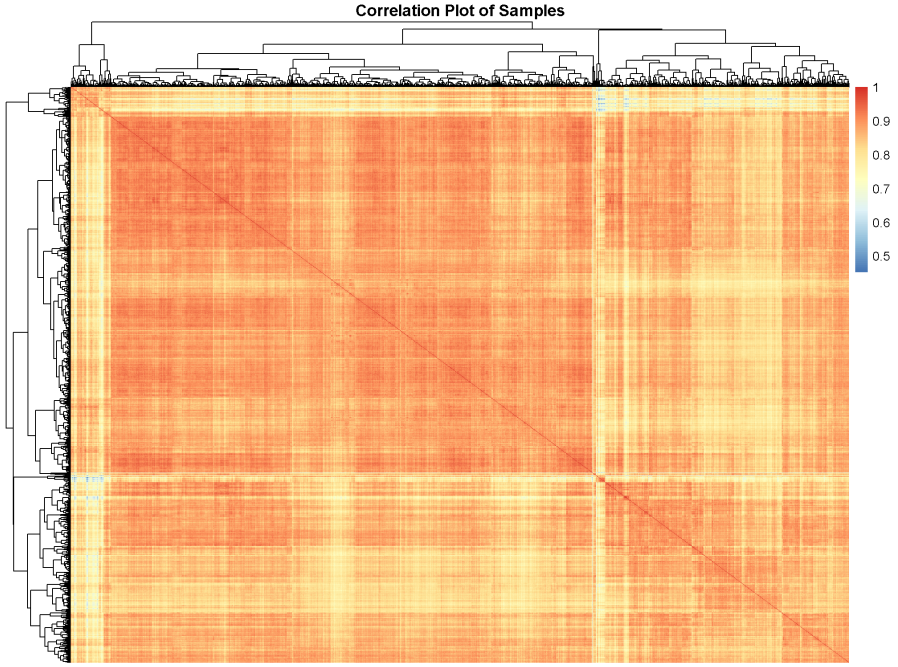

Fig. 3: Correlation plot of METABRIC breast cancer gene expression dataset

The Braak stages [Braak and Braak, 1991] describe how AD progresses in the brain, particularly on the spread of tau tangles. In Stage I, the tangles begin in the part of the brain responsible for memory. By Stage II, they spread to areas involved in memory processing. In Stage III, tangles reach regions related to both memory and emotions. As the disease advances to Stage IV, the tangles spread to more areas of the brain, leading to noticeable memory problems. In Stage V, tangles affect larger parts of the brain, and cognitive issues become more severe. Finally, in Stage VI, the tangles are widespread across the brain, and individuals experience severe dementia, with significant

impairments in memory and cognition. As the Braak stages progress, the severity of memory and thinking problems increases.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	GeneID	GSM3577568	GSM3577569	GSM3577570	GSM3577571	GSM3577572	GSM3577573	GSM3577574	GSM3577575	GSM3577576	GSM3577577	GSM3577578	GSM3577579	GSM3577580	GSM3577581
2	1E+08	15	34	32	50	28	7	18	13	14	16	13	13	7	26
3	653635	494	732	557	813	748	499	896	519	655	566	879	976	416	688
4	1.02E+08	2	13	9	9	16	7	18	8	15	8	13	13	9	21
5	1.08E+08	2	2	1	2	3	2	3	2	4	2	2	2	1	2
6	1E+08	0	0	1	0	0	1	1	0	1	0	0	0	0	1
7	645520	1	2	2	2	6	1	2	3	1	2	2	1	1	3
8	79501	0	0	0	0	0	0	0	0	0	0	0	1	0	0
9	1.01E+08	17	71	59	45	52	48	110	37	124	38	58	91	76	130
10	729737	102	581	289	434	348	110	355	355	315	160	272	380	175	560
11	1.03E+08	12	26	36	49	25	3	11	9	12	12	7	6	6	23
12	1.03E+08	495	775	598	911	811	570	989	555	757	599	913	973	454	775
13	1.02E+08	2	13	9	9	16	7	17	8	15	9	13	13	7	20
14	1.08E+08	2	3	2	7	5	3	11	6	11	4	5	7	7	8
15	1.12E+08	9	15	9	14	11	21	60	19	54	10	13	34	43	43
16	729759	0	1	0	1	0	0	1	0	1	1	0	1	2	0
17	1E+08	76	420	226	216	268	117	238	228	217	130	206	225	149	394
18	1.05E+08	0	5	1	4	1	2	3	1	9	2	1	2	5	4
19	1.02E+08	0	0	1	1	0	1	0	1	0	0	0	0	1	0
20	1.13E+08	28374	54478	44519	49735	48285	21823	36794	71929	15953	37333	40784	24383	18691	21703

Fig. 4: Alzheimer disease gene expression dataset

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID_REF	GSM3577568	GSM3577569	GSM3577570	GSM3577571	GSM3577572	GSM3577573	GSM3577574	GSM3577575	GSM3577576	GSM3577577	GSM3577578
2	stage	braak.score: IV	braak.score: VI	braak.score: IV	braak.score: IV	braak.score: IV	braak.score: IV	braak.score: IV	braak.score: IV	braak.score: IV	braak.score: IV	braak.score: IV

Fig. 5: Alzheimer disease clinical data

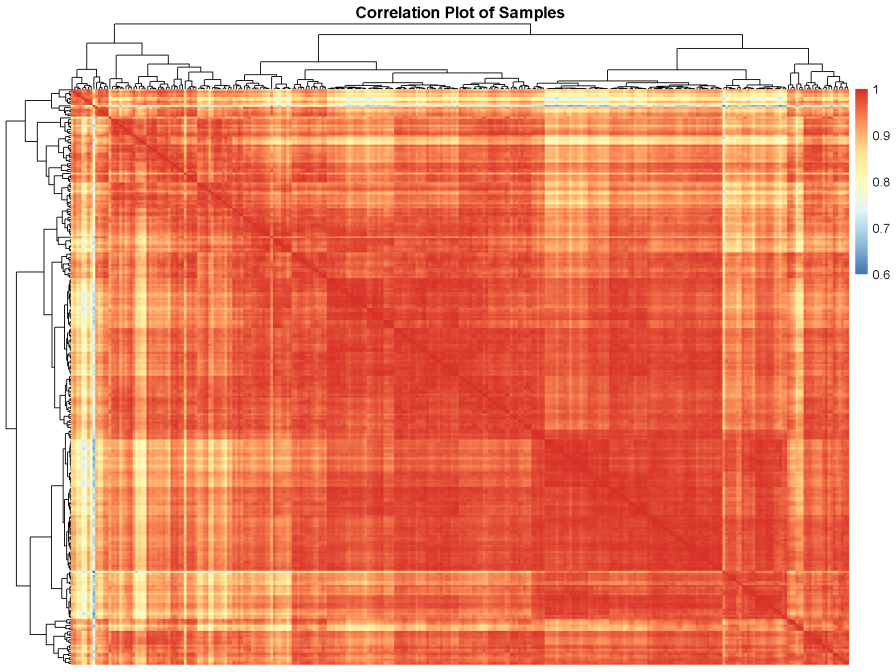

Fig. 6: Correlation plot of Alzheimer disease gene expression dataset

3. Model and Methods Description

The GuidedSparseK-means algorithm enhances traditional sparse K-means clustering by integrating clinical phenotypic data into the clustering process. The model employs a unified objective function that comprises three key components: a weighted K-means approach for effective sample clustering, lasso regularization for gene selection from high-dimensional omics data, and the incorporation of a phenotypic variable to guide the clustering towards biologically meaningful outcomes. This iterative optimization process simultaneously refines both sample clustering and gene selection, allowing the algorithm to uncover subtypes that are not only statistically robust but also aligned with clinical relevance. Through simulations and applications to real-world datasets, the performance of GuidedSparseK-means is benchmarked against existing clustering methods, demonstrating its superiority in capturing the complex relationships inherent in the data.

3.1. K-means Algorithm

The first clustering method the paper described is the normal k means algorithm. It considers X_{gi} to be the gene expression of gene g and subject i. Let the total number of genes be G and the total number of subjects be n. The normal k means algorithm finds K clusters of our n subjects such that the clusters are close together (usually in the euclidean distance). Denote such an optimal clustering as $\hat{C}$. This is done by minimizing the objective function

$$\hat{C} = \text{argmin}_C \sum_{g=1}^{G} WCCS_g = \text{argmin}_C \sum_{g=1}^{G} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{(i,j)\in C_k} (X_{gi} - X_{gj})^2. \tag{1}$$

Where C is a partition of our data set into K clusters and C_k for $1 \leq k \leq K$ is the elements in the kth cluster and n_k is the size of cluster k.

3.2. Sparse K-means Algorithm

Due to the nature of genetics data, we have that there are usually a significantly more number of genes than we have subjects. Thus, there is usually an assumption of sparsity in the number of genes that would actually have a significant impact on the clustering

of subjects. The solution to taking into account this sparsity into our clustering is to maximize the weighted between cluster sum of squares subject to a LASSO like penalty. This usage of between cluster sum of squares is to avoid the issue of a trivial solution to equation (1) from a naive implementation of a lasso penalty. The sparse K means clustering, denoted as $\hat{C}_{Sparse}$, is found by maxizing the objective function

$$\hat{C}_{Sparse} = \text{argmax}_{C,w} w_g \left[\frac{1}{n} \sum_{i,j \in C} (X_{gi} - X_{gj})^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} (X_{gi} - X_{gj})^2 \right] \quad (2)$$

subject to $||w||_2 \leq 1$, $||w||_1 \leq s$, $w \geq 0$ and the weight vector discarded from the argmax. Above, we have that s is a tuning parameter in order to control the number of genes we select. The L_1 penalty is chosen to cause gene selection and the L_2 penalty is additionally used to facilitate selecting more than one gene. Notice that the total sum of squares term is constant for all clusters and our problem thus involves minimizing the within cluster sum of squares with the additional weights.

3.3. GuidedSparseKMeans

The main novelty that this paper introduces is the addition of a guided term to our objective function for sprase k means. This is motivated by a desire to include information that we have from clinical outcomes into our clustering. We denote the association between a gene g and a given clinical outcome variable as U_g. In general the U_g is defined in terms of a function $U : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that $U_g = U(x_g, y)$, where x_g is the expression level of gene g across all subjects and y is the vector of clinical outcomes. From above, we can see that $BCSS_g(C) = \left[\frac{1}{n} \sum_{i,j \in C} (X_{gi} - X_{gj})^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} (X_{gi} - X_{gj})^2 \right]$ and $TSS_g = \frac{1}{n} \sum_{i,j \in C} (X_{gi} - X_{gj})^2$. With these, we can define our objective function fo the GuidedSparseKmeans with the optimal clustering being $\hat{C}_{GSK}$

$$\hat{C}_{GSK} = \text{argmax}_{C,w} \sum_{g=1}^{G} w_g \left[\frac{BCSS_g(C)}{TSS_g} + \lambda U_g \right] \quad (3)$$

subject to the constraints $||w||_2 \leq 1$, $||w||_1 \leq s$, $w \geq 0$. Similarly to above, the weight vector component of the argmax is discarded. Further, λ is a tuning parameter to control how much of the clustering is guided by the clinical outcome and how much is guided

by how much gene g affects the clustering separation. Notice further that the term corresponding to gene g separation ability is normalized by TSS_g to allow this term to be comparable with the scale of U_g. In the paper, the author chose to choose a U_g such that it is based on a regression model f_g which is univariate. Namely $y = f_g(x_g) + \varepsilon$ with ε some error is the relationship between outcomes and gene expression. They defined U_g such that it is the pseudo R-squared of this regression function. Namely,

$$U_g = 1 - \left[\frac{L(f_0)}{L(f_g)}\right]^{2/n},$$

where $L(f_0)$ is the likelihood of the null model and $L(f_g)$ is the likelihood of the f_g model. The GuidedSparseKMeans is computed in Algorithm 1.

Algorithm 1 GuidedSparseKmeans Algorithm

Require: Number of clusters K, λ, s

Input: Number of subjects n, number of subjects G, Gene data $X \in \mathbb{R}^{n \times G}$, clinical outcome U_g

Output: Optimal clustering $\hat{C}_{GSK}$ for GuidedSparseKmeans

1: **Initialization** Let $U_g^* = U_g I(U_g \geq U_{(r)})$, where $U_{(r)}$ is the rth largest value of U_j for $j = 1, \ldots, G$ (r = 400 was chosen for most of the implementation by the author). Initialize $w_g = \frac{U_g^*}{\sum_{g=1}^{G} U_g^*} \times s$

2: Fix w and update C_k with weighted K means. Namely we have $C(w) = \arg\max_C \sum_{g=1}^{G} w_g BCSS_g(C)$.

3: With C fixed, update w as $w_g(C) = \frac{S(a_g, b)}{||S||_2}$, wehre $S(x, c) = max(x - c, 0)$, $a_g = \frac{BCSS_g(C)}{TSS_g} + \lambda U_g$ where b is chosen such that the update weight vector as L_1 norm of s and $b = 0$ is $||w||_1 \leq s$ and $S := (S(a_1, b), \ldots S(a_G, b)) \in \mathbb{R}^G$.

4: Iterate the previous steps 2 and 3 until $\frac{||w^t - w^{t-1}||_1}{||w^{t-1}||_1} \leq 10^{-4}$ with t being the number of iterations and w^t being the weight on the t-th iteration

Note that step 2 is done by clustering our data scaled by multiplying the columns by the square root of the nonnegative weight vectors. The author discusses that this is equivalent to updating the C_k as a weighted k means. To update step 3, the author implements a bisection method starting with $b_1 = 0$ and $b_2 = \max(objective)$ where

objective is the list of the objective values per column. Then we iterate along the interval replacing each endpoint with $\frac{b_1+b_2}{2}$ if our weight vector has either too large 1 norm(b_1 gets replaced) or too low 1 norm(b_2 gets replaced). Lastly, note that to calculate the WCSS and TSS, the author uses $TSS = \sum(X_{gi} = \bar{X}_g)$ and a similar expression for WCSS for simplication of computation.

4. Simulation Studies

The data that was used for the simulation studies was outlined in the paper to best emulate certain characteristics of genetic data. There were 3 subtypes in this example, and the data was simulated to have intrinsic genes, confounding genes and noise genes that have a correlated structure. The ways to generate the 3 types of data are described in algorithm 2.

After generating this data and applying both the clustering methods (the one we implemented and the one implemented in the paper), we implemented the algorithm with tolerance of $1e-6$ and the author implemented the algorithm with tolerance $1e-4$. Both of these clustering methods yielded the same clustering, indicating that the method is not very sensitive to the tolerance level set. Further, our cluster achieved a ARI of 0.804, which is consistent with the ARI levels that the author was getting with the simulated data from their paper. This suggests that our implemented version does a similar clustering that the author does in their algorithm.

5. Results: Application to Alzheimer Disease

This section describes the application of sparse K-means and GuidedSparseKmeans to an RNA-seq dataset from Alzheimer's disease (AD) patients. AD is a severe neurodegenerative disorder that affects older adults. Its key features include neurofibrillary tangles and neuritic plaques. The dataset consisted of 30,727 transcripts from post-mortem fusiform gyrus tissues of 217 individuals. It is available publicly under GEO ID GSE125583. The data was processed by normalizing gene expression values, filtering out the 50% lowest-expressed genes, and standardizing each gene to have a mean of 0 and a standard deviation of 1. This resulted in 15,363 genes used for further analysis.

Algorithm 2 Gene Data Generation

Input: Number of subtypes K, number of gene modules M

Output: A data matrix X containing the simulated gene data and a data vector Y containing clinical response

1: **Initialization** K = 3, M = 20

2: Simulate the number of subjects in each subtype $N_k \sim POI(100)$ for $1 \leq k \leq K$. The total number of subjects we have is thus $N = \sum_{k=1}^{K} N_k$

3: Simulate M gene modules letting $n_m \sim POI(20)$, where n_m is the number of features in module m.

4: Let θ_k be baseline expression of the gene of subtype k and $\theta_k = 2 + 2k$. μ_{km} is the template gene expression of subtype k and module m. We let $\mu_{km} = \alpha_m \theta_k + Z$ with $\alpha_m \sim UNIF((-2, -0.2 \cup (0.2, 2))$, where $Z \sim N(0, 1)$.

5: Generate $X'_{kmi} \sim N(\mu_{km}, 9)$, where 9 is the biological variation in the gene expression.

6: Sample $\Sigma'_{km} \sim W^{-1}(\phi, 60)$, where $\phi = 0.5 I_{n_m \times n_m} + 0.5 * 1_{n_m \times n_m}$, where $1_{x \times y}$ is the $x \times y$ dimensional matrix with elements of all 1.

7: Calculate gene covariance matrix for sutype k and module m Σ_{km} by standardizing Σ'_{km} such that the diagonal is all 1.

8: Intrinsic Gene expression data for subject i in subtype k and module m is $(X_{1kmi}, \ldots, X_{n_m kmi})^T \sim MVN(X'_{kmi}, \Sigma_{km})$

9: Parition your N total subjects randomly into 3 groups.

10: Generate expression data as in steps 3-8 using another 20 modules.

11: Repeat steps 9-10 3 more times to get confounding gene expression data

12: Generate 8000 noise genes with $\mu_g \sim Unif(4, 8)$ and $X_{gi} \sim N(\mu_g, 1)$

13: Simulate clinical response for subject i in subclass k with $Y_{ik} \sim N(\theta_k, 64)$.

The Braak stage was selected as the clinical outcome variable. It is a measure of the severity of neurofibrillary tangles and is widely used in AD research. Since the Braak stage includes six levels (1 to 6), It was treated as a continuous variable. The U_g and R^2 values were calculated from univariate linear regression between each gene and the Braak stage. The number of clusters (K) was set to six, matching the six Braak stages. The paper used the method described above to automatically select the tuning parameter λ. The parameter s was set to select 396 genes, close to the target of 400 genes. The GuidedSparseKmeans algorithm completed the analysis in 7 seconds, while sparse K-means required 22 seconds.

Heatmaps of the selected genes from both methods showed a gradient in expression levels (see Fig. 6 in [Meng et al., 2022]). Expression ranged from low (green) to high (red). Clusters were labeled 1 to 6 to reflect these levels. They evaluated the biological relevance of the AD subtypes by calculating Pearson correlation coefficients between the cluster labels and Braak stages. The GuidedSparseKmeans clusters had a higher correlation (0.282) compared to those from sparse K-means (0.056). This suggests that the subtypes identified by GuidedSparseKmeans were more strongly linked to neurofibrillary tangles.

They also assessed how well the clusters from GuidedSparseKmeans reflected Braak stage information. The relevancy score was high ($Rel = 0.309$), confirming that the algorithm effectively utilized the clinical outcome variable. Pathway enrichment analysis further supported the biological significance of the selected genes. Using the BioCarta database, they found that the genes identified by GuidedSparseKmeans were enriched in the GPCR signaling pathway ($p = 6.59 \times 10^{-6}$) and the NOS1 signaling pathway ($p = 1.61 \times 10^{-5}$). Both pathways are known to play roles in AD. In contrast, sparse K-means did not identify these pathways.

In summary, GuidedSparseKmeans outperformed sparse K-means in identifying AD subtypes and selecting disease-relevant genes. Its ability to incorporate clinical guidance made it more effective in uncovering biologically meaningful insights.

6. Discussion

This paper introduces the outcome-guided sparse K-means algorithm, a novel approach for integrating clinical outcome data with high-dimensional omics data to identify meaningful disease subtypes. The algorithm addresses key challenges in statistical analysis, including gene selection from large datasets, the incorporation of various types of clinical outcome variables (e.g., continuous, binary, ordinal, or survival data), automatic tuning of parameters, and the evaluation of the relevance between the identified subtypes and the clinical outcomes. Its performance has been demonstrated through both simulations and real-world applications, including breast cancer and Alzheimer's disease (AD), where it outperforms traditional sparse K-means methods.

A distinguishing feature of this algorithm is its use of clinical outcomes as guidance, ensuring that the resulting subtypes are both biologically interpretable and clinically relevant. For example, the Braak stage was selected as the guidance variable in the AD application due to its established role in measuring disease severity. Similarly, in breast cancer, HER2-guided clustering produced results with significant survival differences and biologically meaningful pathway analyses. In situations where domain knowledge is limited, the authors recommend testing multiple outcomes and selecting the one that yields the most interpretable clustering results.

The algorithm's parameters, including the number of clusters (K), penalty term (λ), and sparsity level (s), were estimated using gap statistics, sensitivity analysis, and extended gap statistics, respectively. While alternative methods for simultaneous parameter estimation exist, the authors highlight the need for further development of unified approaches. The robustness of the algorithm was evaluated under different scenarios, including varying numbers of selected genes and misspecification of the number of clusters. Simulations indicated that misspecifying K had a significant impact on clustering performance, whereas in real datasets, where cluster boundaries are gradual, the effects were less pronounced.

The computational efficiency of the algorithm is another significant strength. For example, the breast cancer dataset with over 12,000 genes and 1,870 samples required only 31 seconds to analyze. Similarly, the AD dataset with 15,000 genes and 217 samples

was processed in just 7 seconds. The method has been implemented in the R package GuidedSparseKmeans, which is publicly available on GitHub.

7. Conclusion

The outcome-guided sparse K-means algorithm represents a significant advancement in the integration of clinical and omics data. It provides a reliable framework for identifying biologically and clinically meaningful subtypes of diseases. With the growing availability of comprehensive clinical and omics datasets, this method is expected to have broad applicability in understanding complex diseases.

The study by Meng et al. [2022] demonstrated concrete evidence of the advantages of using the outcome-guided sparse K-means algorithm over other methods. They provided both simulation results and analyses from two real-world datasets to support their claims. However, several issues arose when attempting to replicate their results.

Firstly, the data we obtained using the link provided in the paper contained different numbers of genes and subjects than those reported. For instance, in their METABRIC example, they reported 24,368 genes and 1,981 subjects. However, the METABRIC data we obtained from their link contained only 20,603 genes and 2,501 subjects. Similarly, they stated that the Alzheimer's dataset comprised 30,727 transcripts and 217 AD subjects. In contrast, the data we gathered contained 39,376 transcripts and 289 AD subjects.

Secondly, the details of data preprocessing were ambiguous, making it very challenging to arrive at the same final numbers of genes and subjects as reported in the paper. For example, in their motivating example section, they mentioned selecting 408 genes from the 12,180 genes in the METABRIC dataset as input for their algorithm. However, no explanation was provided regarding how these genes were selected.

Thirdly, the code they shared on GitHub was minimal, which made it very difficult for us to replicate their results. It would have been helpful if the code for generating each figure in the paper had been provided.

8. Author Contribution Statement

I Hung: Conceptualization, Investigation, Writing-Original draft preparation. **Nathan:** Conceptualization, Investigation, Methodology, Formal analysis, Writing-Original draft preparation. **Paul:** Conceptualization, Investigation, Data curation, Writing-Original draft preparation, Visualization, Validation.

9. Citations and References

Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G Bragi Walters, Steinunn Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008.

Lingsong Meng, Dorina Avram, George Tseng, and Zhiguang Huo. Outcome-guided sparse k-means for disease subtype discovery via integrating phenotypic data with high-dimensional transcriptomic data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(2):352–375, 2022.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

JL Haybittle, RW Blamey, CW Elston, Jane Johnson, PJ Doyle, FC Campbell, RI Nicholson, and K Griffiths. A prognostic index in primary breast cancer. *British journal of cancer*, 45(3):361–366, 1982.

Elwood V Jensen and Eugene R DeSombre. Estrogen-receptor interaction: Estrogenic hormones effect transformation of specific receptor proteins to a biochemically functional form. *Science*, 182(4108):126–134, 1973.

Pier Paolo Di Fiore, Jacalyn H Pierce, Matthias H Kraus, Oreste Segatto, C Richter King, and Stuart A Aaronson. erb b-2 is a potent oncogene when overexpressed in nih/3t3 cells. *Science*, 237(4811):178–182, 1987.

Karpagam Srinivasan, Brad A Friedman, Ainhoa Etxeberria, Melanie A Huntley, Marcel P van Der Brug, Oded Foreman, Jonathan S Paw, Zora Modrusan, Thomas G Beach, Geidy E Serrano, et al. Alzheimer's patient microglia exhibit enhanced aging and unique transcriptional activation. *Cell reports*, 31(13), 2020.

Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.