

Implementació de models de *deep clustering* en dades metabolòmiques.



Universitat Oberta
de Catalunya

Carles Criado Ninà

Àrea 3

Màster en Bioinformàtica i Bioestadística

Nom del Tutor/a del TF:
Esteban Vegas Lozano

Nom del de/la PRA:
Carles Ventura Royo

15 de gener de 2023



Aquesta obra està subjecta a una llicència de Reconeixement-CompartirIgual
<https://creativecommons.org/licenses/by-nc/3.0/es/>

Fitxa del Treball Final

Títol del treball:	Implementació de models de <i>deep clustering</i> en dades metabolòmiques.
Nom de l'autor/a:	Carles Criado Ninà
Nom del Tutor/a del TF:	Esteban Vegas Lozano
Nom del de/la PRA:	Carles Ventura Royo
Data de lliurament:	15 de gener de 2023
Titulació o programa:	Màster en Bioinformàtica i Bioestadística
Àrea del treball final:	Àrea 3
Idioma del treball:	Català
Paraules clau:	clustering, deep learning, autoencoder, metabolomics

Resum del treball

S'han implementat diversos models de *deep clustering* basats en l'arquitectura Autoencoder amb l'objectiu d'avaluar el seu rendiment en conjunts de dades metabòlomiques. Utilitzant el conjunt de dades MNIST i dos conjunts de dades metabòlomiques, s'ha avaluat el rendiment de diverses variacions de l'arquitectura VAE, DEC i VaDE, utilitzant mètriques de validació interna i externa per mesurar la qualitat de *clustering*. Els resultats s'han contrastat amb mètodes més consolidats com K-means, GMM i *clustering* aglomeratiu. S'ha trobat que l'arquitectura VAE no propicia una bona qualitat de *clustering*. Els clústers obtinguts amb els models DEC, VaDE i les tècniques consolidades mostren un alt nivell de solapament entre ells, obtenen rendiments baixos segons les mètriques de validació. El model DEC destaca sobre la resta en la mètrica de validació interna, però és molt sensible als paràmetres d'inicialització. El model VaDE aconsegueix resultats similars a la resta de tècniques, i presenta el valor afegit tenir capacitat generativa, el que es podria utilitzar en tècniques d'augment artificial de les dades. La distribució multivariant de les covariants (així com la dels metabòlits amb major variabilitat) mostra una distribució diferencial pels clústers obtinguts, tot i que els resultats no són clars. Això suggereix una possible interpretació biològica dels clústers, però serà necessari estudiar-la amb més profunditat per extreure'n conclusions.

Abstract

I implemented several deep clustering models based on the Autoencoder architecture with the aim of evaluating their performance in metabolomic datasets. Using the MNIST dataset and two metabolomic datasets, I evaluated the performance of several variations of the VAE, DEC and VaDE architectures using internal and external validation metrics to measure clustering quality. I compared the results with more established methods such as K-means, GMM and agglomerative clustering. I found found that the VAE architecture is not conducive to good clustering quality. The clusters obtained with the DEC, Vade and consolidated techniques show a high level of overlap with each other, but yield low performances according to the validation metrics. The DEC model excels over the rest in the internal validation metric, but is very sensitive to the initialization parameters. The VaDE model achieves similar results to the rest of the techniques, and has the added value of having generative capacity, which could be used in artificial data augmentation techniques. The multivariate distribution of the covariates (as well as that of the most variable metabolites) shows a differential distribution by the clusters obtained, although the results are not clear. This suggests a possible biological interpretation of the clusters, but it will be necessary to study it in more depth to draw conclusions.

Índex

1	Introducció	9
1.1	Context i justificació del treball	9
1.2	Objectius del treball	10
1.3	Impacte en sostenibilitat, ètic-social i de diversitat	10
1.4	Enfocament i mètode seguit	10
1.5	Planificació del treball	11
1.5.1	Tasques:	11
1.5.2	Calendari:	12
1.5.3	Fites:	12
1.6	Breu sumari de productes obtinguts	12
1.7	Breu descripció dels altres capítols de la memòria	14
2	Estat de l'art	15
2.1	Tècniques de <i>clustering</i>	15
2.2	Reducció de la dimensionalitat	16
2.3	Tècniques de <i>deep clustering</i>	16
2.4	Autoencoder	17
3	Materials i mètodes	19
3.1	Tècniques de <i>clustering</i> clàssiques	19
3.2	Models de <i>deep clustering</i>	20
3.2.1	Deep embedded clustering	20
3.2.2	Variational Autoencoder	22
3.2.3	Optimització dels paràmetres	24
3.2.4	Descens de gradients	25
3.2.5	Implementació del model	25
3.2.6	Variational Deep Embedding	27
3.3	Mètriques d'avaluació	28
3.4	Conjunts de dades	30
3.5	Metodologia	31
3.5.1	MNIST	33
3.5.2	Exposome Data Challenge Event	35
3.5.3	Dades DCH-NG	37
3.6	Eines informàtiques	37

4 Resultats	40
4.1 MNIST	40
4.2 Exposome Data Challenge Event	40
4.3 Dades DCH-NG	42
5 Discussió	49
6 Valoració econòmica	52
7 Conclusions i treballs futurs	53
7.1 Conclusions	53
7.2 Línies de futur	54
7.3 Seguiment de la planificació	55
Glossari i abreviacions	57
Bibliografia	59
A Exposome Data Challenge Event: resultats complets	62
A.1 Taules de mètriques	62
A.2 Heatmaps	74
A.3 Gràfiques radials	76
A.3.1 Conjunt de dades: metaboloma	76
A.3.2 Conjunt de dades: metaboloma (corregit per l'efecte de log)	83
A.3.3 Conjunt de dades: exposoma	88
A.3.4 Conjunt de dades: exposoma (corregit per l'efecte de log)	93

Índex de figures

1.1	Calendari: diagrama de Gantt	13
2.1	Esquema del model Autoencoder	17
3.1	Arquitectura del model DEC mantenint	21
3.2	Alternativa al DEC mantenint descodificador	22
3.3	<i>Variational autoencoder</i> : model probabilístic	23
3.4	Reparametrització de \mathbf{z}	26
3.5	<i>Variational autoencoder</i> : arquitectura	27
3.6	Esquema model VaDE	28
3.7	Exemple gràfica t-SNE	33
3.8	Exemple gràfica radial	33
3.9	Exemple gràfica <i>heatmap</i> de les assignacions de clústers	34
4.1	Comparació dels models VAE, VaDE i DEC	41
4.2	Comparació dels clústers pel conjunt de dades Exposome Data Challenge Event	43
4.3	Distribució diferencial de les variables fenotíp a Exposome Data Challenge Event	43
4.4	Comparació dels clústers pel conjunt de dades DCH-NG	45
4.5	Distribució diferencial de les covariables a DCH-NG	46
4.6	Distribució diferencial dels 20 metabòlits amb més variància a DCH-NG	47
5.1	Comparació dels models VAE+DEC	50

Índex de taules

1.1	Fites i dates clau	12
3.1	Exosome Data Challenge Event: covariables seleccionades	35
3.2	Dades DCH-NG: covariables seleccionades	37
3.3	Tècniques de <i>clustering</i> per conjunt de dades	38
4.1	MNIST: resultats	41
4.2	Exosome Data Challenge Event: resultats (metaboloma)	44
4.3	Exosome Data Challenge Event: resultats (exposoma)	45
4.4	Dades DCH-NG: resultats	48
A.1	Exosome Data Challenge Event: resultats - part 1	63
A.2	Exosome Data Challenge Event: resultats - part 2	65
A.3	Exosome Data Challenge Event: resultats - part 3	66
A.4	Exosome Data Challenge Event: resultats - part 4	67
A.5	Exosome Data Challenge Event: resultats - part 5	68
A.6	Exosome Data Challenge Event: resultats - part6	69
A.7	Exosome Data Challenge Event: resultats - part 7	70
A.8	Exosome Data Challenge Event: resultats - part 8	71
A.9	Exosome Data Challenge Event: resultats - part 9	72
A.10	Exosome Data Challenge Event: resultats - part 10	73

Capítol 1

Introducció

1.1 Context i justificació del treball

Una de les problemàtiques característiques del camp de la bioinformàtica és que les dades que s'estudien habitualment contenen un nombre molt elevat de variables, en comparació al nombre d'observacions. Aquesta elevada dimensionalitat de les dades suposa una dificultat a l'hora de realitzar estudis estadístics.

Entre les tècniques estadístiques utilitzades en dades bioinformàtiques, les tècniques de *clustering* són àmpliament utilitzades. Aquestes tècniques es basen en agrupar les observacions en funció de la seva similitud i permeten extreure informació del conjunt de les dades, com per exemple la seva estructura latent [1, 2].

Les tècniques de *clustering* clàssiques no funcionen bé en dades amb una elevada dimensio-
nalitat, i això suposa una limitació en l'estudi de dades bioinformàtiques.

Una solució que s'utilitza habitualment per combatre aquest problema és aplicar primer una tècnica de reducció de la dimensionalitat, com per exemple anàlisi de components principals (PCA), i aplicar posteriorment les tècniques de *clustering* sobre les dades transformades [2, 3].

En aquest sentit, les xarxes neuronals presenten un avantatge i és que admeten treballar amb dades amb una dimensionalitat elevada. Això ha fet possible desenvolupar un gran nombre de mètodes de *clustering* basats en *deep learning*, que es poden dividir en dos grans grups: mètodes basats reduir la dimensionalitat i aplicar mètodes de *clustering* sobre les dades transformades; i mètodes que desenvolupen un model de *clustering* sobre les dades originals sense transformar [1].

En aquest treball de final de màster (TFM) s'han implementat alguns models de *deep clustering* basats en l'arquitectura *autoencoder* (AE), un tipus de xarxa neuronal profunda utilitzat com a tècnica de reducció de la dimensionalitat.

Els models resultants s'han aplicat sobre diversos conjunts de dades (principalment metabòlmiques) i s'han comparat els resultats obtinguts amb els d'algunes *clustering* clàssiques, més consolidades. S'ha valorat també el grau de solapament entre els clústers trobats pels diferents mètodes, així com la interpretabilitat biològica dels clústers resultants.

1.2 Objectius del treball

1. Implementar mètodes de *deep clustering* basats l'arquitectura AE i aplicar-los a un conjunt de dades metabolòmiques.
 - 1.1 Contextualitzar el problema a resoldre realitzant una recerca bibliogràfica.
 - 1.2 Implementar els models de *deep clustering*.
 - 1.3 Aplicar els models a un conjunt de dades metabolòmiques i estudiar els resultats.

1.3 Impacte en sostenibilitat, ètic-social i de diversitat

A continuació es descriuen els impactes i riscs que s'han detectat en el desenvolupament d'aquest TFM. Donat el perfil altament tècnic del treball, no se n'han detectat d'altres.

Riscs ètic-socials Un dels conjunts de dades que s'ha utilitzat prové d'un estudi clínic realitzats sobre pacients humans. Per tant, és imprescindible garantir el dret a la privacitat de les persones estudiades. Per aquest motiu, el proveïdor ha anonimitzat les dades abans de facilitar-les i ha demanat que es respecti la seva confidencialitat, prohibint expressament la seva publicació. A tal efecte, s'han presentat els resultats obtinguts però serà impossible possibilitar la seva replicabilitat.

Impacte ambiental L'entrenament dels models de deep learning requereix d'un elevat poder de computació, el que es tradueix en elevat consum energètic.

En el cas de models d'avantguarda molt potents, com Stable Diffusion¹, LaMDA² o DALL·E 2³, que requereixen bancs d'equips funcionant durant dies per entrenar els models amb enormes conjunts de dades, aquest consum energètic és significatiu.

No obstant, els models que s'han implementat en aquest treball requereixen d'equips significativament menys potents i només algunes hores d'entrenament. Per tant, s'ha considerat que l'impacte ambiental ha estat mínim.

Per entrenar els models s'ha utilitzat un servei de computació on-line. Només s'han considerat dues alternatives viables: Paperspace Gradient⁴ o Google Colab⁵. Entre els criteris utilitzats per seleccionar un dels dos serveis estan la utilització de font d'energia renovables, així com la possibilitat de medir el consum energètic utilitzar per l'usuari. Malauradament, no s'ha trobat cap informació respecte a cap dels dos criteris.

1.4 Enfocament i mètode seguit

Aquest TFM es divideix en dues parts ben diferenciades. A la primera part, emmarcada per l'objectiu 1 definit a la secció 1.2, s'ha consolidat una base sólida de coneixements i habilitats

¹<https://stability.ai/blog/stable-diffusion-public-release>

²<https://blog.google/technology/ai/lamda/>

³<https://openai.com/dall-e-2/>

⁴<https://docs.paperspace.com/gradient/>

⁵<https://research.google.com/colaboratory/faq.html>

específics que es s'han posat en pràctica a la segona part, que engloba els objectius 2 i 3.

S'ha realitzat una extensa recerca bibliogràfica per tal de contextualitzar el problema a resoldre. S'han caracteritzaran les tècniques de clustering utilitzades en dades metabolòmiques i com s'apliquen les tècniques de deep learning en aquest camp. S'ha estudiat l'arquitectura de diversos models basats en AE i com es poden aplicar per aconseguir models de *deep clustering*.

Finalment, s'ha estudiat el funcionament del software Keras [4] per implementar models de xarxes neuronals profundes. Per familiaritzar-se millor amb aquesta eina, s'ha practicat replicant exemples de models senzills ja desenvolupats. L'estudi s'ha basarà en el manual Deep Learning with Python [5], així com diverses fonts trobades a la web.

Les dades metabolòmiques que s'han utilitzat per avaluar els models s'han obtingut d'un recurs disponible públicament i d'un estudi clínic privat realitzat per un grup de treball de la Universitat de Barcelona associat amb el tutor d'aquest TFM.

A continuació s'ha dut a terme la part central del TFM: la implementació dels model de *deep clustering* basats en l'arquitectura AE, seguida del seu entrenament i aplicació sobre els conjunts de dades obtinguts.

Per tal d'establir un marc de referència contra el que comparar els models de *deep clustering* implementats, s'ha aplicat diverses tècniques *clustering* més consolidades (referides en aquest TFM com a tècniques de *clustering* clàssiques).

Finalment s'han evaluat diverses mètriques de rendiment per comparar els diversos mètodes. S'ha estudiat el grau de solapament entre els clústers trobats per amb diferents tècniques, i la interpretabilitat biològica dels resultats.

1.5 Planificació del treball

1.5.1 Tasques:

A continuació es llisten les tasques realitzades per la consecució dels objectius definits a la secció 1.2. El temps assignat a cada tasca es reflecteix al diagrama de Gantt de la figura 1.1.

1. 1.1 Caracteritzar les tècniques clàssiques de *clustering* aplicades a dades metabolòmiques.
- 1.2 Caracteritzar les tècniques de *clustering* basades en *deep learning*.
- 1.3 Estudiar en profunditat l'arquitectura de *variational autoencoder* i com s'aplica en *clustering*.
- 1.4 Estudiar el funcionament de Keras, practicar utilitzant exemples.
2. 2.1 Obtenir un conjunt de dades adequat per a realitzar l'estudi.
- 2.2 Realitzant un estudi preliminar amb tècniques clàssiques que servirà com a marc de referència.
- 2.3 Desenvolupar un model de *deep clustering*.
- 2.4 Comparar els resultats obtinguts amb el marc de referència.
3. 3.1 Aplicar el model a un conjunt de dades metabolòmiques.
- 3.2 Estudiar la interpretabilitat dels resultats comparant els clústers generats amb els grups reals de les dades.

1.5.2 Calendari:

A la figura 1.1 es mostra un diagrama de Gantt on es representa el temps assignat a cada tasca, així com a altres activitats relacionades amb el TFM com són la redacció dels informes de les diferents PACs, l'elaboració de la memòria i de la presentació i la preparació de la defensa.

En la planificació s'ha tingut en compte el pla docent i els dies festius. S'ha calculat una dedicació de 4 hores per dia laborable, encara que s'ha contemplat la possibilitat de dedicar temps addicional puntualment per assegurar el compliment de les fites. En total, es preveu una dedicació de 312 hores.

1.5.3 Fites:

Les fites definides com a dates clau per planificar el treball es presenten a la taula 1.1.

Data límit	Fita
17/10/2022	Elaboració del pla de treball. Entrega de l'informe de la PAC 1.
28/10/2022	Finalitzar la recerca bibliogràfica (marc teòric).
04/11/2022	Data límit per a l'obtenció d'un conjunt de dades. Consecució de l'objectiu 1.
21/11/2022	Entrega de l'informe de la PAC 2.
09/12/2022	Consecució de l'objectiu 2.
16/12/2022	Consecució de l'objectiu 3.
24/12/2022	Entrega de l'informe de la PAC 3.
15/01/2023	Entrega de la memòria i presentació.
23/01/2023	Defensa pública

Taula 1.1: Fites definides com a dates clau per planificar el treball.

1.6 Breu sumari de productes obtinguts

Com a resultat d'aquest TFM, s'han implementat diversos models de *deep learning* basats en l'arquitectura AE: *variational autoencoder* (VAE), *Deep embedded clustering* (DEC), *Variational Deep Embedding* (VaDE). Aquests models s'han compilat en un mòdul de Python, que es pot carregar fàcilment sense necessitat de re-escriure tot el codi. El mòdul s'ha fet públic a un repositori Github: https://github.com/carlescn/MSc_bioinformatics_thesis.

S'ha obtingut també una sèrie de resultats d'aplicar aquests models i altres tècniques de *clustering* sobre diversos conjunts de dades. Els resultats estan formats per diverses mètriques de rendiment així com les assignacions de les observacions dels conjunts de dades als diversos clústers els resultats per cada mètode. Aquesta informació es pot fer servir en un estudi posterior, i també s'ha fet pública al mateix repositori.

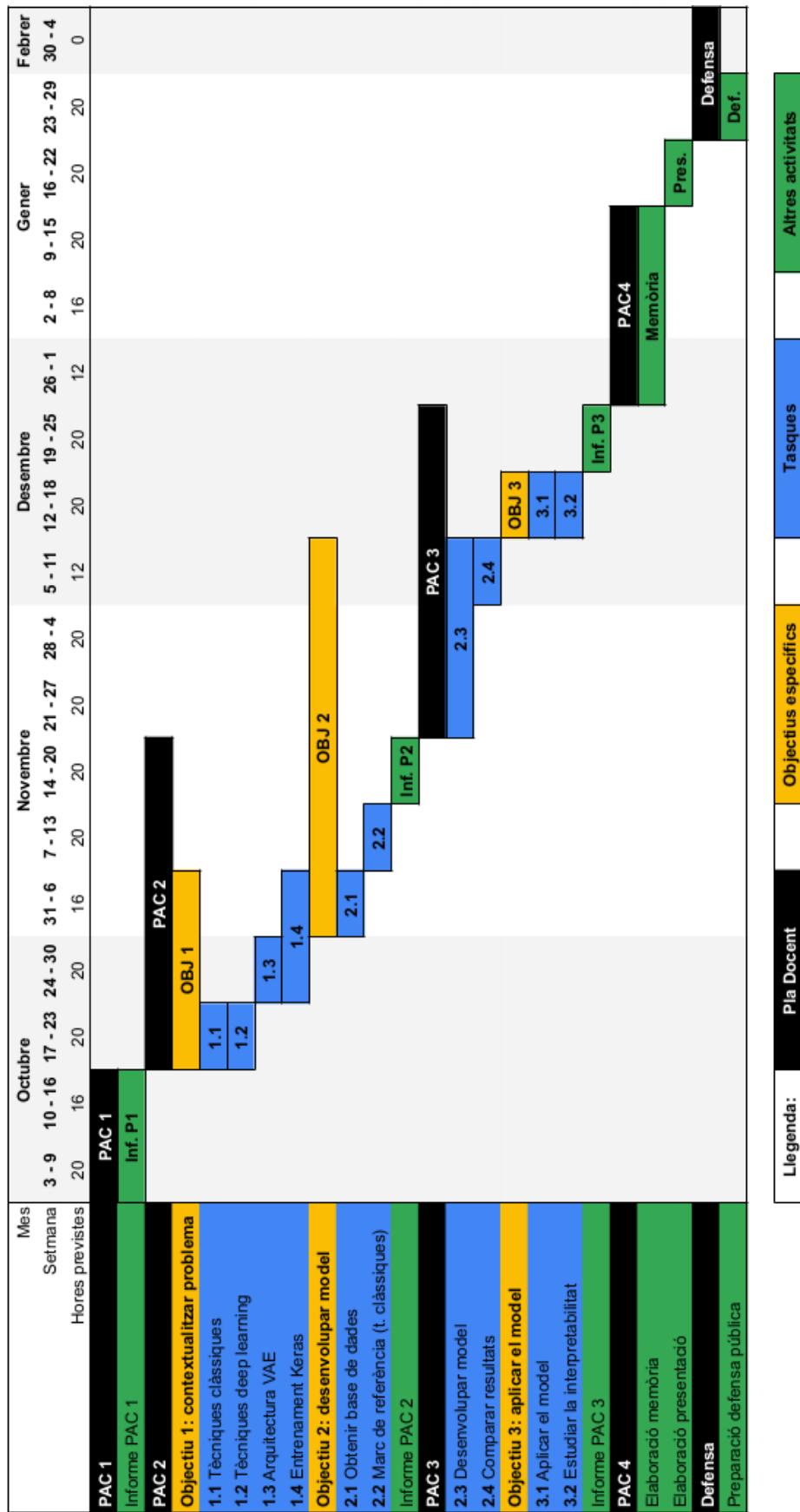


Figura 1.1: Diagrama de Gantt: planificació dels objectius i tasques.

1.7 Breu descripció dels altres capítols de la memòria

Capítol 2 Estat de l'art En aquest primer capítol s'introdueixen les tècniques de *clustering* clàssiques, el concepte d'aprenentatge de característiques (*feature learning*) (FL), i les avantatges que poden suposar els models basats en xarxes neuronals profundes.

Capítol 3 Materials i mètodes Aquí es presenta la base teòrica que sosté els models de *deep clustering* escollits, així com algunes particularitats de la seva implementació. A continuació es detallen les mètriques utilitzades per avaluar el seu rendiment.

Tot seguit es presenten els conjunts de dades utilitzats i s'explica detingudament la metodologia seguida per avaluar i comparar els diversos models.

Finalment, es mencionen breument les eines informàtiques utilitzades per implementar els models de *deep clustering* i avaluar els resultats.

Capítol 4 Resultats En aquest capítol es presenten els resultats obtinguts i se'n fa un breu resum. La totalitat dels resultats es presenten en taules i figures als corresponents apèndixs.

Capítol 5 Discussió Seguidament es discuteixen els resultats. Es discuteix els avantatges i inconvenients dels models implementats i es fa una breu valoració de la interpretabilitat del clústers obtinguts.

Capítol 6 Valoració econòmica En aquest capítol es resumeix breument el petit cost econòmic que ha suposat el desenvolupament d'aquest TFM.

Capítol 7 Conclusions i treballs futurs Finalment, es presenta de manera sintetitzada les conclusions que s'han tret d'aquest TFM i es fa una sèrie de propostes per futurs treballs, que no ha donat temps d'estudiar en el temps disponible.

Capítol 2

Estat de l'art

2.1 Tècniques de *clustering*

Les tècniques de *clustering* són un conjunt de tècniques estadístiques o d'aprenentatge automàtic no supervisat que es basen en agrupar les observacions en funció d'alguna mesura de la seva similitud, sense coneixement previ de l'estructura de les observacions o el nombre de grups que es pretén obtenir.

L'objectiu d'aquestes tècniques és separar les dades en diversos grups que siguin internament homogenis (les observacions dins el mateix grup són similars entre ells) i tinguin característiques diferents a la resta de grups.

Identificar aquests grups permet obtenir informació sobre l'estructura de les dades com la presència de patrons, i pot servir com a un punt de partida en l'exploració de les dades. Per exemple, aplicades a dades metabolòmiques, les tècniques de *clustering* poden permetre identificar diferents tipus cel·lulars en funció de la seva expressió metabòlica [6].

Existeix una gran varietat de tècniques de *clustering* [1–3, 6]. Encara que totes comparteixen un mateix objectiu, aborden el problema aplicant diferents criteris i per tant els grups que formen poden no coincidir. Algunes de les aproximacions més utilitzades històricament es basen en mètodes matemàtics i estadístics [2], com mètodes jeràrquics, mètodes basats en centroides, distribucions o densitats [1].

Mètodes jeràrquics: es basen en crear clústers amb un ordre predeterminat, on els clústers de més baix nivell es combinen iterativament per crear clústers més grans. Això els dota d'una estructura jeràrquica que es pot presentar en forma de dendrograma, però implica que l'assignació de cada punt a un clúster és determinista. Aquests algoritmes no requereixen fixar prèviament un número de clústers, però són sensibles al soroll [1].

Mètodes basats en centroides: (p. ex. *k-means*) es basen assignar cada punt a un número predeterminat de grups i calcular els seus centroides. Després es canvia iterativament l'assignació de cada punt als diferents grups fins a minimitzar la suma de les distàncies de cada punt al centroïde del seu grup. En general aconsegueixen un millor rendiment que els mètodes jeràrquics, però són incapços de trobar grups no convexos [1].

Mètodes basats en distribucions: (p. ex. model de barreja gaussiana (GMM)) es basen en modelar els grups en funció d'una barreja de distribucions probabilístiques. La seva base estadística permet inferir relacions entre les característiques de les dades, però requereix de fortes assumpcions sobre la distribució de les dades i tenen tendència a sobre-ajustament [1].

Mètodes basats en densitats: (p. ex. DBSCAN) defineixen els clústers com àrees amb major densitat comparats amb la resta de les dades. Els punts en regions més disperses es consideren fronteres o soroll. El punt negatiu d'aquests mètodes és precisament que requereixen d'una disminució de la densitat per detectar les fronteres dels clústers i són poc eficaces en separar grups contigus [1].

2.2 Reducció de la dimensionalitat

Un dels problemes que apareix habitualment al analitzar dades bioinformàtiques és la seva elevada dimensionalitat, característica que fa que les tècniques de *clustering* clàssiques no funcionin bé [1, 2]. Per combatre aquest problema, una solució és aplicar tècniques de reducció de la dimensionalitat i posteriorment aplicar les tècniques de *clustering* sobre les dades transformades [2, 3].

La reducció de la dimensionalitat s'aconsegueix mitjançant tècniques d'FL. L'objectiu d'aquestes tècniques és representar les dades originals en un espai dimensional inferior, aplicant una transformació que retengui el màxim d'informació. L'espai reduït resultant s'anomena capa de representació (CR) i les seves dimensions característiques apreses (*learned features*) (LF).

Històricament s'han utilitzat tècniques matemàtiques per realitzar FL, que poden ser transformacions lineals (anàlisi de components principals) o no lineals (mètodes kernel o tècniques espectrals). Posteriorment s'apliquen les tècniques de *clustering* sobre les LF.

Una limitació dels mètodes lineals és que no són capaços de retenir informació sobre relacions no lineals en les dades, el que provoca una reducció de la qualitat de *clustering* (QC). Les tècniques no lineals són més adequades [1].

2.3 Tècniques de *deep clustering*

Les xarxes neuronals profunes possibiliten aplicar tècniques de FL més eficients que les descrites a la secció anterior, aplicant mètodes no lineals complexes que permeten capturar LF més rellevants. En particular, la funció de transformació es pot optimitzar mitjançant l'aprenentatge dels paràmetres de la xarxa, el que permet extreure LF òptimes per obtenir una bona QC [1].

La seva recent popularització ha fet possible desenvolupar un gran nombre de tècniques de *clustering* basades en *deep learning*, anomenades en conjunt tècniques de *deep clustering*. Es poden dividir en dos grans grups: mètodes de dos passos, que es basen en utilitzar tècniques de FL i aplicar mètodes de *clustering* convencionals sobre les LF; i mètodes d'un sol pas que desenvolupen un model de *clustering* sobre les dades originals [1].

A la literatura s'ha descrit una gran varietat de mètodes basats en diferents arquitectures neuronals (*multilayer perceptron* (MLP), *convolutional neural network* (CNN), *deep belief network* (DBN), *generative adversarial network* (GAN), AE). A <https://github.com/rezacsedu/Deep-learning-for-clustering-in-bioinformatics> [1] es pot trobar un llistat amb enllaços als articles originals.

La majoria d'aproximacions actuals utilitzen una arquitectura basada en un AE [1], donat que és capaç d'obtenir LF eficients per realitzar *clustering*.

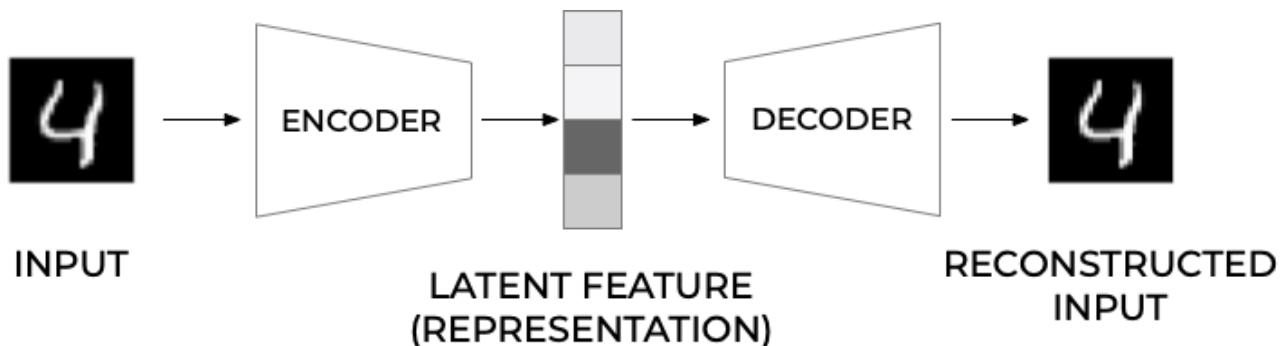


Figura 2.1: Esquema del model AE. Font: Michelucci, 2022 [7]

2.4 Autoencoder

L’arquitectura AE està formada per dues parts: un codificador i un descodificador. La funció del codificador és representar l’entrada en un espai dimensional reduït, sovint un vector, anomenat codi. La funció del descodificador és reconstruir l’entrada original a partir del codi. A la figura 2.1 es mostra esquema de l’arquitectura AE.

Optimitzant els paràmetres d’ambdues parts conjuntament durant l’entrenament de la xarxa es busca aconseguir que el codificador sigui capaç de comprimir la entrada mantenint el màxim d’informació. Aplicant la idea de FL, el codi esdevé la CR.

Per aconseguir un *clustering* eficaç, no és suficient amb aquesta optimització ja que si no s’aplica cap restricció a la CR podria donar lloc a LF que no permeten una bona QC. Per resoldre aquest problema es defineixen dos tipus de funció de cost, que s’optimitzen conjuntament durant l’entrenament [3]:

- La **funció de cost auxiliar**, que depèn de la capacitat del codificador d’obtenir una CR eficient que permeti que el descodificador sigui capaç de regenerar l’entrada. Garanteix obtenir LF rellevants.
- La **funció de cost de clustering**, que depèn de l’algoritme de *clustering* i de que les LF obtingudes siguin adequades per realitzar *clustering*. Assegura obtenir una bona QC.

L’aproximació més bàsica de l’arquitectura AE es pot construir utilitzant MLPs simètrics pel codificador i descodificador. Encara que és fàcil d’implementar, el model generat conté un gran nombre d’hiperparàmetres i esdevé difícil d’optimitzar i balancejar les dues funcions de cost [1]. A més, no s’aplica cap restricció a la funció de representació i per tant no garanteix obtenir una bona QC.

A la literatura s’han descrit diferents aproximacions a la idea del AE, que prenen com a base altres arquitectures. Algunes d’aquestes aproximacions permeten aconseguir un model generatiu, que és capaç de generar dades similars a les d’entrenament. Alguns d’aquestes aproximacions són:

Convolutional autoencoder (CAE): l’arquitectura AE estàndard no dona bons resultats per trobar patrons en imatges (dades amb invariança espacial). Per combatre-ho, es poden combinar amb CNN utilitzant convolucions al codificador del AE (i les corresponents desconvolucions al descodificador), on els filtres de la CNN són un paràmetre que es pot optimitzar (en lloc de construir-los a mà com en una CNN estàndard).

VAE: aquesta aproximació es basa en mètodes Bayesians, el que li atorga certa robustesa estadística que no tenen altres arquitectures. Es tracten les dades com a mostres d'una distribució desconeguda i s'aplica una restricció al codificador perquè les representi com una barreja de distribucions conegeudes (per exemple, gaussianes). Els paràmetres que aprèn la xarxa són els paràmetres de les distribucions (mitjana i variància). El descodificador reconstrueix les mostres originals, el que permet generar mostres aleatòries semblants a la distribució original desconeguda. S'explica amb més detall a la secció 3.2.2.

Long short-term memory autoencoder (LSTM-AE): de manera similar al cas del CAE, l'arquitectura AE estàndard no funciona bé amb dades seqüencials, però es pot combinar amb l'arquitectura *long short-term memory* (LSTM). El codificador es forma combinant capes LSTM, on la última capa codifica un vector que representa l'entrada de la xarxa. El descodificador, també format per capes LSTM, pren aquest vector (replicat per adaptar-lo a l'entrada de la primera capa LSTM) i reconstrueix l'entrada original.

Adversarial autoencoder (AAE): aquesta aproximació pren la idea del discriminador de la arquitectura GAN i la introduceix al AE. L'objectiu és representar les dades originals com una barreja de distribucions conegeudes, de manera similar al VAE. En aquest cas la xarxa està formada per tres parts: el codificador, que redueix l'entrada a un codi, el descodificador, que reconstrueix la entrada a partir del codi, i un discriminador. El discriminador intenta discriminar entre els codis generats i punts mostrejats aleatoriament de la distribució escollida. El resultat és que els codis generats acaben aproximant-se a aquesta distribució [8].

Denoising autoencoder (DAE): aquesta tècnica entraïnament soroll a l'entrada del codificador i reconstruyeix les dades originals amb el descodificador. Això permet aprendre una representació més robusta de les dades i també la reconstrucció de les dades originals a partir d'unes dades parcialment corrompudes.

Stacked autoencoder (SAE): diversos AE es poden apilar per aconseguir representacions més comprimides de les dades d'entrada. S'entrena un primer AE per comprimir i reconstruir les dades d'entrada i es pren el codi comprimit. Aquest codi es pren com l'entrada d'un segon AE, que s'entrena per comprimir i reconstruir el codi. Aquest procés es repeteix fins a aconseguir el nombre de capes desitjat. Finalment, s'ordenen les capes codificadores i descodificadores de cada AE seqüencialment, formant un SAE, i s'entrena una última vegada el model sencer per ajustar tots els pesos. D'aquesta manera s'obté un model capaç de codificar les dades en un espai dimensional més reduït.

En aquest TFM, es desenvoluparà un model de *deep clustering* utilitzant una arquitectura VAE, ja que les seves característiques el fan un model interessant. El model probabilístic li atorga certa robustesa estadística que pot ser útil al extreure conclusions del resultats obtinguts. Alhora, la capacitat generativa pot permetre generar mostres artificials semblants a les originals, el que podria servir com a tècnica d'augment de dades (*data augmentation*).

Capítol 3

Materials i mètodes

3.1 Tècniques de *clustering* clàssiques

A continuació s'expliquen de manera molt sintetizada les tres tècniques de *clustering* clàssiques utilitzades com a referència per comparar amb les tècniques de *deep clustering*.

K-means El mètode K-means és un algoritme de *clustering* que minimitza les distàncies entre cada objecte al centroïde del seu clúster. La distància es sol calcular utilitzant la suma de quadrats de l'error (SSE).

L'algoritme consta de tres passos: s'escull un número k de clústers i k punts en l'espai, que seran els centroïdes inicials. S'assigna cada punt al clúster que minimitza la distància al seu centroïde. S'actualitza la posició dels centroïdes calculant la posició mitjana dels punts associats a cada clúster. Finalment es repeteixen els dos últims passos fins que els centroïdes deixen de moure's.

La principal avantatge d'aquest mètode és que és un mètode senzill i ràpid, però requereix escollir un valor de k arbitrari i no garanteix convergir en un mínim global.

Model de barreja de gaussianes (GMM) El model GMM és un model probabilístic que assumeix que les observacions s'han generat d'una barreja finita de distribucions normals (gaussianes).

Cada distribució està caracteritzada per una mitjana una covariància i una probabilitat π , que defineix la mida de la distribució. La suma de les probabilitats π de totes les distribucions ha de resultar sempre 1.

Així, fixant un nombre de clústers k i utilitzant mètodes Bayesians es poden calcular els paràmetres de les distribucions que maximitzen la probabilitat de les mostres observades. No es considera necessari entrar en detall dels càlculs matemàtics.

Una característica d'aquest model és que no assigna un clúster determinat a cada observació, si no una probabilitat de pertànyer a cada clúster.

El model resultant és un model generatiu: coneixent els paràmetres de les distribucions, es poden calcular mostrejar nous punts que haurien de distribuir-se com les variables estudiada.

Igual que el mètode K-means, presenta l'inconvenient que és necessari escollir prèviament un valor de k arbitrari.

Model aglomeratiu (Aglo.) L'algoritme de *clustering* aglomeratiu es basa en construir una jerarquia de grups en funció de la seva similaritat.

El procés s'inicia assignant un grup unitari a cada observació. A continuació es calcula la distància entre cada parell de grups i s'uneixen els dos que minimitzen aquest valor. Aquest procés es repeteix fins que només queda un grup global.

Per calcular la distància entre punts es poden utilitzar diferents mètriques, com ara la distància Euclidiana, la distància de Manhattan, etc. La distància entre grups també es pot calcular segons diferents criteris: agrupament màxim (la distància entre dos grups es determina per la distància màxima entre dos punts), agrupament mínim (es determina per la mínima distància entre dos punts), etc.

Aquest mètode presenta l'avantatge de no requerir seleccionar un número de clústers arbitrari. A més, el conjunts resultants es poden representar en un dendrograma que representa gràficament la similitud entre els punts i entre grups.

Un inconvenient és que són sensibles al soroll: una petita variació en les dades pot desembocar en grups completament diferents.

3.2 Models de *deep clustering*

En aquest TFM es pretén implementar diversos model de *deep clustering* en dades metabòliques. Com s'ha descrit a les [seccions 2.3 i 2.4](#), existeix una gran varietat d'arquitectures de xarxes neuronals descrites a la literatura. Intentar estudiar totes les possibilitats seria inabastable donat el temps disponible, pel que ha sigut necessari acotar l'estudi a una única arquitectura. S'ha decidit implementar algunes aproximacions basada en AE.

A continuació es fa una breu presentació dels models DEC [9], VAE [10, 11] i VaDE [12] basada en les referències indicades al costat de cada model. No s'entrarà en profunditat en les explicacions matemàtiques. Per una explicació més detallada es recomana llegir les fonts originals.

3.2.1 Deep embedded clustering

El model DEC es basa en un AE al que, després de la fase d'entrenament habitual, s'elimina el descodificador i es substitueix per una nova xarxa neuronal, la sortida de la qual es pot tractar com un vector de probabilitats d'assignacions a clústers, com s'il·lustra a la [figura 3.1](#).

Primerament, s'entrena un model AE aplicant únicament una **funció de cost de reconstrucció**, que busca minimitzar la distància entre la sortida del descodificador i l'entrada original del codificador. Poden utilitzar-se diferents funcions, com ara la SSE, la mitjana de quadrats de l'error (MSE) o la entropia creuada. A la [secció 3.2.2](#) s'entra en més detall sobre la funció escollida i com es calcula.

Un cop entrenat, s'obté un model amb un codificador capaç de representar les dades d'entrada \mathbf{x} en un espai latent \mathbf{z} , i un descodificador capaç de generar mostres simulades a partir de \mathbf{z} . És a dir, el model ha après unes LF rellevants de les dades que permeten obtenir una CR eficient.

El següent pas és aplicar una tècnica de *clustering* sobre les representacions de l'espai latent \mathbf{z} . Per aconseguir-ho, es construeix un model de *clustering* basat en una nova capa neuronal.

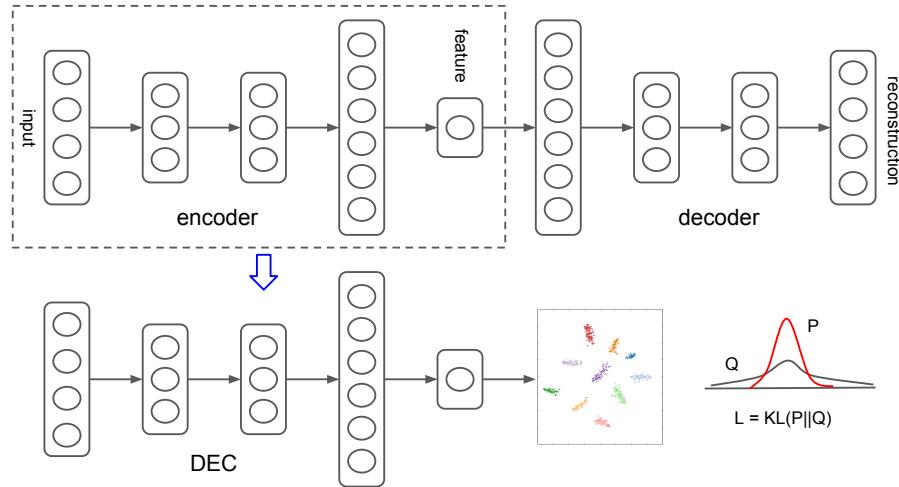


Figura 3.1: Esquema de l'arquitectura del model DEC. Font: Xie *et. al.*, 2015 [9]

Els pesos de cada neurona d'aquesta capa corresponen a la posició del centroide d'un clúster. Una limitació d'aquesta tècnica és que requereix establir el nombre k de clústers prèviament.

La capa pren com entrada un punt de \mathbf{z} i calcula la distància del punt a cada centroide. La sortida de la capa és un vector de longitud k que suma 1, on cada valor representa la probabilitat d'assignació del punt a cada clúster. Aquesta assignació s'interpreta com una distribució Q .

A continuació es defineix una nova funció de cost, la **funció de cost de clustering**, com la distància Kullback-Leibler (KL) entre la distribució Q i una distribució auxiliar P . Escollint una distribució auxiliar adequada, al minimitzar aquesta distància s'aconsegueix millorar la QC: es minimitza la distància de cada punt al seu centroide alhora que es maximitza la distància a la resta de centroides.

La distribució auxiliar P es calcula a partir de Q :

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_j q_{ij}^2 / f'_j} \quad (3.1)$$

on p_{ij} i q_{ij} són respectivament la probabilitat de P i Q per la observació i i el clúster j , i f_j és la freqüència del clúster j .

El model DEC es construeix prenent només el codificador del AE entrenat prèviament, i afegint la capa de *clustering*. Abans d'iniciar l'entrenament del nou model, és necessari establir uns paràmetres inicials (els pesos i biaixos de les neurones de la capa de *clustering*).

Es passen les dades d'entrenament pel codificador per obtenir les seves representacions en l'espai \mathbf{z} . Seguidament s'aplica un mètode de *clustering* (l'autor proposa *K-means*) i es pren la posició dels centroides.

Finalment, s'entrena la xarxa utilitzant el descens de gradients estocàstic optimitzant únicament per la funció de cost de *clustering*. El model aprendrà una nova representació en l'espai latent \mathbf{z} , juntament amb una posició òptima dels centroides que defineixen cada clúster, que en conjunt optimitzin la QC. És important mencionar que al descartar el descodificador i no utilitzar la funció de cost de reconstrucció, no es pot assegurar que el model mantingui les LF apreses.

Alternativament, es pot construir un model alternatiu que sí manté el descodificador [3] (figura 3.2). En aquest cas el model s'entrena aplicant conjuntament les dues funcions de cost

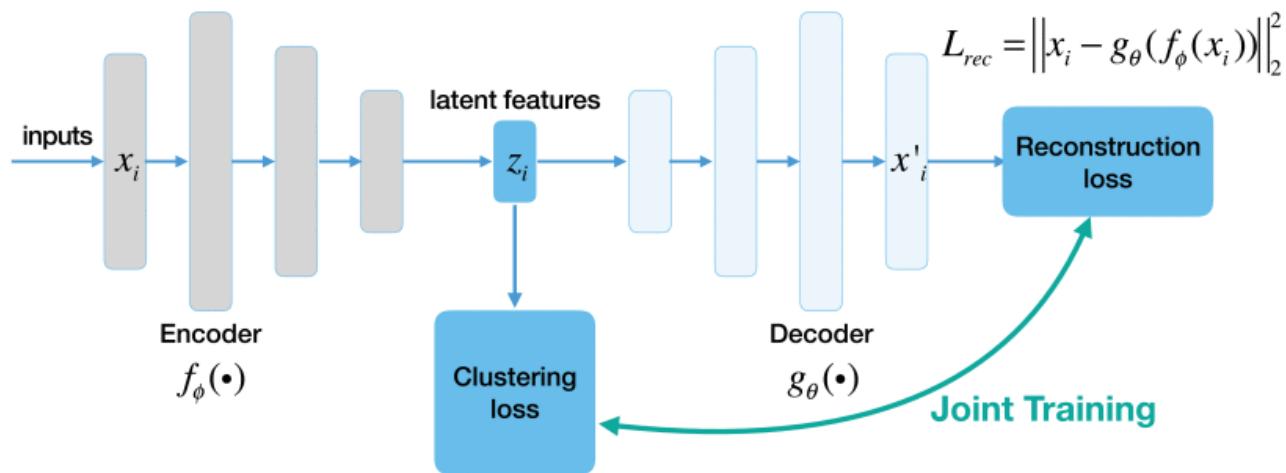


Figura 3.2: Esquema del model DEC alternatiu, que manté el descodificador i s'entrena aplicant conjuntament les funcions de cost de reconstrucció i de *clustering*. Font: Min et. al., 2018 [3]

(de reconstrucció i de *clustering*). D'aquesta manera s'assegura mantenir mantenir unes LF rellevants a la CR.

3.2.2 Variational Autoencoder

Com s'ha mencionat a la secció 2.4, l'arquitectura VAE és una implementació específica de AE basada en mètodes Bayesians. Això li atorga al model certa robustesa estadística que no trobem en altres arquitectures. A més, la CR apresa permet la generació de mostres artificials similars a les dades d'entrenament. Això últim pot resultar útil com a tècnica d'augment de dades.

En l'arquitectura AE, el codificador aprèn una representació de l'entrada \mathbf{x} en un espai dimensional reduït \mathbf{z} aplicant una transformació no lineal. Des de la perspectiva de FL, aquest espai reduït es correspon amb la CR. El descodificador aprèn a reconstruir \mathbf{x} a partir de \mathbf{z} aplicant una segona transformació no lineal. L'entrenament de les dues parts conjuntament aconsegueix que les LF siguin una bona representació de les dades originals.

En l'aproximació que pren el VAE, s'assumeix que l'entrada \mathbf{x} és un conjunt d'observacions de variables aleatòries, que provenen d'un procés desconegut amb una distribució $p^*(\mathbf{x})$. Aquesta distribució és desconeguda, de manera que intentem aproximar-la amb un model $p_\theta(\mathbf{x})$, on θ són els paràmetres del model.

L'espai \mathbf{z} es considera un conjunt de variables latents del model que segueixen una distribució $p_\theta(\mathbf{z})$. Les variables latents són variables que estan presents al model però no observem en les dades. Es pot fixar que $p_\theta(\mathbf{z})$ sigui una distribució fàcil de computar, per exemple una barreja de gaussianes.

Sota aquesta premissa, el model a inferir és sobre una distribució conjunta $p_\theta(\mathbf{x}, \mathbf{z})$ de les variables observades i latents. L'objectiu del codificador és aprendre els paràmetres θ de la distribució $p_\theta(\mathbf{z}|\mathbf{x})$ (la distribució posterior de \mathbf{z} condicionada a \mathbf{x}). De la mateixa manera, l'objectiu del descodificador és aprendre els paràmetres θ de la distribució $p_\theta(\mathbf{x}|\mathbf{z})$. A la figura 3.3 s'il·lustra com el model relaciona els espais \mathbf{x} i \mathbf{z} .

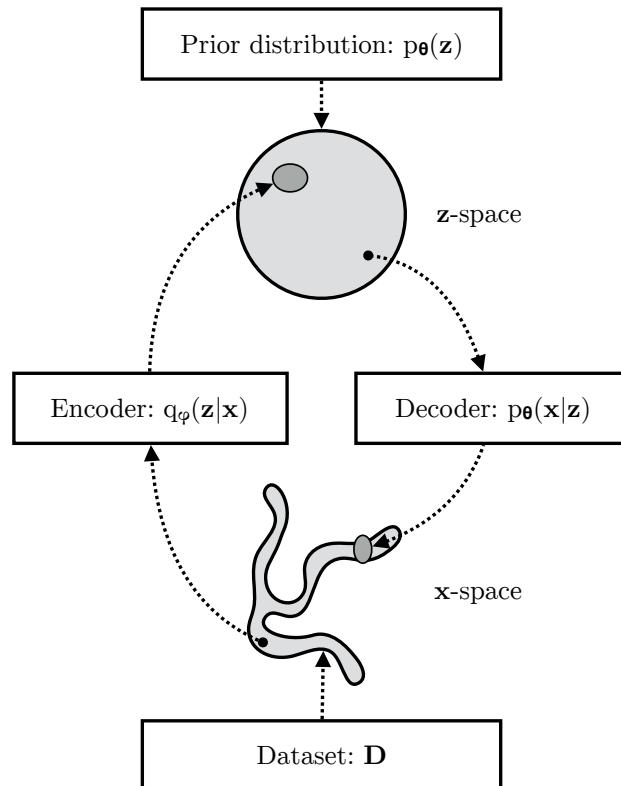


Figura 3.3: Esquema del model VAE. Les dades observades \mathbf{D} (*dataset*) són una mostra d'unes variables de l'espai \mathbf{x} , que segueix una distribució intractable. L'espai latent \mathbf{z} s'escull que segueixi una distribució $p_\phi(\mathbf{z})$ fàcil de computar. El codificador (*encoder*) troba la distribució condicional de \mathbf{z} donada \mathbf{x} , aconseguint una representació de les mostres a l'espai latent. El decodificador (*decoder*) troba la distribució condicional de \mathbf{x} donada \mathbf{z} , aconseguint la reconstrucció de les dades a partir de l'espai latent. Font: Kingma and Welling, 2019 [11]

3.2.3 Optimització dels paràmetres

El codificador pretén aprendre els paràmetres θ òptims de la distribució $p_\theta(\mathbf{z}|\mathbf{x})$, que està condicionada per la distribució marginal $p_\theta(\mathbf{x})$ segons la equació:

$$p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{x})} \quad (3.2)$$

Aquesta distribució marginal en el model bé donada per:

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \quad (3.3)$$

La integral en la equació (3.3) provoca que la distribució marginal $p_\theta(\mathbf{x})$ no tingui un estimador eficient i sigui intractable computacionalment. Això provoca que no es pugui diferenciar respecte als paràmetres θ i optimitzar-la pel mètode del descens de gradients. En conseqüència, tampoc es poden optimitzar els paràmetres de la distribució $p_\theta(\mathbf{z}|\mathbf{x})$.

En el seu lloc, es pot inferir un model $q_\phi(\mathbf{z}|\mathbf{x})$ que aproximi la distribució intractable $p_\theta(\mathbf{z}|\mathbf{x})$, on els paràmetres ϕ s'aprenen a través de la optimització dels pesos i biaixos de la xarxa neuronal.

La diferència entre aquestes dues distribucions es pot calcular mitjançant la divergència de KL: $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x}))$. L'objectiu és minimitzar aquesta divergència, però no és possible calcular-la directament donada la intractabilitat de $p_\theta(\mathbf{z}|\mathbf{x})$.

Aquesta divergència es relaciona amb la distribució marginal segons la següent igualtat:

$$\log p_\theta(\mathbf{x}) = \mathcal{L}_{\theta,\phi}(\mathbf{x}) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})) \quad (3.4)$$

Per tant, es pot minimitzar la divergència maximitzant el terme $\mathcal{L}_{\theta,\phi}(\mathbf{x})$. Aquest terme s'anomena límit inferior de la evidència (*evidence lower bound*) (ELBO) ja que la divergència KL és sempre no negativa i per tant ELBO és el límit inferior de $\log p_\theta(\mathbf{x})$ (evidència de \mathbf{x}):

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{\theta,\phi}(\mathbf{x}) \quad (3.5)$$

El terme ELBO es calcula segons la següent equació:

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z})} \right] \quad (3.6)$$

La maximització del ELBO implica:

- Maximitzar el primer terme de la equació, la esperança de $\log p_\theta(\mathbf{x}|\mathbf{z})$, que es correspon amb optimitzar els paràmetres del descodificador. És a dir, s'aconsegueix una millor reconstrucció.
- Minimitzar el segon terme, la esperança de $\log(q_\phi(\mathbf{z}|\mathbf{x})/p_\theta(\mathbf{z}))$, que és la divergència KL entre la distribució inferida pel codificador $q_\phi(\mathbf{z}|\mathbf{x})$ i la distribució $p_\theta(\mathbf{z})$ fixada de l'espai latent. És a dir, s'aconsegueix millorar la representació obtinguda pel codificador.

3.2.4 Descens de gradients

Per optimitzar els paràmetres de la xarxa neuronal de tal manera que es maximitzi el ELBO, s'utilitza el descens de gradients. Per tant, cal calcular els gradients del ELBO amb respecte als paràmetres θ i ϕ .

El ELBO es pot reformular de la següent manera, el que facilitarà el càlcul dels gradients:

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})] \quad (3.7)$$

I els gradients es calculen segons:

$$\nabla_{\theta} \mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_{\theta} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})] \quad (3.8)$$

$$\nabla_{\phi} \mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})] \quad (3.9)$$

Per la regla de la integral de Leibniz, el gradient en la [equació \(3.8\)](#) es pot moure a dins de la esperança i això permet la seva computació. Però el mateix no és possible en la [equació \(3.9\)](#), ja que l'esperança està en funció de ϕ .

Per resoldre aquest problema es planteja la següent solució: reparametritzar la variable $\mathbf{z} \sim q_{\theta}(\mathbf{z}|\mathbf{x})$ com una funció d'una altra variable aleatòria ϵ , donades \mathbf{x} i ϕ :

$$\epsilon \sim p(\epsilon) \quad (3.10)$$

$$\mathbf{z} = g(\phi, \mathbf{x}, \epsilon) \quad (3.11)$$

on la funció $g(\cdot)$ és una transformació determinista. Aquesta reparametrització s'il·lustra a la [figura 3.4](#). Llavors, el ELBO es pot reescriure en funció de $\mathbb{E}_{p(\epsilon)}$ en lloc de $\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}$, de manera que “s'externalitza” la aleatorietat de la variable \mathbf{z} a la nova variable ϵ . La nova forma del gradient del ELBO en respecte a ϕ és:

$$\nabla_{\phi} \mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_{\phi} \mathbb{E}_{p(\epsilon)}[\log p_{\theta}(\mathbf{x}, \mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})] \quad (3.12)$$

Després d'aquesta reparametrització, es poden calcular els gradients del ELBO respecte als paràmetres ϕ i θ , i així optimitzar-los mitjançant el descens de gradients.

3.2.5 Implementació del model

Un cop presentada la base teòrica, a continuació es resumeix com s'ha implementat el model VAE que s'ha construït en aquest TFM. S'ha escollit que la variable latent \mathbf{z} segueixi una distribució normal multivariant:

$$\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$$

on μ és el vector de mitjanes de les distribucions i σ^2 és la matriu de variàncies. Per simplificar els càlculs s'ha escollit una matriu diagonal, de mode que σ^2 es pot escriure com un vector.

Així, el codificador i el descodificador del VAE són dues xarxes neuronals amb paràmetres θ i ϕ , respectivament. Aquests paràmetres són l'arquitectura de la xarxa neuronal, juntament amb els pesos i biaixos de cada neurona.

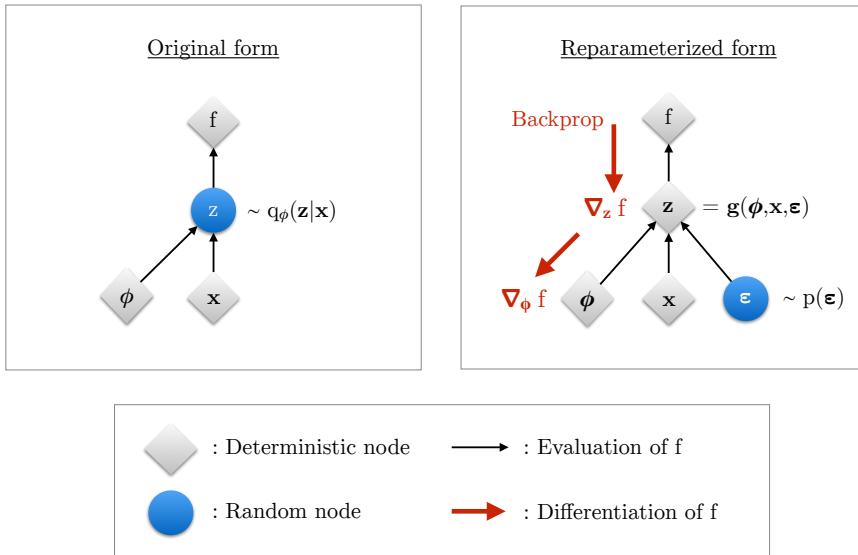


Figura 3.4: Il·lustració de la reparametrització de \mathbf{z} . Per optimitzar els paràmetres θ es necessita calcular els gradients de la funció d'optimització f amb respecte a θ . En la figura de l'esquerra (forma original) no es pot diferenciar f amb respecte a θ perquè no es poden propagar endarrere els gradients a través de la variable aleatoria \mathbf{z} . Després de la reparametrització (figura de la dreta), la aleatorietat queda “externalitzada” a la variable ϵ , i \mathbf{z} depèn d'una funció $g(\cdot)$ determinista. Aquesta configuració sí permet propagar endarrere els gradients. Font: Kingma and Welling, 2019 [11]

El codificador pren com entrada els valors de les variables observades, en forma d'un vector \mathbf{x} per cada observació, i té com a sortida dos vectors: un vector μ amb les mitjanes de la barreja de gaussianes, i un vector σ^2 amb les corresponents variàncies (codificades com a $\log \sigma^2$ per motius de computació).

El descodificador pren com entrada un punt de l'espai latent \mathbf{z} i té com a sortida un vector amb la mateixa forma que \mathbf{x} .

En l'arquitectura AE, durant la fase d'entrenament es connecta la sortida del codificador a l'entrada del descodificador. Així, les dues xarxes s'entrenen conjuntament utilitzant el descens de gradient estocàstic. En el cas de l'arquitectura VAE, és necessari aplicar la reparametrització explicada en la secció 3.2.4. Un cop aplicada, es pot reescriure \mathbf{z} com:

$$\epsilon \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathbf{z} = \mu + \sigma \odot \epsilon$$

on \odot és el producte per elements.

En la implementació del VAE, s'insereix una capa neuronal sense pesos o biaixos entre el codificador i el descodificador. Aquesta capa pren com entrada la sortida del codificador (vectors μ i σ^2), obté una mostra aleatòria de la distribució de ϵ i calcula un valor de \mathbf{z} , que passa a l'entrada del descodificador. A la [figura 3.5](#) es mostren esquemàticament les diferents parts del VAE.

Per entrenar la xarxa neuronal en conjunt, s'han definit dues funcions de cost que corresponen als dos termes del ELBO, segons la [equació \(3.6\)](#):

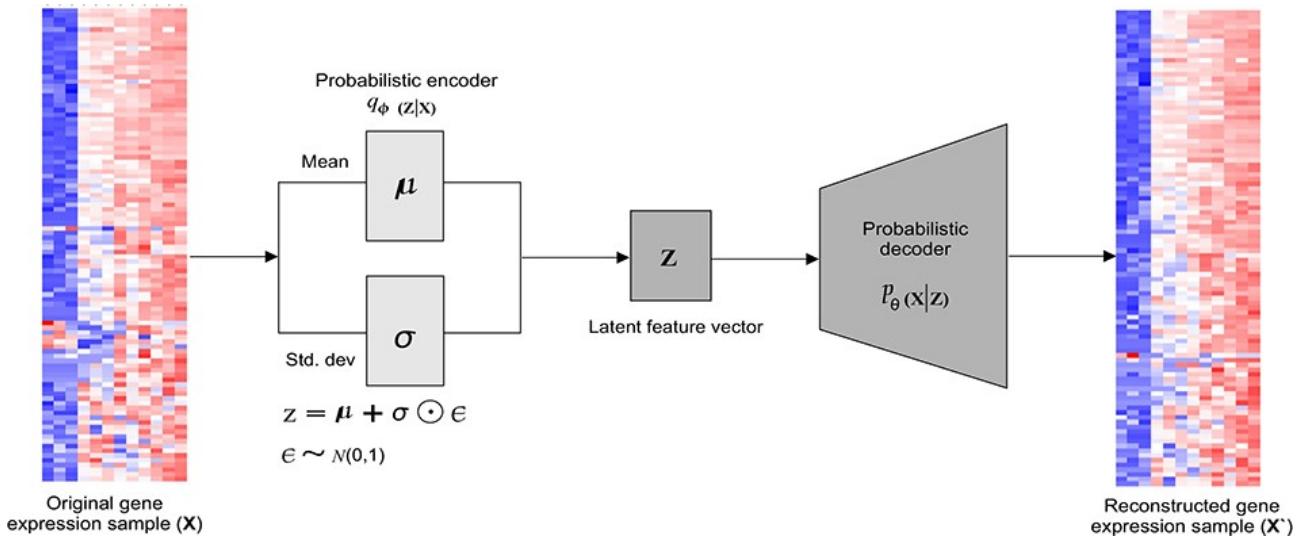


Figura 3.5: Representació esquemàtica d'una arquitectura VAE. El codificador aprèn a representar l'entrada \mathbf{x} en un espai latent \mathbf{z} que segueix una distribució normal multivariant, amb un vector de mitjanes μ i un vector de variàncies σ . S'utilitza la reparametrització de \mathbf{z} en funció de ϵ . El descodificador reconstrueix l'entrada a partir de la representació a l'espai latent. Font: Karim *et. al.*, 2021 [1]

- **Funció de cost de la reconstrucció:** computa la diferència entre l'entrada del codificador i la sortida del descodificador (la reconstrucció de l'entrada). Minimitzant-la, es maximitza el terme primer terme del ELBO: $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$.
- **Funció de cost de regularització:** computa la distància KL entre la distribució de $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ i una distribució $\mathcal{N}(0, \mathbf{I})$. Minimitzant-la, es minimitza el segon terme del ELBO: $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(q_\phi(\mathbf{z}|\mathbf{x})/p_\theta(\mathbf{z}))]$. S'anomena així ja que actua com a regularitzador de la sortida del codificador, assegurant que l'espai latent segueixi la distribució desitjada.

3.2.6 Variational Deep Embedding

El model VaDE generalitza el model VAE utilitzant un model GMM en lloc d'una normal multivariant, com s'il·lustra a la figura 3.6. Això fa que sigui un model molt més apropiat per realitzar *clustering*, i manté la propietat generadora del VAE.

Així, la principal diferència és que la CR està formada per π distribucions normals multivarians. El codificador aproxima una distribució conjunta de \mathbf{x} , \mathbf{z} i c (el clúster):

$$p(\mathbf{x}, \mathbf{z}, c) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|c)p(c) \quad (3.13)$$

on les probabilitats són:

$$p(c) = Cat(c|\pi) \quad (3.14)$$

$$p(\mathbf{z}|c) = \mathcal{N}(\mu_c, \sigma_c^2 \mathbf{I}) \quad (3.15)$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mu_x, \sigma_x^2 \mathbf{I}) \quad (3.16)$$

on $Cat(c|\pi)$ és la distribució dels clústers i μ_c i σ_c^2 són la mitjana i variància corresponents al clúster c .

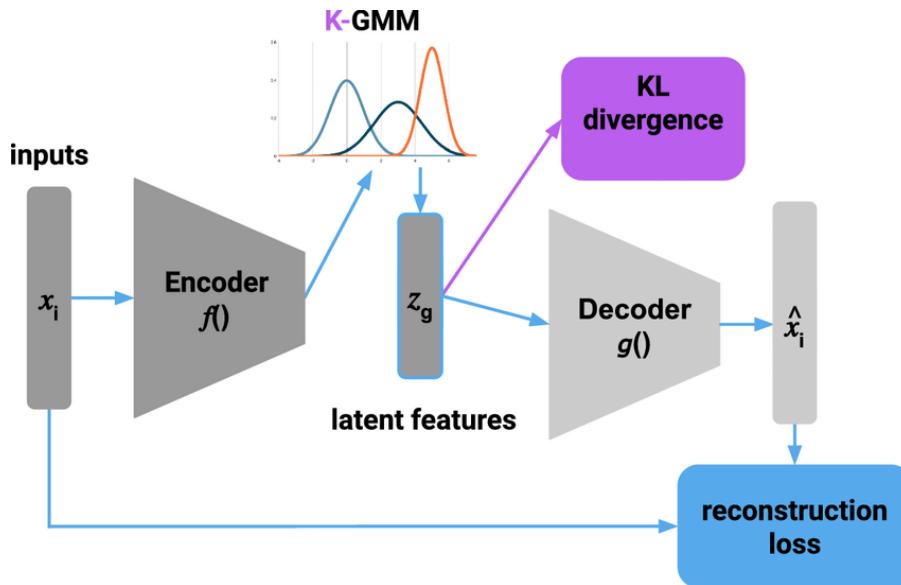


Figura 3.6: Representació esquemàtica de l'arquitectura del model VaDE. Font: Lafabregue *et. al.*, 2022 [13]

El procés generatiu es realitza escollint un clúster c , i realitzant el mateix procés que en el model VAE per obtenir la reconstrucció, utilitzant la corresponent distribució normal.

Durant la fase d'entrenament, el model s'optimitza utilitzant les mateixes funcions de cost de reconstrucció i regularització, utilitzant el mètode de la reparametrització de \mathbf{z} i una fórmula més complexa per calcular el ELBO. Donat que el procés és similar al VAE, no s'ha considerat necessari desenvolupar les fórmules matemàtiques.

3.3 Mètriques d'avaluació

Per validar la QC s'han utilitzat dos tipus de mètriques: validació externa i validació interna. Les mètriques de validació externa es basen en comparar els clústers obtinguts amb una partició de les dades coneguda prèviament, que es considera la partició *correcta*. Les mètriques de validació interna no utilitzen dades externes i es basen en mesures observables en la pròpia estructura dels clústers.

Donat que la tècnica de *clustering* és una tècnica d'aprenentatge no supervisada, normalment s'utilitzen exclusivament les mètriques de validació interna ja que no es coneix l'estruccura real de les dades. En el cas d'aquest TFM, les dades utilitzades provenen d'un estudi clínic i estan etiquetades segons el diagnòstic de cada pacient, per tant s'han pogut utilitzar també mètriques de validació externa.

Cap mètrica per si sola és eficaç per avaluar la QC [14], de manera que s'han seleccionat quatre mètriques: una basada en validació interna i tres basades en validació externa. Es descriuen a continuació:

Mètriques de validació interna La QC es pot mesurar en funció de la cohesió (la proximitat entre els elements dins d'un mateix clúster), i el nivell de separació entre els diferents clústers.

Una de les mètriques més comunes per avaluar aquestes dues característiques és el coefficient

silueta. Es computa de la següent manera, segons les equacions a continuació: per cada punt i es computen la distància mitja $a(i)$ a tots els punts j dins el mateix cluster C_a , la mínima distància mitja $b(i)$ a tots els punts j dins cada cluster C_b diferent de C_a i el seu coeficient silueta $s(i)$. Finalment, es computa la silueta global S prenent la mitjana de les siluetes de tots els punts.

$$a(i) = \frac{1}{|C_a|} \sum_{j \in C_a, i \neq j} d(i, j) \quad (3.17)$$

$$b(i) = \min_{C_b \neq C_a} \frac{1}{|C_b|} \sum_{j \in C_b} d(i, j) \quad (3.18)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.19)$$

$$Sil. = \frac{1}{n} \sum_{i=1}^n s(i) \quad (3.20)$$

El coeficients silueta pren un valor en l'interval $[-1, 1]$, on valors positius indiquen una bona separació entre clústers, valors negatius indiquen que els clústers estan barrejats entre ells, i un coeficient de zero indica que les dades estan distribuïdes uniformement.

Mètriques de validació externa Si es té una partició de referència P , la QC es pot mesurar comparant-la amb els clústers C resultants de l'algoritme de *clustering*. Es construeix un taula de contingència i es comparen les parelles d'observacions trobades en el mateix o diferents clústers en les particions P i C . S'estreuen els següents indicadors:

TP nombre de parelles trobades al mateix clúster tant en C com en P .

FP nombre de parelles trobades al mateix clúster en C , però en diferents clústers en P .

TN nombre de parelles trobades en diferents clústers tant en C com en P .

FN nombre de parelles trobades en diferents clústers en C , però en el mateix clúster en P .

Amb aquests indicadors es pot computar una gran varietat de mètriques, que es poden classificar diverses famílies [14].

- La família de conjunts coincidents es basa en assignar una correspondència entre els clusters C i les particions P , i mesurar la similaritat entre els conjunts.
- La família d'igual a igual parteix de l'assumpció que les observacions que es troben a la mateixa partició a P haurien d'estar també al mateix clúster C , i es basen en les correlacions entre parelles d'observacions.
- La tercera família es basa en conceptes de la teoria de la informació.

S'ha seleccionat una mètrica de cada família, respectivament: l'exactitud, l'índex de Rand ajustat (*Adjusted Rand Index*) (ARI) i el coeficient d'informació mútua ajustat (*Adjusted Mutual Information*) (AMI).

Exactitud L'exactitud (*accuracy*) mesura la proporció d'elements classificats correctament, és a dir la proporció de coincidències entre assignacions dels clústers i les classes reals:

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.21)$$

Pel seu càlcul, és necessari primer associar cada clúster a una de les classes reals. S'ha associat a cada clúster la classe que obté la major freqüència.

ARI Calcula la similaritat entre les particions C i P com la proporció d'el nombre d'elements classificats correctament respecte el nombre total d'elements, corregit per la probabilitat de cada clúster.

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{c_i}{2} \sum_j \binom{p_j}{2}] / \binom{n}{2}}{\frac{1}{2}[\sum_i \binom{c_i}{2} + \sum_j \binom{p_j}{2}] - [\sum_i \binom{c_i}{2} \sum_j \binom{p_j}{2}] / \binom{n}{2}} \quad (3.22)$$

on n_{ij} són els elements d'una matriu de contingència de les particions C i P , p_i i c_i són el número d'elements de les particions de P i C .

AMI Mesura la reducció en la incertesa de l'assignació dels clústers C donada la partició coneguda P , corregit per la probabilitat de cada clúster.

$$MI(P, C) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j} \quad (3.23)$$

$$H(P) = - \sum_{j=1}^p P_P(j) \log P_P(j) \quad (3.24)$$

$$H(C) = - \sum_{j=1}^c P_C(j) \log P_C(j) \quad (3.25)$$

$$AMI = \frac{MI(P, C) - \mathbb{E}(MI(P, C))}{\max\{H(P), H(C)\} - \mathbb{E}(MI(P, C))} \quad (3.26)$$

on p i c són el número de particions de P i C .

3.4 Conjunts de dades

Per comparar els resultats dels diferents mètodes de *clustering* s'han utilitzat tres conjunts de dades, dos d'accés públic i un privat.

MNIST El conjunt de dades *Modified National Institute of Standards and Technology database* (MNIST) [15] és un conjunt de dades freqüentment utilitzat per l'entrenament de models de *machine learning*, especialment amb els relacionats amb el reconeixement o processat d'imatges.

Es tracta d'un conjunt de 70.000 imatges de dígits escrits a mà, etiquetades en funció del dígit que representen. Cada imatge té una resolució de 28 x 28 píxels. En aquest TFM, s'ha tractat cada imatge com una observació i els 784 píxels com 784 variables independents.

Exposome Data Challenge Event Es tracta de diversos conjunts de dades provinents d'un concurs anomenat *Exposome Data Challenge Event* [16], que va tenir lloc a l'abril de 2021. En aquest concurs es van presentar una sèrie de conjunts de dades metabolòmiques, proteòmiques, de metilació del ADN i d'exposició ambiental observades en nens entre 6 i 11 anys, a més de dades fenotípiques i de covariables o possibles factors de confusió.

L'objectiu del concurs era que els competidors presentessin formes innovadores d'analitzar les dades i trobar relacions rellevants en l'àmbit de la salut.

S'han seleccionat dos conjunts de dades sobre els que comparar les diferents tècniques de *clustering*:

- Dades metabolòmiques, formades per dos conjunts de dades, extretes de mostres de sèrum (177 metabolits) i orina (44 metabolits), formant un total de 221 variables.
- Dades d'exposició ambiental o exposoma, compostes per 222 indicadors d'exposició relacionats amb l'exposició a factors ambientals (contaminació de l'aire, soroll), l'exposició a productes químics (metalls, pesticides) i l'estil de vida dels individus (dieta, consum de tabac).

Per mesurar les mètriques de validació externa, s'han seleccionat algunes variables fenotípiques i de les covariables que s'han trobat que poden ser descriptives.

En tots els casos, s'ha seleccionat només aquelles observacions que contenen dades a tots els conjunts de dades. En total, es disposa de 1152 observacions.

Dades DCH-NG En tercer lloc, s'ha utilitzat un conjunt de dades metabolòmiques facilitat pel grup de recerca *Biomarkers and Nutritional & Food Metabolomics*¹, del Departament de Nutrició, Ciències de l'Alimentació i Gastronomia de la Universitat de Barcelona².

Les dades provenen d'un estudi clínic titulat *Diet, Cancer and Health - Next Generations* (DCH-NG) [17], que pretén desenvolupar tècniques innovadores per estudiar la relació entre els gens, la dieta i l'estil de vida amb càncer i altres malalties, aplicant una perspectiva transgeneracional.

Es tracta d'un conjunt de 1120 observacions sobre 411 metabòlits obtingudes en tres moments diferents sobre els mateixos pacients, acompanyades de 16 covariables de les quals dues són categòriques (gènere, índex de risc de malalties cardiovasculars) i la resta numèriques (edat i diversos indicadors de salut).

L'equip de recerca que ha proveït les dades ha anonimitzat la identitat dels pacients i ha prohibit expressament la publicació de les dades per protegir la seva confidencialitat.

3.5 Metodologia

A continuació es descriu la metodologia que s'ha seguit per avaluar els diferents mètodes de *clustering* utilitzant els tres conjunts de dades estudiats.

¹<http://www.nutrimetabolomics.com/>

²https://www.ub.edu/web/ub/ca/universitat/campus_fac_dep/departaments/n/depnutricioCAiG.html?

Tot el codi utilitzat per implementar els models de *deep clustering* i per dur a terme la metodologia que s'explica a continuació s'ha fet públic a un repositori de GitHub: https://github.com/carlescn/MSc_bioinformatics_thesis.

Per cada conjunt de dades, s'han seguit els següents passos amb algunes variacions que s'expliquen a continuació. A la [taula 3.3](#) es resumeixen totes les tècniques avaluades per cada conjunt de dades.

Seleccionar classes de referència Primerament s'ha seleccionat una o diverses variables categòriques independents, que permeten separar les observacions en grups de referència contra els que comparar els clústers obtinguts i calcular les mètriques de validació externa.

El nombre de classes de cada una d'aquestes variables ha determinat el nombre de clústers que s'ha fixat per cada una de les tècniques. En els casos en que s'han escollit variables amb diferents nombres de classes, s'han repetit tots els mètodes pels respectius números de clústers.

En el cas que les variables escollides siguin numèriques i contínues, s'han codificat artificialment en quatre grups utilitzant quantils.

Normalitzar les dades Els conjunts de dades s'han normalitzat aplicant la funció *min-max* per acotar tots els valors dins d'un rang entre 0 i 1. D'aquesta manera s'aconsegueix adaptar les dades a un format que els models de *deep learning* puguin tractar, i reduir els possibles efectes d'escala de les diferents variables.

Tècniques clàssiques A continuació s'han aplicat les tècniques de *clustering* clàssiques seleccionades (K-means, GMM, Aglo.) sobre les dades sense tractar.

Després s'ha aplicat PCA i s'han seleccionat les primeres components principals que expliquen el 80% de la variància. S'han repetit les tècniques clàssiques sobre les dades transformades.

Les mètriques obtingudes amb aquest mètodes s'han pres com a mesura de referència contra la que comparar el rendiment de les tècniques de *deep clustering*.

Mètodes de *deep learning* Seguidament s'han avaluat diversos mètodes de *deep clustering*. Tots els models implementats es basen en un AE amb codificador i descodificador simètrics basats en MLP. El primer pas ha sigut buscar una configuració òptima del AE.

Donat que (tret de les dades MNIST) els conjunts de dades són relativament petits, s'ha intentat reduir al màxim el nombre de paràmetres entrenables de la xarxa neuronal. S'han provat diverses configuracions de nombre i mida de les capes internes i s'ha seleccionat la que aconsegueix reduir el nombre de paràmetres sense afectar negativament a la puntuació de la funció de cost.

El nombre de neurones de la capa latent (així com de la capa de *clustering* quan aplica) s'ha fixat al nombre de clústers que es vol obtenir.

En tots els casos, per totes les neurones s'ha establert la funció d'activació ReLU, excepte les capes d'entrada del codificador i descodificador, així com la sortida del codificador (capa latent), que no tenen funció d'activació, i la sortida del descodificador (sortida de reconstrucció de l'autoencoder), que té funció d'activació sigmoide.

Un cop seleccionada la configuració, s'ha entrenat i avaluat els models DEC i VaDE per cada un dels conjunts de dades i números de clústers escollits.

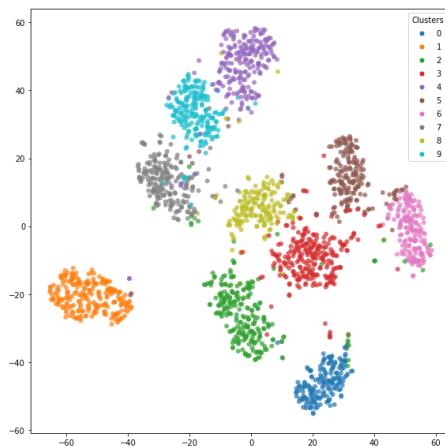


Figura 3.7: Exemple de gràfica t-SNE del model Va-DE entrenat sobre les dades MNIST. Es representa la CR utilitzant les dues primeres components de la reducció t-SNE. Els colors mostren les assignacions de clústers del model.

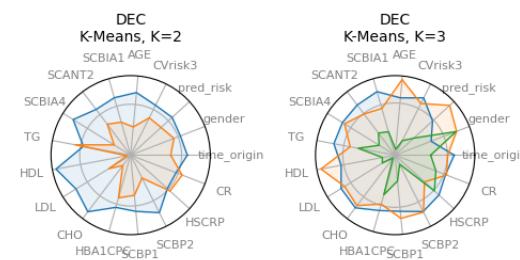


Figura 3.8: Exemple de gràfica radial. Es representa la distribució multivariant de les covariables del conjunt de dades DCH-NG en funció de les assignacions de clústers trobades pel model DEC, fixant el nombre de clústers a 2 i 3. Cada radi del diagrama representa una variable, i els polígons del centre representen la distribució multivariant que pren cada clúster, on cada vèrtex és la mitjana dels valors per cada variable.

Avaluació dels resultats Finalment, s'han compilat totes les mètriques i s'han comparat els clústers obtinguts mitjançant representacions gràfiques.

Per cada tècnica i número de clústers s'han visualitzat el conjunt de dades (originals o transformades, segons la tècnica utilitzada) en un núvol de punts de dues dimensions mitjançant la tècnica *t-Distributed Stochastic Neighbor Embedding* (t-SNE), assignant un color diferent a cada clúster ([figura 3.7](#)).

Mitjançant gràfiques radials ([figura 3.8](#)), s'ha representat la distribució multivariant de les variables independents pels clústers trobats amb cada combinació de tècnica i nombre de clústers.

S'han seleccionat les 20 variables amb més variabilitat de cada conjunt de dades i s'ha representat la seva distribució multivariant pels diferents clústers de la mateixa manera.

Per últim, s'han comparat entre elles les assignacions de clústers de cada un dels mètodes mitjançant un gràfic tipus *heatmap* ([figura 3.9](#)), on cada fila representa una de les tècniques avaluades, les columnes representen les observacions del conjunt de dades, i el color representa l'assignació de cada mostra a un dels clústers.

3.5.1 MNIST

El conjunt de dades MNIST s'ha seleccionat per ser un exemple molt estudiat i sobre el que es coneix que es poden aconseguir bons resultats de *clustering* amb diferents mètodes, clàssics i basats en *deep learning*.

S'ha utilitzat com a marc de referència per validar la funcionalitat dels diferents models de *deep clustering* implementats, i seleccionar aquells que funcionin millor.

El nombre de clústers s'ha fixat en 10, ja que només s'han comparat amb les etiquetes de les imatges.

Mètodes clàssics Només s'han avaluat les tècniques K-means i GMM.

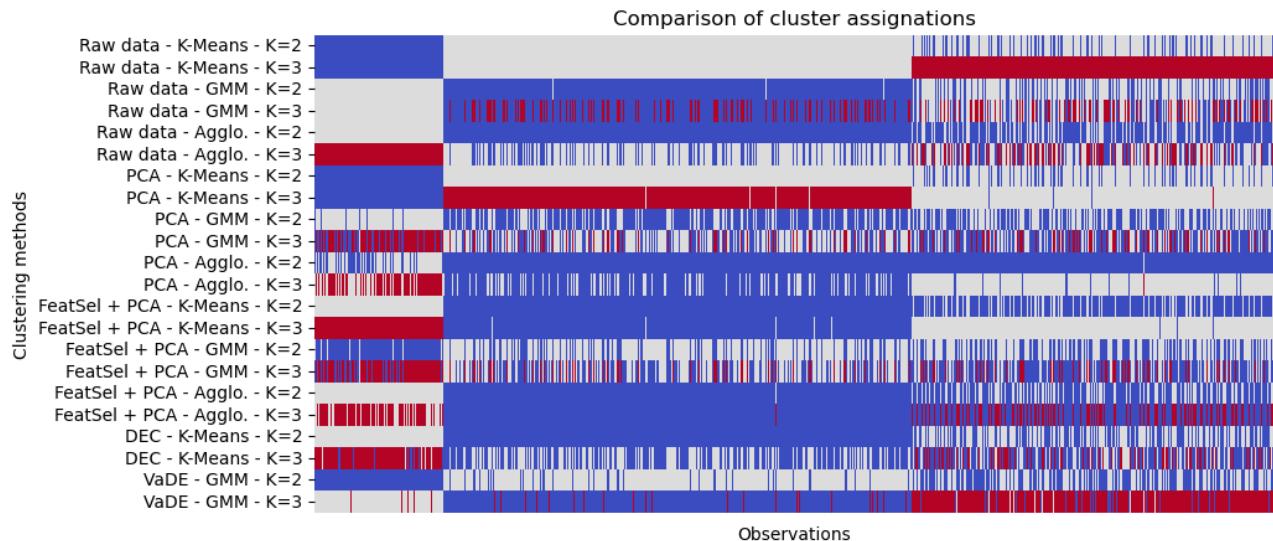


Figura 3.9: Exemple de gràfica *heatmap* on es comparen els clústers trobats pels diferents mètodes evaluats sobre el conjunt de dades DCH-NG. Cada fila representa una de les tècniques, les columnes representen les observacions i el colors representen a quin clúster s'ha assignat cada observació. Per poder comparar les tècniques, totes les files s'han ordenat segons el mateix índex.

Mètodes de deep clustering En tots els casos s'ha utilitzat el mateix subcojunt de 60.000 imatges per entrenar el model i les 10.000 restants per avaluar el seu rendiment.

S'ha escollit una configuració del AE, que es coneix dona bons resultats en aquest conjunt de dades [9]: tres capes ocultes amb mides 512, 512 i 2048.

S'han contrastat els següents mètodes:

- Entrenar un model AE amb les dades d'entrenament, codificar les dades de validació i aplicar K-means i GMM sobre les dades transformades (capa latent).
- Repetir la mateixa operació amb un model VAE.
- Entrenar i avaluar un model DEC, inicialitzant els paràmetres del model primer amb K-means i amb GMM.
- Entrenar i avaluar un model mixt VAE+DEC, que reemplaça el AE de l'arquitectura DEC per un VAE. Els paràmetres s'han inicialitzat amb K-means i amb GMM.
- Repetir la mateixa operació, però mantenint el descodificador del model VAE i afegint la funció de cost de reconstrucció durant l'entrenament.
- Entrenar i avaluar un model VaDE, inicialitzant els paràmetres amb GMM.

Avaluació dels mètodes Finalment, s'han comparat totes les mètriques i s'han seleccionat aquells models de *deep learning* que mostren un bon rendiment en relació amb la seva complexitat: DEC i VaDE.

S'han comparat els clústers trobats per les diferents tècniques utilitzant les representacions t-SNE de les dades. No s'han representat els gràfics radials multivariants.

3.5.2 Exposome Data Challenge Event

Amb aquest conjunt de dades s'han comparat tres tècniques de *clustering* clàssiques amb els dos models de *deep clustering* seleccionats al pas anterior (DEC, VaDE). Donat que en aquest cas sí es tracta de dades metabolòmiques, s'espera que el resultat sigui extrapolable a altres conjunts de dades del mateix tipus.

S'han seleccionat 11 variables independents (veure [taula 3.1](#)). Per tant, s'han avaluat les tècniques pels fixant el nombre de clústers a 2, 3, 4, 6 i 7.

Grup de covars.	Nom	Núm. classes	Descripció
Fenotip	birth_weight	4*	Pes al néixer.
Fenotip	iq	4*	Quocient d'intel·ligència.
Fenotip	behaviour	4*	Comportament neurològic (índex).
Fenotip	asthma	2	Incidència d'asma.
Fenotip	bmi	4*	Índex de massa corporal.
Covariables	cohort	6	Cohort d'inclusió a l'estudi.
Covariables	age	7	Edat en anys.
Covariables	sex	2	Gènere.
Covariables	education	3	Nivell d'estudis de la mare.
Covariables	native	3	Pares natius del país de naixement.
Covariables	parity	3	Nombre d'embarassos previs (mare).

Taula 3.1: Variables independents seleccionades del conjunt de dades Exposome Data Challenge Event.

*Variables numèriques contínues, convertides en 4 grups mitjançant quantils.

S'han estudiat independentment dos conjunts de dades: les dades metabolòmiques (s'ha unit les lectures en sèrum i en orina per obtenir un conjunt de dades més gran) i les dades del exposoma.

Per les tècniques de *deep clustering*, la configuració seleccionada del AE és de tres capes internes amb mides 16, 16 i 128.

Dades metabolòmiques

Tècniques clàssiques S'han avaluat (K-means, GMM, Aglo.) sobre les dades originals i les transformacions PCA.

Addicionalment, motivat pels resultats poc prometedors obtinguts, s'han seleccionat les variables amb més variabilitat (desviació estàndard superior a la mitjana del dataset, avaluades per separat en les dades de sèrum i orina). S'ha repetit l'avaluació dels mètodes sobre el conjunt de dades reduït.

Mètodes de *deep clustering* S'ha entrenat i avaluat els dos models de *deep clustering* seleccionats (DEC, VaDE). Donat que els resultats inicials no han semblat molt prometedors, s'ha intentat millorar la relació entre la mida de les dades i el nombre de paràmetres de dues maneres:

- Augmentar artificialment el nombre de dades. S'han generat mostres aleatòries d'una distribució normal amb mitjana 0 i desviació estàndard 0.01 (un ordre de magnitud inferior a la de les dades).

A continuació s'han sumat aquestes mostres als valors de les dades originals. S'ha repetit el procés 10 vegades per obtenir un conjunt de dades amb 11.520 observacions, que presenten una distribució similar a les dades originals.

Amb aquestes dades, s'han entrenat i avaluat els dos models de *deep clustering*.

- Reduir el nombre de paràmetres entrenables de la xarxa neuronal. S'ha substituït les capes internes del tipus MLP del AE per capes convolucionals, amb una i dues dimensions. Això permet reduir dràsticament el nombre de paràmetres del model, però assumeix que les dades presenten algun tipus d'estructura invariable.

S'ha necessitat adaptar les dades a una mida d'entrada que aquests models puguin tractar. Pel model amb capes convolucionals 1D, s'han convertit els vectors de 221 valors de cada observació en un vector 1D amb 224. Pel model amb capes 2D, s'han transformat en matrius bidimensionals de mida 16 x 16 (256 punts). En ambdós casos s'han omplert amb zeros les posicions buides.

Per cada una de les dues configuracions (1D, 2D), s'han seleccionat els paràmetres òptims del AE. Pel model 1D, una capa convolucional de mida 4. i mida del kernel 3. Pel model 2D, dues capes convolucionals de mides 4 i 4, i mida del kernel 3.

Finalment, s'han implementat i avaluat els quatre models de *deep learning* (DEC i VaDE, amb capes convolucionals 1D i 2D).

Dades de l'exposoma

Tècniques clàssiques S'han avaluat (K-means, GMM, Aglo.) sobre les dades originals i les transformacions PCA. No s'ha reduït la mida del conjunt de dades.

Mètodes de *deep clustering* S'han avaluat els models DEC i VaDE basats en MLP. No s'han aplicat tècniques d'augment artificial de les dades ni models amb capes convolucionals.

Correcció de l'efecte lot Els resultats obtinguts suggereixen la presència d'un fort efecte de lot, ja que totes les tècniques aconsegueixen un solapament gairebé perfecte amb els grups de la variable *cohort*.

Per aquest motiu, s'ha decidit aplicar una correcció per aquest efecte lot als dos conjunts de dades. Per cada classe de la variable *cohort*, s'han estandarditzat els valors de cada variable restant la mitjana i dividint per al desviació estàndard. Les dades corregides obtingudes s'han normalitzat de nou amb la funció *min-max*.

Finalment, s'han tornat a avaluar els mateixos cinc mètodes de clustering (K-means, GMM, Aglo., DEC i VaDE).

Avaluació dels models S'han compilat les mètriques obtingudes amb totes les tècniques i s'han comparat les assignacions de clústers obtingudes mitjançant gràfiques t-SNE, gràfiques tipus heatmap i gràfiques radials per les distribucions multivariants de les covariables i de les 20 variables amb major variabilitat de cada conjunt de dades.

3.5.3 Dades DCH-NG

Com a grups objectiu contra els que avaluar les mètriques s'han seleccionat dues variables categòriques de les covariables, que s'ha trobat mostren una distribució diferencial per la resta de covariables (indicadors biològics relacionats amb la salut) (veure [taula 3.2](#)). Tots els mètodes s'han avaluat fixant el número de clústers a 2 i 3.

Nom	Núm. grups	Descripció
gender	2	Gènere de l'individu.
CVrisk3	3	Risk de patir malalties cardiovasculars.

Taula 3.2: Variables independents seleccionades del conjunt de dades DCH-NG.

Tècniques clàssiques S'han avaluat (K-means, GMM, Aglo.) sobre les dades originals i les transformacions PCA.

Mètodes de deep clustering S'han avaluat els models DEC i VaDE basats en MLP. La configuració escollida per l'AE de base és de dues capes amb 16 i 32 neurones.

Avaluació dels models Donat que les dades contenen mesures repetides sobre els mateixos individus, i s'espera que les variacions entre individus siguin majors que entre les observacions del mateix individu, s'ha calculat una mètrica addicional: la freqüència amb que les observacions del mateix individu s'assignen a un únic clúster (corregida per la mida del clúster).

S'han compilat les mètriques obtingudes amb totes les tècniques i s'han comparat les assignacions de clústers obtingudes mitjançant gràfiques t-SNE, gràfiques tipus heatmap i gràfiques radials per les distribucions multivariants de les covariables i de les 20 variables amb major variabilitat de cada conjunt de dades.

3.6 Eines informàtiques

Software Els models de *deep clustering* s'han implementat utilitzant el software Keras [4], una eina basada en TensorFlow [18] que es va desenvolupar per l'àmbit de la recerca. Permet definir i entrenar models de xarxes neuronals profunds de manera relativament senzilla, però alhora possibilita interactuar a més baix nivell amb TensorFlow per realitzar operacions més complexes [5].

Pràcticament la totalitat de l'estudi s'ha realitzat mitjançant el llenguatge de programació Python [19], documentat en llibretes Jupyter Notebook [20]. Com s'ha mencionat al inici de la secció, tot el codi està disponible a un repositori a GitHub.

Conjunt de dades	Tipus	Feature learning	Clustering	Incialització	Variacions
MNIST	Clàssic	Cap, PCA AE VAE DEC VAE+DEC VaDE	K-means, GMM K-means, GMM K-means, GMM DEC VAE+DEC VaDE	- - - K-means, GMM K-means, GMM GMM	- - - - - -
	Deep learning				Mantenir descodificador
ExposomeChall. (metaboloma)	Clàssic	Cap, PCA	K-means, GMM	-	Selecció de variables, Correcció efecte lot
	Deep learning	DEC VaDE	DEC VaDE	K-means, GMM GMM	Correcció efecte lot, augment de dades, conv 1D, 2D
ExposomeChall. (exposoma)	Clàssic	Cap, PCA DEC VaDE	K-means, GMM DEC VaDE	- K-means, GMM GMM	Correcció efecte lot
	Deep learning				
Dades DCH-NG	Clàssic	Cap, PCA DEC VaDE	K-means, GMM DEC VaDE	- K-means, GMM GMM	- - -
	Deep learning				

Taula 3.3: Resum de les tècniques de *clustering* avaluades per cada conjunt de dades.

Hardware i serveis de computació *on-line*. L'equip informàtic del que disposa l'estudiant, si bé té capacitat per executar les tècniques de *clustering* clàssiques, presenta limitacions a l'hora d'entrenar models basats en xarxes neuronals profundes, que consumeixen una gran quantitat de memòria RAM i requereixen de GPUs potents per una execució més o menys àgil.

Per aquest motiu, ha sigut imprescindible utilitzar un servei de computació *on-line*. Concretament, s'ha utilitzat el servei de pagament Paperspace Gradient³. A la capítol 6 es desglossa el cost d'aquest servei.

³<https://docs.paperspace.com/gradient/>

Capítol 4

Resultats

4.1 MNIST

A la taula 4.1 es mostren les mètriques obtingudes per cada un dels models avaluats. Les tècniques clàssiques aplicades sobre les dades sense transformar i sobre la transformació PCA han aconseguit el mateix rendiment, mentre que els models de *deep clustering* en general han aconseguit un rendiment notablement més elevat. Això suggereix que les tècniques de *deep learning* són capaces de trobar relacions no lineals que es perdren al aplicar PCA.

Dintre dels models de *deep clustering*, en referència a les mètriques de validació externa (és a dir, el grau de coincidència amb les etiquetes reals) cap destacat que els models basats en VAE són els que obtenen pitjor rendiment.

Per altra banda, el model DEC pràcticament obté els mateixos resultats que aplicar tècniques clàssiques a la capa de representació del AE. Això suggereix una alta dependència dels model DEC en els paràmetres amb què s'inicialitza durant l'entrenament. Ambdós obtenen els segons millors resultats.

Els millors resultats per les mètriques de validació externa s'obtenen amb el model VaDE, que aconsegueix una exactitud de 94%.

Per últim, per la mètrica de validació interna el model DEC destaca sobre la resta. El model VAE+DEC on s'ha eliminat el descodificador obté una puntuació similar al DEC, mentre que el model VAE+DEC que manté el descodificador perd aquesta capacitat.

A la figura 4.1 s'il·lustra la diferència entre els clústers obtinguts pels diferents models. El model VAE, encara que té certa capacitat per distingir els diferents grups, té dificultats per separar-los en la CR donat que tots els punts comparteixen la mateixa distribució normal. En conseqüència, els clústers aconseguits aplicant GMM mostren solapament. En comparació, els models VaDE i DEC són capaços de separar els clústers i obtenir una millor QC.

4.2 Exposome Data Challenge Event

Les mètriques de validació mesurades en aquest conjunt de dades han resultat poc informatives. S'ha trobat que tots els models han aconseguit un solapament gairebé perfecte amb la covariable *cohort*, suggerint un fort efecte lot en les dades.

Tipus	Feature learning	Clustering	Incialització	Acc.	ARI	AMI	Sil.
Clàssic	-	K-means	-	0.59	0.41	0.53	0.06
		GMM	-	0.44	0.22	0.34	0.02
	PCA	K-Means	-	0.59	0.41	0.53	0.09
		GMM	-	0.47	0.23	0.43	0.02
Deep learning	AE	K-means	-	0.83	0.69	0.73	0.19
		GMM	-	0.77	0.57	0.68	0.14
	DEC	DEC	K-means	0.83	0.69	0.74	0.93
		DEC	GMM	0.76	0.56	0.69	0.93
	VAE	K-means	-	0.59	0.41	0.54	0.16
		GMM	-	0.48	0.25	0.39	-0.01
	VAE+DEC ¹	VAE+DEC ¹	K-Means	0.61	0.42	0.54	0.90
		VAE+DEC ¹	GMM	0.57	0.31	0.47	0.87
	VAE+DEC ²	VAE+DEC ²	K-means	0.57	0.37	0.49	0.20
		VAE+DEC ²	GMM	0.59	0.36	0.49	0.12
	VaDE	VaDE	GMM	0.94	0.88	0.88	0.23

Taula 4.1: Resum dels resultats obtinguts sobre els conjunt de dades MNIST.

¹ Model sense descodificador i per tant sense funció de cost de reconstrucció.

² Model amb descodificador i funció de cost de reconstrucció.

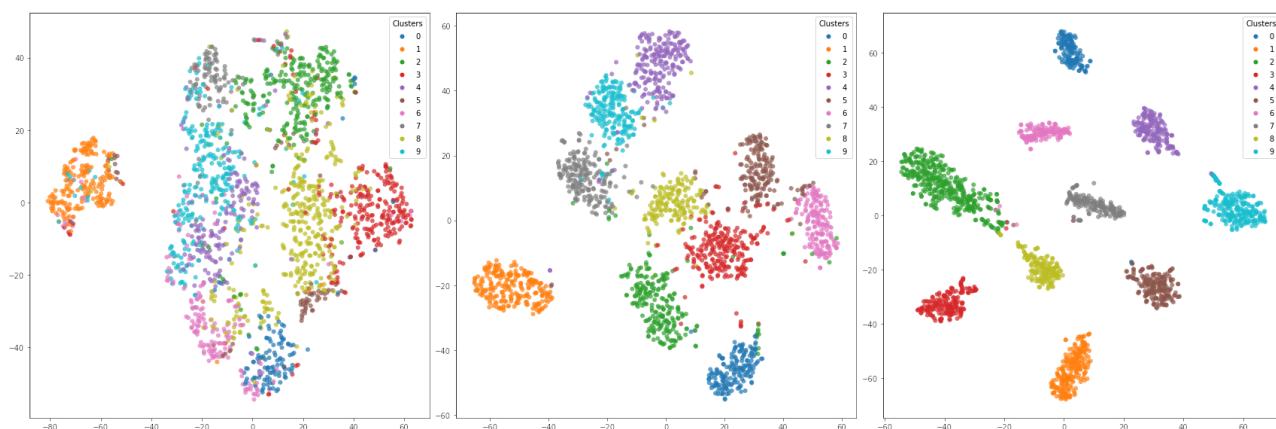


Figura 4.1: Comparació dels models VAE + GMM (esquerra), VaDE (centre) i DEC (dreta). Representació t-SNE de la CR dels models. Els colors dels punts representen l'assignació a diferents clústers.

No obstant, tot i aplicant una tècnica per corregir per aquest efecte, no s'ha trobat una relació clara entre els clústers i les covariables estudiades.

Els mètodes addicionals que s'han estudiat (models convolucionals, selecció de variables amb més variació, augment artificial de les dades) no han aconseguit una millora en el rendiment.

De nou, s'observa que el mètode DEC aconsegueix destacar-se reiteradament en quant a la mètrica silueta.

A les [taules 4.2 i 4.3](#) es mostra per cada tècnica la mitjana de les mètriques obtingudes sobre els diferents números de clústers fixats. A la [secció 3.5](#) s'explica amb més detall com interpretar aquesta gràfica. Els resultats complets s'adjunten al [apèndix A](#).

Al comparar els clústers trobats per les diferents tècniques, s'observa un bon nivell de solapament entre totes les tècniques, amb l'excepció de GMM aplicada sobre la transformació PCA, que sembla trobar una estructura alternativa.

A la [figura 4.2](#) es mostra la comparació dels clústers trobats per les diferents tècniques avaluades sobre subconjunt de dades metabolòmiques corregides per l'efecte de lot. La resta de gràfiques s'adjunte al [apèndix A](#).

Finalment, respecte a la distribució multivariant de les covariables, s'observa una possible distribució diferencial en funció dels lots que podria ser indicativa d'una interpretació biològica, però seria necessari un estudi multivariant posterior per treure'n conclusions.

La distribució multivariant de les variables amb més variació no sembla mostrar una distribució diferencial en funció dels clústers, però com s'ha indicat abans caldria estudiar-ho amb més profunditat.

A la [figura 4.3](#) es mostra la distribució diferencial de les variables fenotípiques en funció dels clústers trobats pels diferents mètodes, fixant el número de clústers a 4. La resta de gràfiques s'adjunte al [apèndix A](#).

4.3 Dades DCH-NG

De nou, les mètriques de validació mesurades en aquest conjunt de dades han resultat poc informatives: cap dels mètodes ha aconseguit un bon rendiment respecte a les mètriques de validació externes. Respecte a la mètrica silueta, de nou el model DEC destaca significativament.

Encara que els clústers obtinguts no mostren correlació amb les covariables categòriques preses com a referència, sí mostren un alt nivell de solapament entre les diferents tècniques ([figura 4.4](#)).

Respecte a la distribució multivariant de les covariables ([figura 4.5](#)), sembla que els grups trobats per les tècniques basades en K-means es relacionen amb valors més o menys elevats dels indicadors de salut (variables numèriques), mentre que les basades en GMM, a més contrasten aquests valors contra la variable *gènere*.

En quant a la distribució multivariant de les variables amb més variació ([figura 4.6](#)), sembla que els valors dels metabòlits alanina, isoleucina, leucina, i en menor grau l'oxaloacètic, són els que presenten majors diferències entre clústers.

Sembla, per tant, que els clústers trobats poden tenir certa interpretabilitat biològica. Seria necessari estudiar-ho amb més profunditat.

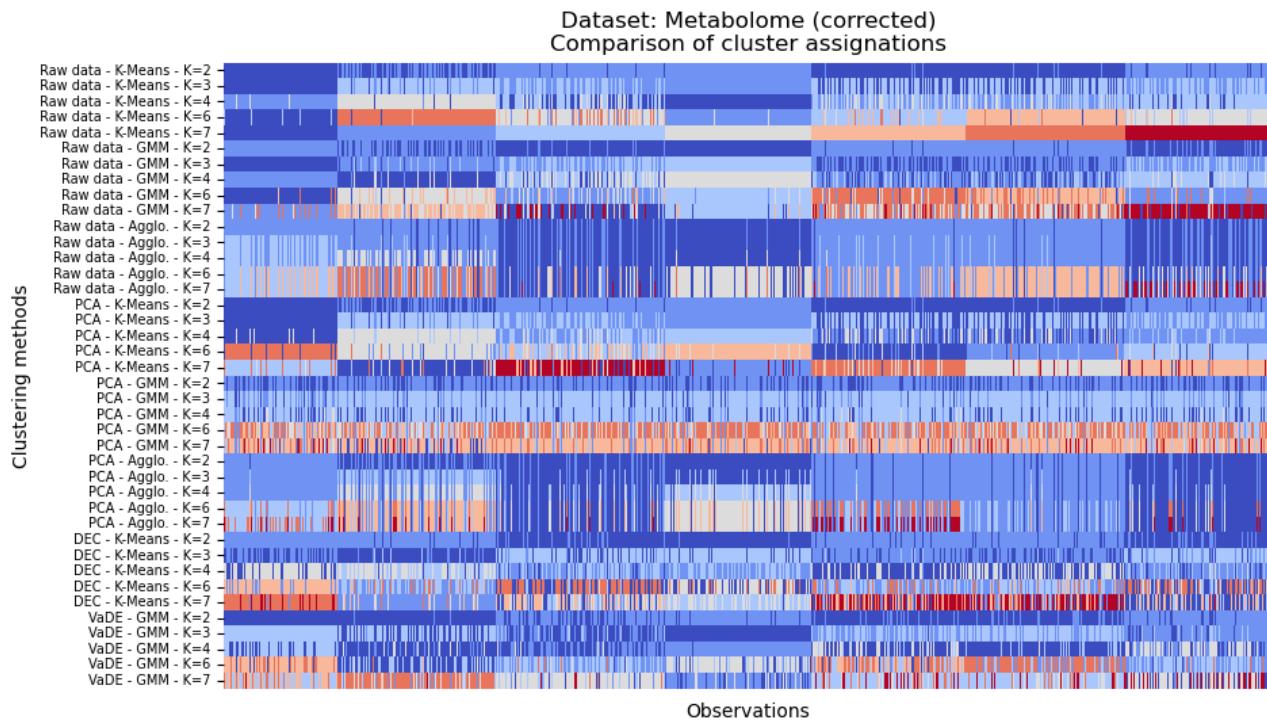


Figura 4.2: Comparació dels clústers trobats pels diferents mètodes evaluats sobre el conjunt de dades Exposome Data Challenge Event, subconjunt dades metabolòmiques corregides per l'efecte de lot. S'observa un bon nivell de solapament entre totes les tècniques, excepte PCA + GMM que sembla trobar una estructura alternativa.

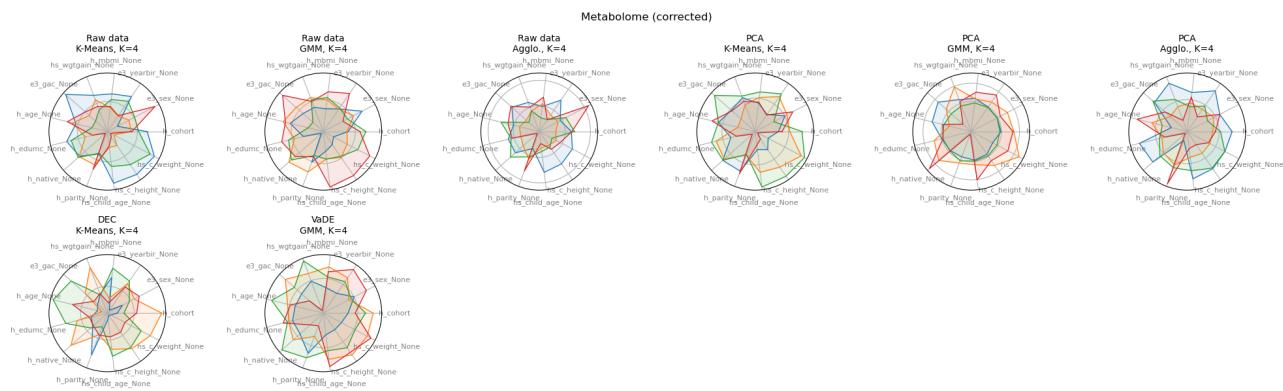


Figura 4.3: Representació de la distribució diferencial de les variables fenotípiques al conjunt de dades Exposome Data Challenge Event en funció dels clústers trobats pels diferents mètodes, fixant el nombre de clústers a 4.

Modificació dades	Tipus	Feature learning	Clustering	Inicialització	Acc.	ARI	AMI	Sil.
Clàssic	Raw data	PCA	Aggro.	-	0.46	0	0	0.06
			GMM	-	0.46	0	0	0
			K-Means	-	0.46	0	0	0.09
			Aggro.	-	0.46	0	0	0.05
	DEC	Raw data	GMM	-	0.46	0	0	0.06
			K-Means	-	0.46	0	0	0.07
	Deep clustering	DEC + D.A.	DEC	K-Means	0.48	0	0	0.54
		VaDE	VaDE	GMM	0.48	0	0	0.17
		VaDE + D.A.	VaDE	GMM	0.48	0	0	0.19
		DEC (conv. 1D)	DEC	K-Means	0.48	0	0	0.69
		DEC (conv. 2D)	DEC	K-Means	0.48	0	0	0.46
		VaDE (conv. 1D)	VaDE	GMM	0.48	0	0	0.21
Selecció de variables	Raw data	PCA	Aggro.	-	0.46	0	0	0.07
			GMM	-	0.46	0	0	0.05
			K-Means	-	0.46	0	0	0.08
			Aggro.	-	0.46	0	0	0.06
	DEC	Deep clustering	GMM	-	0.46	0	0	0.05
			K-Means	-	0.46	0	0	0.07
	DEC (D.A.)	DEC	DEC	K-Means	0.48	0	0	0.5
		DEC (D.A.)	DEC	K-Means	0.48	0	0	0.77
		VaDE	VaDE	GMM	0.48	0	0	0.17
		VaDE (D.A.)	VaDE	GMM	0.48	0	0	0.28
Correcció efecte lot	Raw data	PCA	Aggro.	-	0.46	0	0	0.07
			GMM	-	0.46	0	0	0.02
			K-Means	-	0.46	0	0	0.09
			Aggro.	-	0.46	0	0	0.06
	DEC	Deep clustering	GMM	-	0.46	0	0	0.07
			K-Means	-	0.46	0	0	0.07
	VaDE	DEC	DEC	K-Means	0.48	0	0	0.49
		VaDE	VaDE	GMM	0.48	0	0	0.22

Taula 4.2: Resum dels resultats obtinguts sobre el subconjunt de dades metabolòmiques del conjunt de dades Exposome Data Challenge Event (les puntuacions són la mitjana per tots els números de clústers evaluats). (D.A.): augment de dades.

Modificació dades	Tipus	Feature learning	Clustering	Inicialització	Acc.	ARI	AMI	Sil.
-	Clàssic	PCA	Agglo.	-	0.6	0.17	0.18	0.13
			GMM	-	0.59	0.16	0.18	0.11
			K-Means	-	0.59	0.16	0.18	0.13
		Raw data	Agglo.	-	0.6	0.17	0.18	0.1
			GMM	-	0.59	0.17	0.19	0.1
	Deep clustering	DEC	K-Means	-	0.59	0.16	0.18	0.1
			DEC	K-Means	0.6	0.16	0.17	0.79
		VaDE	VaDE	GMM	0.58	0.12	0.14	0.45
			Agglo.	-	0.46	0	0	0.02
			PCA	GMM	-	0.46	0	0.02
Correcció efecte lot	Clàssic	PCA	K-Means	-	0.47	0	0	0.03
			Agglo.	-	0.46	0	0	0.02
			Raw data	GMM	-	0.49	0.02	0.02
		Raw data	K-Means	-	0.47	0	0	0.02
			DEC	K-Means	0.48	0	0	0.64
	Deep clustering	DEC	VaDE	GMM	0.48	0	0	0.27
			Agglo.	-	0.46	0	0	0.02
		VaDE	PCA	GMM	-	0.46	0	0.02
			K-Means	-	0.47	0	0	0.03
			Agglo.	-	0.46	0	0	0.02

Taula 4.3: Resum dels resultats obtinguts sobre el subconjunt de dades de l'exposoma del conjunt de dades Exposome Data Challenge Event (les puntuacions són la mitjana per tots els números de clústers evaluats).

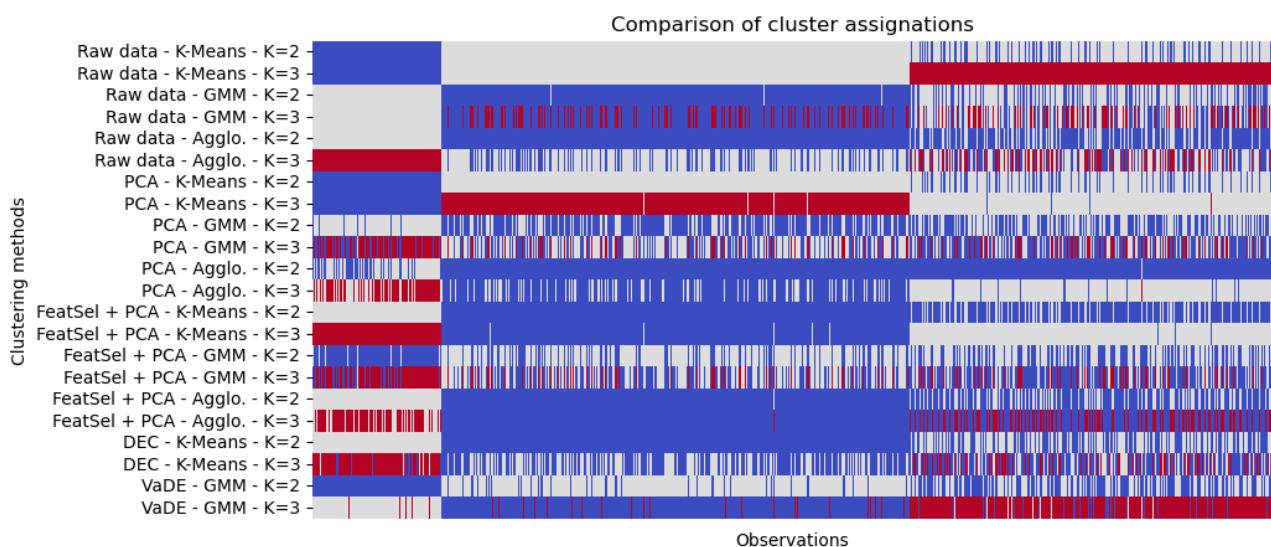


Figura 4.4: Comparació dels clústers trobats pels diferents mètodes evaluats sobre el conjunt de dades DCH-NG.

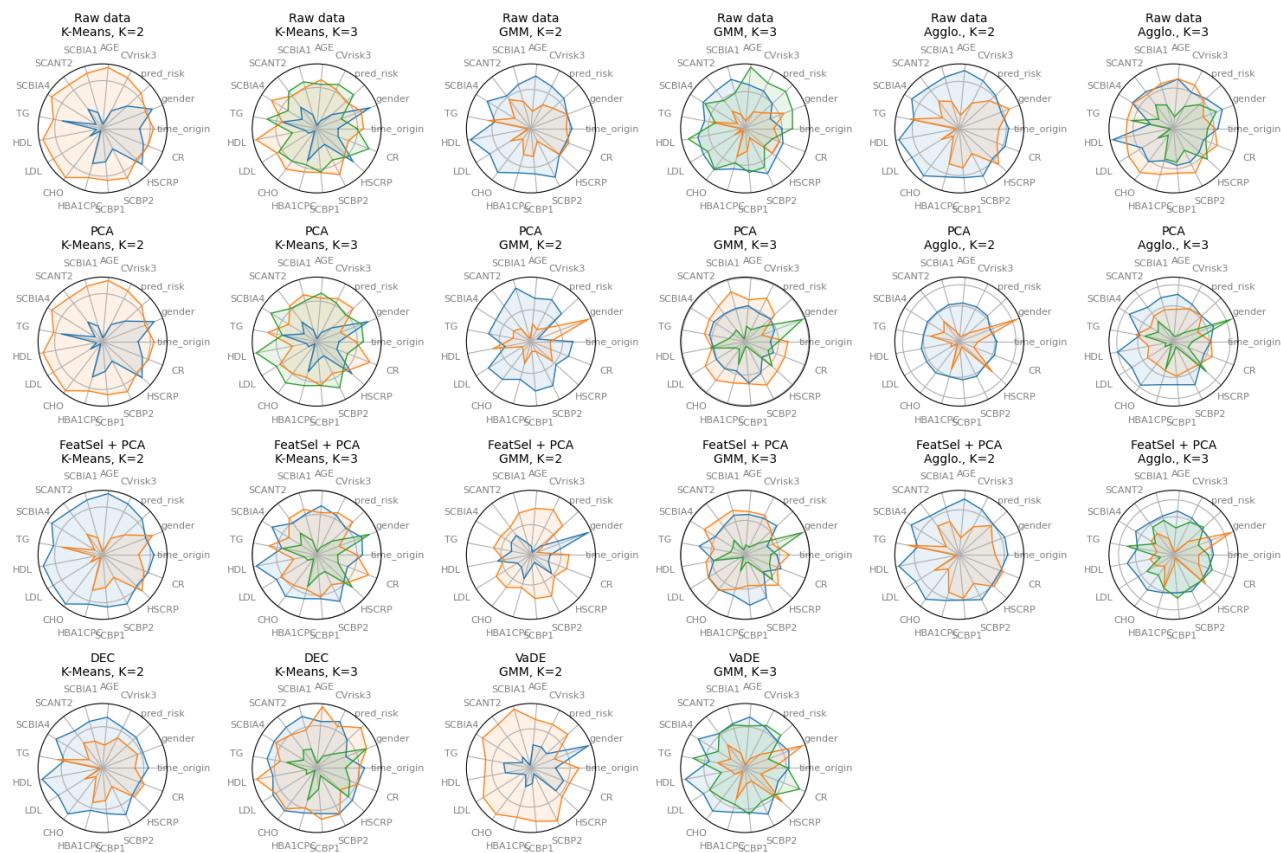


Figura 4.5: Representació de la distribució diferencial de les covariables del conjunt de dades DCH-NG en funció dels clústers trobats pels diferents mètodes avaluats.

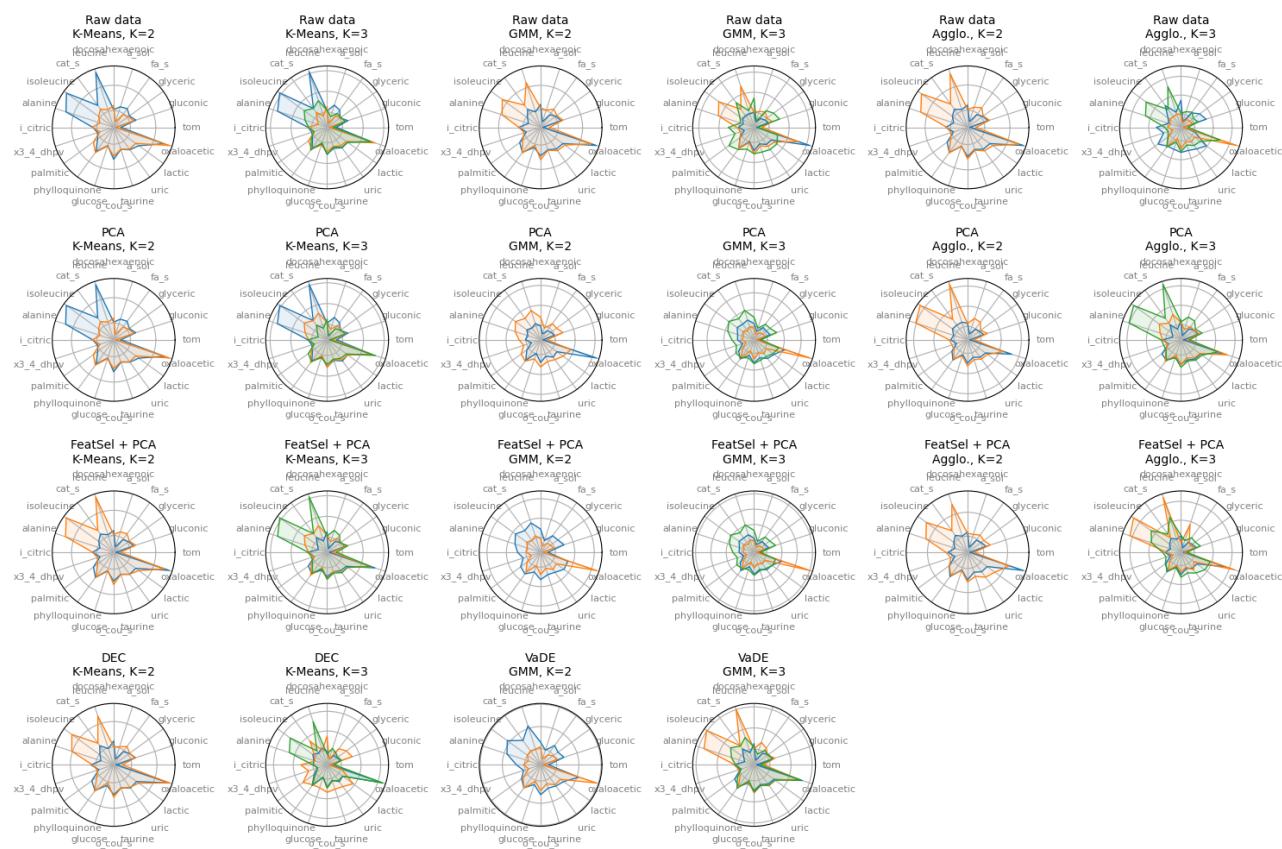


Figura 4.6: Representació de la distribució diferencial dels metabòlits amb més variància del conjunt de dades DCH-NG en funció dels clústers trobats pels diferents mètodes avaluats.

Tipus	Feature learning	Clustering	Incialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.	Mateix clúst.	
	K-Means	-	2	gender	0.55	0	0	0.25	0.47		
		-	3	CVrisk3	0.36	0	0	0.09	0.21		
	GMM	-	2	gender	0.55	0	0	0.17	0.33		
		-	3	CVrisk3	0.36	0	0	0.1	0.17		
	Agglo.	-	2	gender	0.55	0	0	0.24	0.46		
		-	3	CVrisk3	0.36	0	0	0.11	0.18		
	K-Means	-	2	gender	0.55	0	0	0.29	0.47		
		-	3	CVrisk3	0.36	0	0	0.12	0.2		
	GMM	-	2	gender	0.55	0.01	0.01	0.14	0.24		
		-	3	CVrisk3	0.37	0	0	0.02	0.12		
Clàssic	PCA	-	2	gender	0.55	0	0	0.38	0.75		
		-	3	CVrisk3	0.35	0	0	0.1	0.24		
	Agglo.	-	2	gender	0.55	0	0	0.32	0.48		
		-	3	CVrisk3	0.36	0	0	0.14	0.2		
	K-Means	-	2	gender	0.55	0.01	0.01	0.16	0.25		
		-	3	CVrisk3	0.36	0	0	0.02	0.14		
	Selecció de variables	GMM	-	2	gender	0.55	0	0	0.28	0.41	
	+ PCA	-	2	gender	0.55	0	0	0.22	0.39		
	Agglo.	-	3	CVrisk3	0.36	0	0	0.96	0.35		
	DEC	K-Means	2	gender	0.55	0	0	0.93	0.15		
	DEC	K-Means	3	CVrisk3	0.36	0	0	0.57	0.3		
	VaDE	GMM	2	gender	0.55	0	0	0.45	0.2		
	VaDE	GMM	3	CVrisk3	0.36	0	0				

Taula 4.4: Resum dels resultats obtinguts sobre el conjunt de dades metabolòmiques DCH-NG. (*Mateix clust.: freqüència amb que les observacions del mateix individu s'assignen a un únic clúster (corregida per la mida del clúster)*)

Capítol 5

Discussió

Els resultats obtinguts amb les dades MNIST mostren que els models de *deep clustering* obtenen en general millor QC que les tècniques clàssiques. Això suggereix que són capaços d'aprendre LF més eficients degut a la seva capacitat de trobar relacions no lineals.

El model DEC obté uns clústers molt similars als aconseguits aplicant tècniques clàssiques a la CR d'un AE, però aconsegueix millor QC. Els models basats en VAE no aconsegueixen un bon rendiment.

Els models que han obtingut millor rendiment són els DEC i VaDE. La resta de models s'han descartat de la resta de l'estudi.

Model DEC El model DEC ha aconseguit generar els clústers amb millors puntuacions per la mètrica silueta, amb diferència. Això indica que la CR té una gran capacitat de separar les observacions en funció dels clústers aconseguits.

Per altra banda, aquest model mostra ser especialment sensible als paràmetres amb que s'inicialitza. En conseqüència, el seu rendiment depèn de que les LF apreses durant la fase de pre-entrenament propiciïn una bona QC amb els mètodes clàssics utilitzats per inicialitzar els paràmetres.

Durant la fase d'ajustament del model, s'accentua el criteri utilitzat per separar els clústers inicials, de manera que no s'obté nova informació, però s'aconsegueixen clústers més comprimits, resultant en una millor QC.

Model VAE La principal característica del model VAE és que la seva CR es basa en un model probabilístic i per tant no determina punts fixos en un espai dimensional, sinó els paràmetres d'una distribució normal multivariant de la qual es mostren punts aleatoris. Això aconsegueix que la CR sigui contínua i per tant pugui funcionar com un model generatiu, mostrent punts de la distribució latent i passant-los al descodificador per obtenir una nova observació artificial que hauria de compartir la distribució de les variables originals.

S'ha trobat que els models basats en VAE no aconsegueixen una bona QC. Això és degut a l'aleatorietat del model probabilístic i el fet que tots els punts comparteixin la mateixa distribució latent. En conseqüència, encara que el model pugui ser capaç de diferenciar grups, és difícil aconseguir clústers que no presentin solapament.

Aquesta idea queda il·lustrada a la [figura 5.1](#), on es comparen els models VAE+DEC que mantenen o descartan el descodificador: el model que el manté el descodificador, i per tant

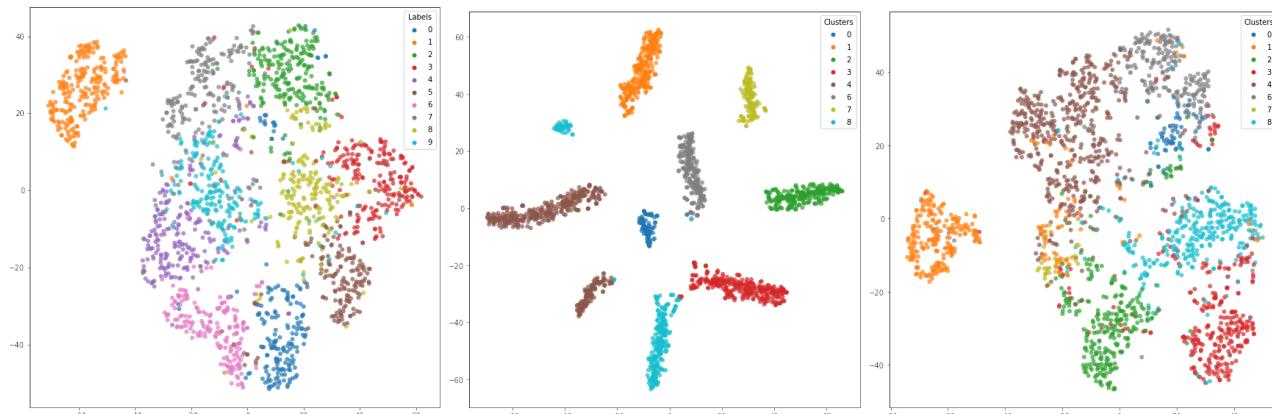


Figura 5.1: Comparació dels models VAE+DEC. Representació t-SNE de la CR del model VAE pre-entrenat (**esquerra**), del model combinat que perd el descodificador (**centre**) i del model combinat que manté el descodificador (**dreta**). S'observa una clara millora en la QC obtinguda pel model que perd el descodificador.

manté la funció de cost de regularització durant l'entrenament obté clústers molt més laxos i que presenten major solapament (el que es tradueix en puntuacions més baixes per la mètrica si-lueta) en comparació amb el model que descarta el descodificador, i per tant s'entrena únicament amb la funció de cost de *clustering*.

Aquest problema es pot entendre des d'un punt de vista de les funcions de cost, utilitzades per entrenar el model. El model VAE+DEC que manté el descodificador, durant l'entrenament se li aplica una funció de cost que es pot dividir en tres parts: funció de cost de reconstrucció, de regularització i de clustering.

La funció de cost de regularització minimitza la distància de la distribució obtinguda a la capa latent amb la distribució objectiu (normal multivariant), mentre que la funció de cost de *clustering* minimitza la distància de cada punt al centroide del seu clúster, alhora que maximitza la distància als centroides de la resta de clústers. Al maximitzar la distància entre centroides, la distribució s'allunya d'una distribució normal, i viceversa. En conseqüència, minimitzar una funció de cost penalitza a l'altra i el model resultat no aconsegueix optimitzar cap de les dues.

Model VaDE El model VaDE resol aquest problema assignant una distribució independent a cada grup. D'aquesta manera, augmentar la distància entre clústers no penalitza a la funció de regularització i es poden optimitzar les dues simultàniament.

Així, s'obté un model de *deep clustering* que manté la capacitat generativa del model VAE. A la figura 4.1 es mostra gràficament la diferència entre les representacions obtingudes pels dos models sobre les dades MNIST.

No s'han explorat les característiques generatives del model VaDE, que juntament amb la seva capacitat d'obtenir una bona QC el converteixen en un model interessant.

Dades metabolòmiques i interpretació biològica Al aplicar les tècniques sobre les dades metabolòmiques, el rendiment de les tècniques de *deep clustering* no sembla millorar respecte a les tècniques clàssiques. Segurament això és degut a que aquestes dades presenten una estructura molt més complexa que és difícil de modelar inclús pels models de *deep clustering*.

A més, per la seva naturalesa es disposa d'un número de dades molt petit que els que es

soLEN utilitzar per entrenar models de *deep learning*, com per exemple les dades MNIST. Això a provocat també la mida dels models implementats s'hagin acotat molt, limitant així la seva capacitat d'aprenentatge.

S'ha observat que els clústers trobats per les diferents tècniques (i números de clústers) presenten en general un bon solapament, és a dir que agrupen les observacions de manera similar. Això seria indicatiu de que existeix algun tipus d'estructura latent a les dades.

La major dificultat s'ha presentat alhora de donar una interpretació als grups trobats, donat que presenten distribucions multivariants complexes. Encara que les dades DCH-NG mostren uns resultats que apunten a una possible interpretació biològica, seria necessari un estudi addicional per confirmar-ho.

Un estudi més profund d'aquestes distribucions podria ajudar a trobar un criteri més clar amb el que valorar l'efectivitat de les diferents tècniques.

Per tant, es considera que des del punt de vista de la interpretació biològica els resultats són poc concloents i requeririen un estudi més profund.

Capítol 6

Valoració econòmica

La realització d'aquest TFM ha requerit d'una única despesa econòmica: una subscripció Pa-perspace Gradient, el servei de computació *on-line* que es menciona a la secció 3.6.

Encara que aquesta plataforma ofereix un servei gratuït, s'ha considerat insuficient ja que no garanteix la disponibilitat de màquines virtuals en qualsevol moment.

S'ha subscript el pla de pagament més econòmic, que ha cobert els requisits de poder de computació i disponibilitat. La subscripció ha tingut un cost de \$8 al mes i s'ha activat durant els mesos de novembre, desembre i generdurant, sumant per tant un total de \$24 (equivalent a aproximadament 22 € en el moment de redactar aquest informe).

Aquest servei només ha sigut necessari ja que no es disposa d'una màquina prou potent per entrenar els models neuronals que s'han implementat.

Donat que el cost mensual d'aquest servei és petit, que no és estrictament necessari si es disposa d'una màquina suficientment potent, i que l'objectiu no ha sigut desenvolupar un producte econòmicament sostenible, si no implementar i avaluar una sèrie de models prototip, es considera que el cost ha sigut perfectament assumible.

Capítol 7

Conclusions i treballs futurs

7.1 Conclusions

De manera sintetitzada, les conclusions que s'han extret en aquest TFM són:

- Els models de *deep clustering* tenen la capacitat de superar el rendiment de les tècniques clàssiques, a costa d'un augment significatiu de la complexitat del model.
- El número reduït d'observacions de les que es sol disposar en conjunts de dades metabòlomiques afecta negativament a la capacitat d'aprenentatge dels models de *deep learning*.
- El model DEC destaca per aconseguir clústers molt estrets, però és molt sensible als paràmetres amb què s'inicialitza.
- El model VaDE aconsegueix una bona QC, i a més té la capacitat de funcionar com un model generatiu.
- L'arquitectura VAE no propicia una bona QC donat que la CR està acotada a una única distribució latent, compartida per totes les observacions.
- Tots els mètodes evaluats sobre les dades metabòlomiques (K-means, GMM, Aglo., DEC, VaDE) mostren un alt nivell de solapament dels clústers trobats. És a dir, troben una estructura latent de les dades similar.
- Encara que els clústers obtinguts mostren certes diferències en la distribució multivariant de les covariables d'interès, la seva interpretació és complexa i requeriria d'un estudi posterior.

Els resultats obtinguts es consideren satisfactoris i es considera que s'han aconseguit els objectius inicials.

La implementació dels models de *deep learning* ha resultat una tasca més complicada del que s'esperava, però finalment s'ha aconseguit.

El temps disponible per realitzar aquest TFM ha fet necessari limitar l'anàlisi que s'ha fet dels resultats obtinguts. A la següent secció es presenten una sèrie de idees que no ha donat temps d'estudiar.

7.2 Línies de futur

Durant el desenvolupament d'aquest TFM han sorgit diverses idees que no s'han pogut explorar donada la limitació de temps disponible.

Algunes d'aquestes idees es podrien desenvolupar en estudis posteriors. Es fan les següents propostes:

Dades DCH-NG En l'estudi del conjunt de dades DCH-NG no s'ha tingut en compte que es tracta de mesures repetides en el temps. Concretament, es disposa de tres mesures realitzades sobre els mateixos individus.

Es proposa estudiar les tècniques de *deep clustering* avaluades, entrenant els models amb les dades d'un dels temps iavaluant els resultats amb les dades dels altres dos temps. Fent el mateix per els altres dos grups, es pot realitzar validació creuada.

Una segona proposta és separar les dades en tres subconjunts en funció del temps, aplicar les tècniques de *clustering* sobre cada subconjunt per separat i comparar les assignacions obtingudes. S'esperaria observar major variació entre individus que entre les observacions del mateix individu. Per tant, s'esperaria que trobar un fort solapament entre els clústers: el mateix individu s'hauria d'assignar al mateix clúster en els diferents temps.

Per últim, es proposa estudiar la interpretació biològica dels clústers obtinguts mitjançant un estudi multivariant per mesures repetides.

Exosome Data Challenge Event Es tracta d'un conjunt de dades molt complexe compost per dades de diversos orígens, pel que les possibilitats d'estudi són molt àmplies.

Una aproximació que no s'ha realitzat en aquest TFM per falta de temps és implementar un model basat en un AE que tingui dues entrades. El model es compondria per dos (o múltiples) codificadors paral·lels. Les capes de representació es combinarien concatenant les respectives capes neuronals, i posteriorment es podria aplicar *clustering* mitjançant tècniques clàssiques o un model DEC.

Cada un dels codificadors tindrien com entrada un subconjunt de dades, per exemple les dades metabolòmiques de sèrum i d'orina, les dades metabolòmiques i les dades del exposoma, o algun dels subconjunts de dades òmiques i un dels subconjunts de covariables. D'aquesta manera es podria obtenir informació sobre una estructura latent condicionada simultàniament a diversos conjunts de dades.

Els codificadors es podrien aconseguir entrenant dos AE per separat, i després unint-los, o implementant directament un AE amb dues entrades i dues sortides.

Models generatius Els models VAE i VaDE es caracteritzen per la seva capacitat generativa: es poden generar noves dades artificials prenent mostres de la distribució de la seva capa latent i passant-les pel descodificador del model.

Donat que una de les limitacions de les dades metabolòmiques (i òmiques en general) és que soleten tenir un número reduït de mostres, es proposa estudiar la viabilitat d'utilitzar aquests models com a tècnica d'augment de les dades, ja sigui per realitzar estudis de *clustering* o aplicar altres mètodes d'anàlisi.

Model barreja d'experts Una altra idea que no ha donat temps d'estudiar en aquest TFM és desenvolupar un model tipus *mixture of experts*. Un cop implementats i seleccionats els models individuals de *clustering*, es podria aplicar un mètode per unificar les assignacions de clústers a un sol criteri.

Això podria realitzar-se de manera més o menys senzilla, assignant a cada observació el clúster que mostri més solapament entre les diferents tècniques, o mitjançant una xarxa neuronal més complexa que combini els diversos models en una sola sortida.

Interpretació dels clústers En aquest TFM no s'ha estudiat amb profunditat la possible interpretació biològica dels clústers. Es proposa realitzar un estudi més rigorós analitzant la distribució multivariant diferencial en funció dels clústers obtinguts.

Número de clústers Per últim, es planteja una limitació important de totes les tècniquesavaluades en aquest TFM: la selecció d'un nombre de clústers òptim.

Excepte el mètode aglomeratiu, tots els mètodes utilitzats requereixen fixar prèviament un nombre de clústers. Sovint no és factible conèixer el número de clústers per endavant, precisament perquè l'objectiu de les tècniques de *clustering* és buscar una estructura latent desconeguda.

En aquest sentit, s'ha de manera superficial l'eina *clValid*¹, un paquet de R que permet avaluar simultàniament diverses tècniques de *clustering* clàssiques mesurant diferents mètriques internes, d'estabilitat i biològiques.

Es proposa estudiar el conjunt de dades Exposome Data Challenge Event amb l'objectiu de trobar un número de clústers òptim que admeti la seva interpretació. Posteriorment es podria tornar a avaluar els diferents mètodes en funció del seu alineament amb aquesta interpretació.

7.3 Seguiment de la planificació

Si bé es considera que s'han complert tots els objectius globals i específics especificats a l'inici del projecte, no s'han respectat les dates del pla de treball. Això ha estat causat per dos motius.

Compaginar recerca teòrica i treball pràctic La fase inicial del TFM, en la que es pretenia obtenir una base sólida de coneixements teòrics i pràctics sobre la que desenvolupar la resta del treball, es va definir inicialment com un bloc independent de la resta del projecte.

A mesura que es va anar avançant, i sobretot després de fer les primeres proves d'implementar xarxes neuronals amb certa complexitat amb Keras, es va trobar més productiu compaginar les tasques de formació amb la implementació de les primeres versions dels models de *deep clustering*. Realitzar aquestes dues tasques de manera simultània va ajudar a consolidar els coneixements teòrics i pràctics, i a trobar limitacions dels models que no s'havia tingut en compte inicialment.

Això va tenir l'efecte que, per una banda, s'endarrerí el compliment del primer objectiu específic, alhora que es va avançar l'inici del segon.

¹<https://www.rdocumentation.org/packages/clValid/versions/0.7/topics/clValid>

Augment de complexitat del problema Inicialment, aquest TFM es va titular *Desenvolupament d'un model de deep clustering en dades metabolòmiques utilitzant un Variational Autoencoder*, ja que en una primera presa de contacte amb la literatura es va pensar que el VAE, degut a la seva capacitat com a model generatiu, podia ser una base interessant sobre la que construir un model de *deep clustering*.

Si bé la idea no estava equivocada (el model VaDE parteix precisament d'aquest concepte), la implementació que es va intentar inicialment va resultar problemàtica.

Un cop entès el funcionament de l'arquitectura VAE, i després d'implementar un model i comprovar el seu correcte funcionament, es va procedir a afegir al model una capa de *clustering* similar al model mixte VAE + DEC que es va avaluar amb el conjunt de dades MNIST.

Finalment es va comprendre que aquest model no és adequat per realitzar aquesta tasca, com s'explica al [capítol 5](#).

En resposta a aquest problema, es van estudiar i implementar dos models de *deep clustering* addicionals: DEC i VaDE, el que va suposar una dedicació extra.

Tot això va portar a canviar lleugerament la direcció del TFM: en lloc d'implementar un model VAE per realitzar *clustering* en dades metabolòmiques, s'han implementat diversos models i s'han comparat els resultats obtinguts.

La conseqüència negativa més significativa és que no s'ha disposat del temps que s'hauria desitjat per analitzar més detingudament els resultats i obtenir unes conclusions més interessants.

Glossari i abreviaccions

AAE *adversarial autoencoder*. 18

Acc. exactitud (*accuracy*). 41, 44, 45, 48, 63, 64, 66–73

AE *autoencoder*. 9–12, 16–18, 20–22, 26, 32, 34–37, 40, 49, 54

Aglo. model aglomeratiu. 20, 32, 35–37, 53

AMI informació mútua ajustat (*Adjusted Mutual Information*). 29, 30, 41, 44, 45, 48, 63, 64, 66–73

ARI índex de Rand ajustat (*Adjusted Rand Index*). 29, 30, 41, 44, 45, 48, 63, 64, 66–73

CAE *convolutional autoencoder*. 17, 18

CNN *convolutional neural network*. 16, 17

CR capa de representació. 16, 17, 20, 22, 27, 33, 40, 41, 49, 50, 53

DAE *denoising autoencoder*. 18

DBN *deep belief network*. 16

DCH-NG *Diet, Cancer and Health - Next Generations*. 31, 33, 34, 37, 45–48, 51, 54, 61

DEC *Deep embedded clustering*. 12, 20–22, 32–37, 40–42, 49, 53, 54, 56

ELBO límit inferior de la evidència (*evidence lower bound*). 24–28

FL aprenentatge de característiques (*feature learning*). 14, 16, 17, 22

GAN *generative adversarial network*. 16, 18

GMM model de barreja de gaussianes. 19, 27, 32–37, 40–43, 53

K-means K-means. 19, 32–37, 42, 53

KL Kullback-Leibler. 21, 24, 27

LF característiques apreses (*learned features*). 16, 17, 20–22, 49

LSTM *long short-term memory.* 18

LSTM-AE *long short-term memory autoencoder.* 18

MLP *multilayer perceptron.* 16, 17, 32, 36, 37

MNIST *Modified National Institute of Standards and Technology database.* 30, 32, 33, 41, 49–51, 56

MSE mitjana de quadrats de l'error. 20

PCA anàlisi de components principals. 9, 32, 35–37, 40, 42, 43

QC qualitat de *clustering.* 16, 17, 21, 28, 29, 40, 49, 50, 53

ReLU unitat lineal rectificada, (*rectified linear unit*). 32

SAE *stacked autoencoder.* 18

Sil. coeficient Silueta. 41, 44, 45, 48, 63, 64, 66–73

SSE suma de quadrats de l'error. 19, 20

t-SNE *t-Distributed Stochastic Neighbor Embedding.* 33, 34, 37, 41, 50

TFM treball de final de màster. 9–12, 14, 18, 20, 25, 28, 30, 52–56, 61

VaDE *Variational Deep Embedding.* 12, 20, 27, 28, 32–37, 40, 41, 49, 50, 53, 54, 56

VAE *variational autoencoder.* 12, 18, 20, 22, 23, 25–28, 34, 40, 41, 49, 50, 53, 54, 56

Bibliografia

- [1] M. R. Karim, O. Beyan, A. Zappa, I. G. Costa, D. Rebholz-Schuhmann, M. Cochez, and S. Decker, “Deep learning-based clustering approaches for bioinformatics,” *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 393–415, 2021.
- [2] M. A. Masood, M. N. A. Khan, S. Zulfikar, and A. Bhutto, “Clustering Techniques in Bioinformatics,” in *Modern Education and Computer Science*, vol. 1, pp. 38–46, 2015.
- [3] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, “A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture,” *IEEE Access*, vol. 6, pp. 39501–39514, 2018.
- [4] F. Chollet, “Keras.” <https://github.com/fchollet/keras>, 2015.
- [5] N. Ketkar and J. Moolayil, *Deep Learning with Python*. 2021.
- [6] G. Blekherman, R. Laubenbacher, D. F. Cortes, P. Mendes, F. M. Torti, S. Akman, S. V. Torti, and V. Shulaev, “Bioinformatics tools for cancer metabolomics,” *Metabolomics*, vol. 7, pp. 329–343, sep 2011.
- [7] U. Michelucci, “An Introduction to Autoencoders,” *Preprint at arXiv:2201.03898*, jan 2022.
- [8] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, “Adversarial Autoencoders,” *Preprint at arXiv:1511.05644*, nov 2015.
- [9] J. Xie, R. Girshick, and A. Farhadi, “Unsupervised deep embedding for clustering analysis,” in *33rd International Conference on Machine Learning, ICML 2016*, vol. 1, pp. 740–749, International Machine Learning Society (IMLS), nov 2016.
- [10] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, dec 2014.
- [11] D. P. Kingma and M. Welling, “An introduction to variational autoencoders,” *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.
- [12] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, “Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering,” *Preprint at arXiv:1611.05148*, nov 2016.

- [13] B. Lafabregue, J. Weber, P. Gançarski, and G. Forestier, “End-to-end deep representation learning for time series clustering: a comparative study,” *Data Mining and Knowledge Discovery*, vol. 36, pp. 29–81, jan 2022.
- [14] J.-O. Palacio-Niño and F. Berzal, “Evaluation Metrics for Unsupervised Learning Algorithms,” *Preprint at arXiv:1905.05667*, 2019.
- [15] L. Deng, “The MNIST database of handwritten digit images for machine learning research,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
- [16] L. Maitre, J. B. Guimbaud, C. Warembourg, N. Güil-Oumrait, P. M. Petrone, M. Chadeau-Hyam, M. Vrijheid, X. Basagaña, J. R. Gonzalez, R. Alfano, S. Basu, J. Benavides, L. Broséus, C. Brunius, A. Caceres, M. Carli, R. Cazabet, S. Chattopadhyay, Y. H. Chen, L. Chillrud, D. Conti, C. Gennings, R. Gouripeddi, S. H. Iyer, P. Jedynak, H. Li, G. McGee, V. Midya, S. Mistry, C. Moccia, S. D. Mork, L. J. Pearce, M. Peruzzi, J. M. Pescador, B. Reimann, J. C. Roscoe, X. Shen, N. Stratakis, Z. Wang, C. Wang, D. Wheeler, A. Wilson, Q. Wu, M. Yu, Y. Zhao, F. Zugna, R. Chen, Y. C. Chung, J. Jang, and M. Turyk, “State-of-the-art methods for exposure-health studies: Results from the exposome data challenge event,” *Environment International*, vol. 168, p. 107422, oct 2022.
- [17] K. E. Petersen, J. Halkjær, S. Loft, A. Tjønneland, and A. Olsen, “Cohort profile and representativeness of participants in the Diet, Cancer and Health—Next Generations cohort study,” *European Journal of Epidemiology*, vol. 37, pp. 117–127, jan 2022.
- [18] S. Ghemawat, X. Zheng, S. Moore, M. Abadi, J. Dean, Z. Chen, M. Kudlur, P. Warden, G. Irving, J. Chen, P. Barham, Y. Yu, V. Vasudevan, M. Devin, P. Tucker, A. Davis, B. Steiner, R. Monga, M. Wicke, D. Murray, J. Levenberg, and M. Isard, “TensorFlow: A system for large-scale machine learning,” in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, pp. 265–283, 2016.
- [19] G. Van Rossum, F. L. Drake, C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, *Python 3 Reference Manual*, vol. 585. 2009.
- [20] T. Kluyver, B. Ragan-Kelley, B. Granger, and E. al., “Jupyter Notebooks - a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. ,” pp. 87–90, 2016.

Agraïments

M'agradaria expressar el meu agraïment a Esteban Vegas Lozano, el meu tutor en aquest TFM, pel temps dedicat a llegir esborranys de la memòria i els meus e-mails amb múltiples capítols, així com pel suport que m'ha donat aportant idees molt valuoses i tranquil·litzant-me en els moments complicats en que els resultats no semblaven sortir.

També vull agrair a l'equip de recerca *Biomarkers and Nutritional & Food Metabolomics* del Departament de Nutrició, Ciències de l'Alimentació i Gastronomia de la Universitat de Barcelona que hagi compartit desinteressadament el conjunt de dades DCH-NG.

Apèndix A

Exosome Data Challenge Event: resultats complets

A.1 Taules de mètriques

En les següents pàgines es mostren els resultats complets de les mètriques mesurades en el conjunt de dades Exosome Data Challenge Event.

<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
K-Means	-	-	2	asthma	0.89	0	0	0.1
			2	sex	0.53	0	0	0.1
			3	education	0.51	0	0	0.07
			3	native	0.84	0	0	0.07
			3	parity	0.45	0	0	0.07
			4	birth_weight	0.27	0	0	0.06
			4	iq	0.3	0	0	0.06
			4	behaviour	0.27	0	0	0.06
			6	cohort	0.21	0	0	0.05
			7	age	0.32	0	0	0.05
GMM	-	-	2	asthma	0.89	0	0	0.1
			2	sex	0.53	0	0	0.1
			3	education	0.51	0	0	0.06
			3	native	0.84	0	0	0.06
			3	parity	0.45	0	0	0.06
			4	birth_weight	0.27	0	0	0.06
			4	iq	0.3	0	0	0.06
			4	behaviour	0.27	0	0	0.06
			6	cohort	0.22	0	0	0.04
			7	age	0.32	0	0	0.04
Agglo.	-	-	2	asthma	0.89	0	0	0.09
			2	sex	0.53	0	0	0.09
			3	education	0.51	0	0	0.05
			3	native	0.84	0	0	0.05
			3	parity	0.46	0	0	0.05
			4	birth_weight	0.27	0	0	0.04
			4	iq	0.3	0	0	0.04
			4	behaviour	0.27	0	0	0.04
			6	cohort	0.21	0	0	0.02
			7	age	0.32	0	0	0.02
K-Means	-	-	2	asthma	0.89	0	0	0.13
			2	sex	0.53	0	0	0.13
			3	education	0.51	0	0	0.08
			3	native	0.84	0	0	0.08
			3	parity	0.45	0	0	0.08
			4	birth_weight	0.27	0	0	0.08
			4	iq	0.3	0	0	0.08
			4	behaviour	0.27	0	0	0.08
			6	cohort	0.22	0	0	0.06
			7	age	0.32	0	0	0.06
PCA	GMM	-	2	asthma	0.89	0	0	0.02
			2	sex	0.53	0	0	0.02
			3	education	0.51	0	0	0.02
			3	native	0.84	0	0	0.02
			3	parity	0.45	0	0	0.02
			4	birth_weight	0.28	0	0	0
			4	iq	0.32	0.01	0.01	0
			4	behaviour	0.28	0	0	0
			6	cohort	0.2	0	0	-0.02
			7	age	0.32	0	0	-0.02
Agglo.	-	-	2	asthma	0.89	0	0	0.11
			2	sex	0.53	0	0	0.11
			3	education	0.51	0	0	0.06
			3	native	0.84	0	0	0.06
			3	parity	0.45	0	0	0.06
			4	birth_weight	0.27	0	0	0.05
			4	iq	0.29	0	0	0.05
			4	behaviour	0.27	0	0	0.05
			6	cohort	0.2	0	0	0.03
			7	age	0.32	0	0	0.03

Taula A.1: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt metabòloma, dades originals, tècniques clàssiques).

<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
DEC	DEC	K-Means	2	asthma	0.89	0	0	0.7
			2	sex	0.53	0	0	0.7
			3	education	0.51	0	0	0.64
			3	native	0.84	0	0	0.64
			3	parity	0.45	0	0	0.64
			4	birth_weight	0.27	0	0	0.45
			4	iq	0.29	0	0	0.45
			4	behaviour	0.27	0	0	0.45
			4	bmi	0.69	0	0	0.45
			6	cohort	0.21	0	0	0.4
			7	age	0.32	0	0	0.42
VaDE	VaDE	GMM	2	asthma	0.89	0	0	0.29
			2	sex	0.53	0	0	0.29
			3	education	0.51	0	0	0.16
			3	native	0.84	0	0	0.16
			3	parity	0.45	0	0	0.16
			4	birth_weight	0.28	0	0	0.15
			4	iq	0.3	0	0	0.15
			4	behaviour	0.27	0	0	0.15
			4	bmi	0.69	0	0	0.15
			6	cohort	0.21	0	0	0.11
			7	age	0.32	0	0	0.09
DEC (D.A.)	DEC	K-Means	2	asthma	0.89	0	0	0.95
			2	sex	0.53	0	0	0.95
			3	education	0.51	0	0	0.91
			3	native	0.84	0	0	0.91
			3	parity	0.45	0	0	0.91
			4	birth_weight	0.27	0	0	0.77
			4	iq	0.31	0.01	0	0.77
			4	behaviour	0.3	0	0	0.77
			4	bmi	0.69	0	0	0.77
			6	cohort	0.2	0	0	0.77
			7	age	0.32	0	0.01	0.71
VaDE (D.A.)	VaDE	GMM	2	asthma	0.89	0	0	0.35
			2	sex	0.53	0	0	0.35
			3	education	0.51	0	0	0.18
			3	native	0.84	0	0	0.18
			3	parity	0.47	0.01	0	0.18
			4	birth_weight	0.28	0	0	0.18
			4	iq	0.31	0	0	0.18
			4	behaviour	0.27	0	0	0.18
			4	bmi	0.69	0	0	0.18
			6	cohort	0.22	0	0	0.1
			7	age	0.32	0	0	0.07
DEC (conv. 1D)	DEC	K-Means	2	asthma	0.89	0	0	0.64
			2	sex	0.53	0	0	0.64
			3	education	0.51	0	0	0.59
			3	native	0.84	0	0	0.59
			3	parity	0.45	0	0	0.59
			4	birth_weight	0.25	0	0	1
			4	iq	0.28	0	0	1
			4	behaviour	0.25	0	0	1
			4	bmi	0.69	0	0	1
			6	cohort	0.2	0	0	0.28
			7	age	0.32	0	0	0.22
VaDE (conv. 1D)	VaDE	GMM	2	asthma	0.89	0	0	0.38
			2	sex	0.53	0	0	0.38
			3	education	0.51	0	0	0.2
			3	native	0.84	0	0	0.2
			3	parity	0.46	0	0	0.2
			4	birth_weight	0.27	0	0	0.18
			4	iq	0.31	0.01	0	0.18

			4	behaviour	0.27	0	0	0.18
			4	bmi	0.69	0	0	0.18
			6	cohort	0.21	0	0	0.12
			7	age	0.32	0	0	0.09
<hr/>								
			2	asthma	0.89	0	0	0.6
			2	sex	0.53	0	0	0.6
			3	education	0.51	0	0	0.56
			3	native	0.84	0	0	0.56
			3	parity	0.45	0	0	0.56
			DEC (conv. 2D)	birth_weight	0.27	0	0	0.39
			DEC	4	iq	0.3	0	0.39
			K-Means	4	behaviour	0.27	0	0.39
				4	bmi	0.69	0	0.39
				6	cohort	0.2	0	0.33
				7	age	0.32	0	0.29

Taula A.2: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt metaboloma, dades sense originals, tècniques de *deep clustering*).

<i>Feature learning</i>	<i>Clustering</i>	<i>Inicialització</i>	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
K-Means	-		2	asthma	0.89	0	0	0.08
			2	sex	0.53	0	0	0.08
			3	education	0.51	0	0	0.08
			3	native	0.84	0	0	0.08
			3	parity	0.45	0	0	0.08
			4	birth_weight	0.26	0	0	0.07
			4	iq	0.3	0	0	0.07
			4	behaviour	0.29	0	0	0.07
			6	cohort	0.21	0	0	0.06
			7	age	0.32	0	0	0.06
GMM	-		2	asthma	0.89	0	0	0.06
			2	sex	0.53	0	0	0.06
			3	education	0.51	0	0	0.05
			3	native	0.84	0	0	0.05
			3	parity	0.45	0	0	0.05
			4	birth_weight	0.28	0	0	0.04
			4	iq	0.3	0	0	0.04
			4	behaviour	0.28	0	0	0.04
			6	cohort	0.21	0	0	0.03
			7	age	0.32	0	0	0.04
Agglo.	-		2	asthma	0.89	0	0	0.1
			2	sex	0.53	0	0	0.1
			3	education	0.51	0	0	0.06
			3	native	0.84	0	0	0.06
			3	parity	0.45	0	0	0.06
			4	birth_weight	0.26	0	0	0.05
			4	iq	0.3	0	0	0.05
			4	behaviour	0.28	0	0	0.05
			6	cohort	0.2	0	0	0.03
			7	age	0.32	0	0	0.03
K-Means	-		2	asthma	0.89	0	0	0.09
			2	sex	0.53	0	0	0.09
			3	education	0.51	0	0	0.09
			3	native	0.84	0	0	0.09
			3	parity	0.45	0	0	0.09
			4	birth_weight	0.27	0	0	0.08
			4	iq	0.3	0	0	0.08
			4	behaviour	0.29	0	0	0.08
			6	cohort	0.21	0	0	0.07
			7	age	0.32	0	0	0.07
PCA	GMM		2	asthma	0.89	0	0	0.06
			2	sex	0.53	0	0	0.06
			3	education	0.51	0	0	0.05
			3	native	0.84	0	0	0.05
			3	parity	0.45	0	0	0.05
			4	birth_weight	0.27	0	0	0.06
			4	iq	0.3	0	0	0.06
			4	behaviour	0.28	0	0	0.06
			6	cohort	0.2	0	0	0.01
			7	age	0.32	0	0	0.03
Agglo.	-		2	asthma	0.89	0	0	0.11
			2	sex	0.53	0	0	0.11
			3	education	0.51	0	0	0.07
			3	native	0.84	0	0	0.07
			3	parity	0.45	0	0	0.07
			4	birth_weight	0.27	0	0	0.05
			4	iq	0.33	0.01	0.01	0.05
			4	behaviour	0.28	0	0	0.05
			6	cohort	0.21	0	0	0.06
			7	age	0.32	0	0	0.04

Taula A.3: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt metabòloma, subset de variables amb major variància, tècniques clàssiques).

<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
DEC	DEC	K-Means	2	asthma	0.89	0	0	0.69
			2	sex	0.53	0	0	0.69
			3	education	0.51	0	0	0.54
			3	native	0.84	0	0	0.54
			3	parity	0.45	0	0	0.54
			4	birth_weight	0.27	0	0	0.48
			4	iq	0.31	0	0	0.48
			4	behaviour	0.27	0	0	0.48
			4	bmi	0.69	0	0	0.48
			6	cohort	0.21	0	0	0.4
VaDE	VaDE	GMM	7	age	0.32	0	0	0.22
			2	asthma	0.89	0	0	0.31
			2	sex	0.53	0	0	0.31
			3	education	0.51	0	0	0.22
			3	native	0.84	0	0	0.22
			3	parity	0.45	0	0	0.22
			4	birth_weight	0.28	0	0	0.13
			4	iq	0.3	0	0	0.13
			4	behaviour	0.28	0	0	0.13
			4	bmi	0.69	0	0	0.13
DEC (D.A.)	DEC	K-Means	6	cohort	0.21	0	0	0.13
			7	age	0.32	0	0	0
			2	asthma	0.89	0	0	0.93
			2	sex	0.53	0	0	0.93
			3	education	0.51	0	0	0.87
			3	native	0.84	0	0	0.87
			3	parity	0.45	0	0	0.87
			4	birth_weight	0.28	0	0	0.71
			4	iq	0.3	0.01	0	0.71
			4	behaviour	0.27	0	0	0.71
VaDE (D.A.)	VaDE	GMM	4	bmi	0.69	0	0	0.71
			6	cohort	0.21	0	0	0.61
			7	age	0.32	0	0	0.57
			2	asthma	0.89	0	0	0.48
			2	sex	0.54	0	0	0.48
			3	education	0.51	0	0	0.32
			3	native	0.84	0	0	0.32
			3	parity	0.45	0	0	0.32
			4	birth_weight	0.27	0	0	0.24
			4	iq	0.3	0	0	0.24

Taula A.4: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt metabòloma, subset de variables amb major variància, tècniques de *deep clustering*).

<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
K-Means	-	-	2	asthma	0.89	0	0	0.12
			2	sex	0.53	0	0	0.12
			3	education	0.51	0	0	0.07
			3	native	0.84	0	0	0.07
			3	parity	0.45	0	0	0.07
			4	birth_weight	0.28	0	0	0.06
			4	iq	0.29	0	0	0.06
			4	behaviour	0.27	0	0	0.06
			6	cohort	0.21	0	0	0.05
			7	age	0.32	0	0	0.04
GMM	-	-	2	asthma	0.89	0	0	0.12
			2	sex	0.53	0	0	0.12
			3	education	0.51	0	0	0.06
			3	native	0.84	0	0	0.06
			3	parity	0.45	0	0	0.06
			4	birth_weight	0.29	0	0	0.05
			4	iq	0.29	0	0	0.05
			4	behaviour	0.26	0	0	0.05
			6	cohort	0.22	0	0	0.05
			7	age	0.32	0	0	0.04
Agglo.	-	-	2	asthma	0.89	0	0	0.1
			2	sex	0.53	0	0	0.1
			3	education	0.51	0	0	0.06
			3	native	0.84	0	0	0.06
			3	parity	0.45	0	0	0.06
			4	birth_weight	0.27	0	0	0.05
			4	iq	0.29	0	0	0.05
			4	behaviour	0.26	0	0	0.05
			6	cohort	0.21	0	0	0.02
			7	age	0.32	0	0	0.02
K-Means	-	-	2	asthma	0.89	0	0	0.14
			2	sex	0.53	0	0	0.14
			3	education	0.51	0	0	0.08
			3	native	0.84	0	0	0.08
			3	parity	0.45	0	0	0.08
			4	birth_weight	0.27	0	0	0.08
			4	iq	0.29	0	0	0.08
			4	behaviour	0.27	0	0	0.08
			6	cohort	0.21	0	0	0.07
			7	age	0.32	0	0	0.06
PCA	GMM	-	2	asthma	0.89	0	0	0.03
			2	sex	0.53	0	0	0.03
			3	education	0.51	0	0	0.03
			3	native	0.84	0	0	0.03
			3	parity	0.46	0	0	0.03
			4	birth_weight	0.28	0	0	0.02
			4	iq	0.29	0	0	0.02
			4	behaviour	0.26	0	0	0.02
			6	cohort	0.2	0	0	0
			7	age	0.32	0	0	-0.01
Agglo.	-	-	2	asthma	0.89	0	0	0.12
			2	sex	0.53	0	0	0.12
			3	education	0.51	0	0	0.07
			3	native	0.84	0	0	0.07
			3	parity	0.45	0	0	0.07
			4	birth_weight	0.28	0	0	0.06
			4	iq	0.3	0	0	0.06
			4	behaviour	0.26	0	0	0.06
			6	cohort	0.21	0	0	0.04
			7	age	0.32	0	0	0.04

Taula A.5: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt metabòloma, dades corregides, tècniques clàssiques).

<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
DEC	DEC	K-Means	2	asthma	0.89	0	0	0.69
			2	sex	0.53	0	0	0.69
			3	education	0.51	0	0	0.43
			3	native	0.84	0	0	0.43
			3	parity	0.45	0	0	0.43
			4	birth_weight	0.27	0	0	0.44
			4	iq	0.3	0	0	0.44
			4	behaviour	0.28	0	0	0.44
			4	bmi	0.69	0	0	0.44
			6	cohort	0.21	0	0	0.61
			7	age	0.32	0	0	0.32
VaDE	VaDE	GMM	2	asthma	0.89	0	0	0.35
			2	sex	0.53	0	0	0.35
			3	education	0.51	0	0	0.24
			3	native	0.84	0	0	0.24
			3	parity	0.45	0	0	0.24
			4	birth_weight	0.27	0	0	0.18
			4	iq	0.3	0	0	0.18
			4	behaviour	0.27	0	0	0.18
			4	bmi	0.69	0	0	0.18
			6	cohort	0.2	0	0	0.16
			7	age	0.32	0.01	0	0.11

Taula A.6: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt metaboloma, dades corregides, tècniques de *deep clustering*).

<i>Feature learning</i>	<i>Clustering</i>	<i>Inicialització</i>	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
K-Means	-		2	asthma	0.89	0	0	0.1
			2	sex	0.53	0	0	0.1
			3	education	0.53	0.05	0.02	0.09
			3	native	0.84	0	0	0.09
			3	parity	0.45	0	0	0.09
			4	birth_weight	0.33	0.02	0.03	0.11
			4	iq	0.43	0.1	0.14	0.11
			4	behaviour	0.34	0.02	0.05	0.11
			6	cohort	0.99	0.98	0.97	0.12
			7	age	0.59	0.48	0.58	0.11
GMM	-		2	asthma	0.89	0	0	0.07
			2	sex	0.53	0	0	0.07
			3	education	0.53	0.05	0.02	0.09
			3	native	0.84	0	0	0.09
			3	parity	0.45	0	0	0.09
			4	birth_weight	0.33	0.02	0.03	0.11
			4	iq	0.47	0.2	0.26	0.11
			4	behaviour	0.31	0.01	0.02	0.11
			6	cohort	0.99	0.99	0.98	0.12
			7	age	0.59	0.48	0.58	0.11
Agglo.	-		2	asthma	0.89	0	0	0.09
			2	sex	0.53	0	0	0.09
			3	education	0.56	0.06	0.03	0.1
			3	native	0.84	0	0	0.1
			3	parity	0.47	0.01	0.01	0.1
			4	birth_weight	0.33	0.02	0.03	0.11
			4	iq	0.43	0.1	0.14	0.11
			4	behaviour	0.34	0.02	0.05	0.11
			6	cohort	0.99	0.98	0.97	0.12
			7	age	0.59	0.48	0.58	0.1
K-Means	-		2	asthma	0.89	0	0	0.12
			2	sex	0.53	0	0	0.12
			3	education	0.53	0.05	0.02	0.12
			3	native	0.84	0	0	0.12
			3	parity	0.45	0	0	0.12
			4	birth_weight	0.33	0.02	0.03	0.14
			4	iq	0.43	0.1	0.14	0.14
			4	behaviour	0.34	0.02	0.05	0.14
			6	cohort	0.99	0.98	0.97	0.16
			7	age	0.58	0.47	0.57	0.14
PCA	GMM		2	asthma	0.89	0	0	0.05
			2	sex	0.53	0	0	0.05
			3	education	0.54	0.02	0.03	0.09
			3	native	0.84	0	0	0.09
			3	parity	0.47	0.01	0.01	0.09
			4	birth_weight	0.32	0.02	0.02	0.13
			4	iq	0.43	0.09	0.14	0.13
			4	behaviour	0.34	0.02	0.05	0.13
			6	cohort	0.99	0.97	0.96	0.16
			7	age	0.58	0.47	0.56	0.13
Agglo.	-		2	asthma	0.89	0	0	0.12
			2	sex	0.53	0	0	0.12
			3	education	0.56	0.06	0.03	0.13
			3	native	0.84	0	0	0.13
			3	parity	0.47	0.01	0.01	0.13
			4	birth_weight	0.33	0.02	0.03	0.13
			4	iq	0.43	0.1	0.14	0.13
			4	behaviour	0.34	0.02	0.05	0.13
			6	cohort	0.99	0.97	0.96	0.16
			7	age	0.58	0.47	0.58	0.14

Taula A.7: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt exposoma, dades sense corregir, tècniques clàssiques).

<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
DEC	DEC	K-Means	2	asthma	0.89	0	0	0.89
			2	sex	0.53	0	0	0.89
			3	education	0.51	0.13	0.11	0.8
			3	native	0.84	0	0	0.8
			3	parity	0.45	0	0	0.8
			4	birth_weight	0.33	0.03	0.03	0.77
			4	iq	0.43	0.11	0.17	0.77
			4	behaviour	0.33	0.03	0.03	0.77
			4	bmi	0.69	0	0	0.77
			6	cohort	0.99	0.98	0.97	0.74
			7	age	0.58	0.47	0.56	0.73
VaDE	VaDE	GMM	2	asthma	0.89	0	0	0.52
			2	sex	0.53	0	0	0.52
			3	education	0.55	0.07	0.03	0.49
			3	native	0.84	0	0	0.49
			3	parity	0.45	0	0	0.49
			4	birth_weight	0.32	0.01	0.02	0.36
			4	iq	0.41	0.07	0.11	0.36
			4	behaviour	0.34	0.03	0.04	0.36
			4	bmi	0.69	0	0	0.36
			6	cohort	0.82	0.71	0.82	0.45
			7	age	0.57	0.46	0.54	0.58

Taula A.8: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt exposoma, dades sense corregir, tècniques de *deep clustering*).

<i>Feature learning</i>	<i>Clustering</i>	<i>Inicialització</i>	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
K-Means	-		2	asthma	0.89	0	0	0.03
			2	sex	0.53	0	0	0.03
			3	education	0.51	0	0	0.03
			3	native	0.84	0	0	0.03
			3	parity	0.45	0	0	0.03
			4	birth_weight	0.29	0	0	0.03
			4	iq	0.31	0	0	0.03
			4	behaviour	0.29	0.01	0.01	0.03
			6	cohort	0.23	0	0	0.02
			7	age	0.33	0.01	0	0.02
GMM	-		2	asthma	0.89	0	0	0.01
			2	sex	0.53	0	0	0.01
			3	education	0.51	0	0	0.01
			3	native	0.84	0	0	0.01
			3	parity	0.45	0	0	0.01
			4	birth_weight	0.31	0.01	0.01	-0.01
			4	iq	0.32	0.01	0.02	-0.01
			4	behaviour	0.31	0.01	0.01	-0.01
			6	cohort	0.28	0.07	0.09	0
			7	age	0.4	0.11	0.11	-0.01
Agglo.	-		2	asthma	0.89	0	0	0.04
			2	sex	0.54	0	0	0.04
			3	education	0.51	0	0	0.02
			3	native	0.84	0	0	0.02
			3	parity	0.45	0	0	0.02
			4	birth_weight	0.28	0	0	0.01
			4	iq	0.31	0	0	0.01
			4	behaviour	0.28	0	0	0.01
			6	cohort	0.22	0	0	0.01
			7	age	0.32	0.01	0	0.01
K-Means	-		2	asthma	0.89	0	0	0.04
			2	sex	0.53	0	0	0.04
			3	education	0.51	0	0	0.03
			3	native	0.84	0	0	0.03
			3	parity	0.45	0	0	0.03
			4	birth_weight	0.28	0	0	0.03
			4	iq	0.31	0	0	0.03
			4	behaviour	0.29	0	0.01	0.03
			6	cohort	0.23	0	0	0.02
			7	age	0.32	0	0.01	0.02
PCA	GMM		2	asthma	0.89	0	0	0.01
			2	sex	0.53	0	0	0.01
			3	education	0.51	0	0	0.04
			3	native	0.84	0	0	0.04
			3	parity	0.45	0	0	0.04
			4	birth_weight	0.27	0	0	0.03
			4	iq	0.3	0	0	0.03
			4	behaviour	0.27	0	0	0.03
			6	cohort	0.21	0	0	0.01
			7	age	0.32	0	0	-0.01
Agglo.	-		2	asthma	0.89	0	0	0.04
			2	sex	0.53	0	0	0.04
			3	education	0.51	0	0	0.02
			3	native	0.84	0	0	0.02
			3	parity	0.45	0	0	0.02
			4	birth_weight	0.27	0	0	0.01
			4	iq	0.3	0	0	0.01
			4	behaviour	0.28	0	0	0.01
			6	cohort	0.22	0	0	0.01
			7	age	0.33	0	0.01	0.01

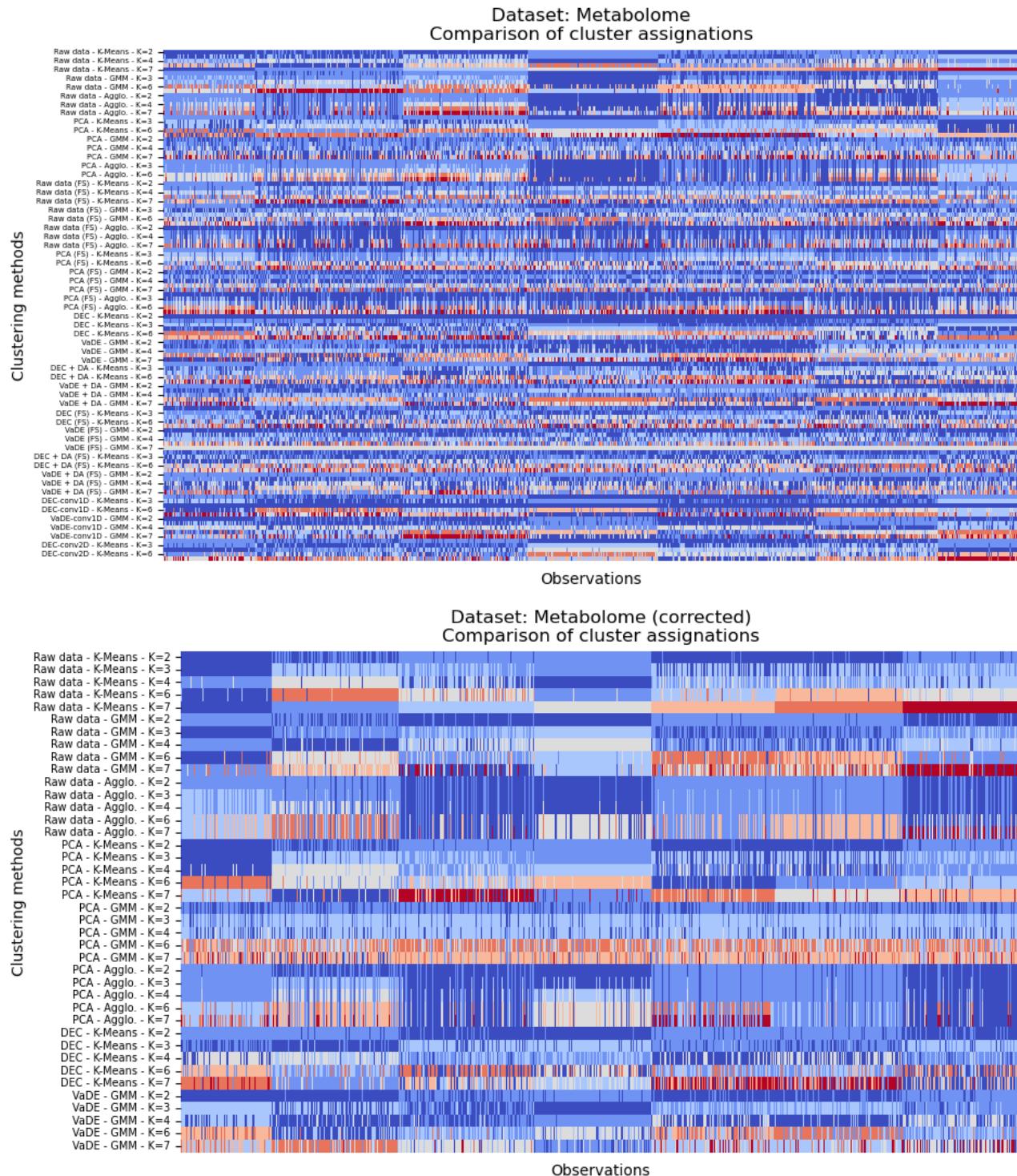
Taula A.9: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt exposoma, dades corregides, tècniques clàssiques).

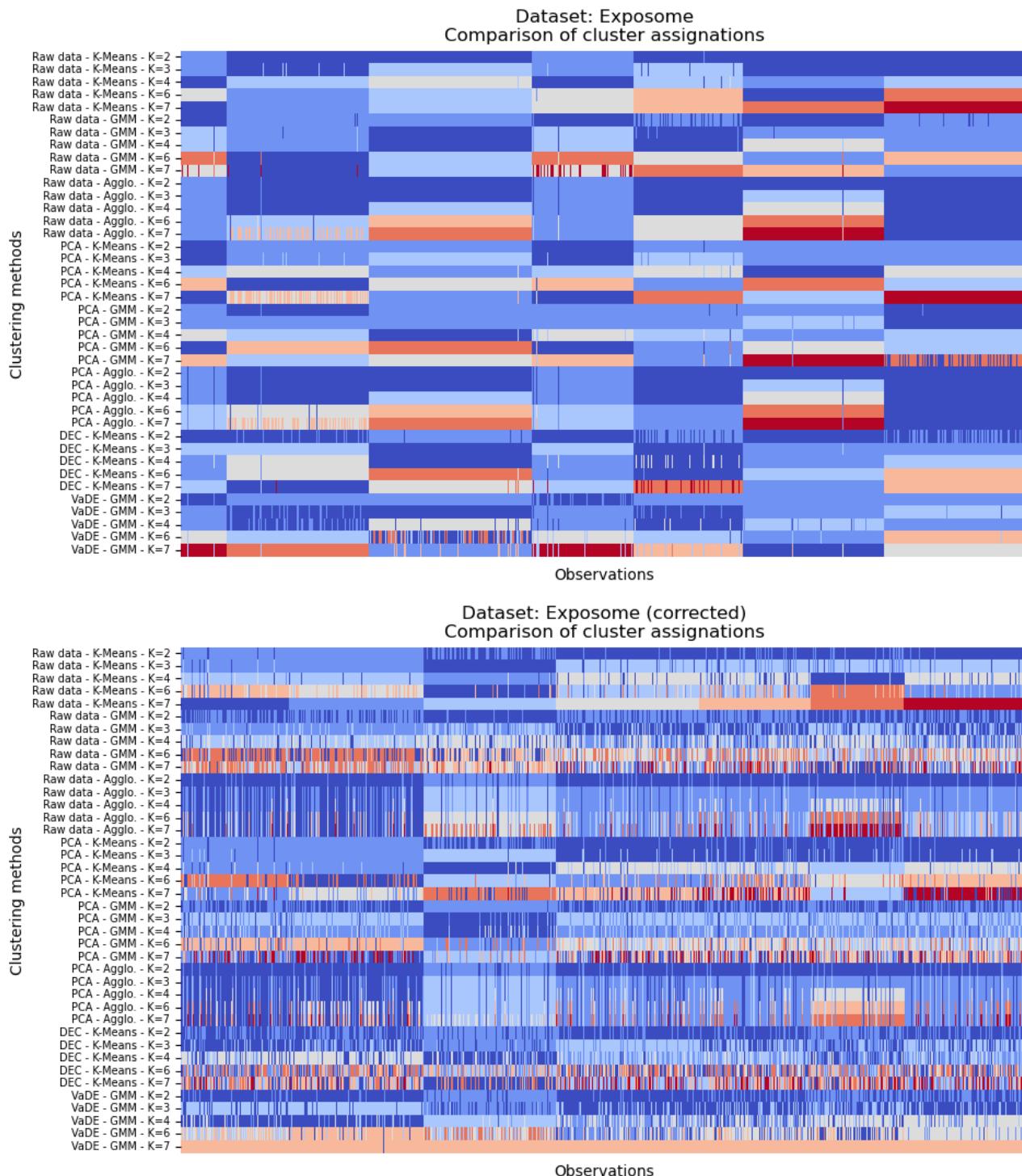
<i>Feature learning</i>	<i>Clustering</i>	Inicialització	Núm. clústers	Covariable referència	Acc.	ARI	AMI	Sil.
DEC	DEC	K-Means	2	asthma	0.89	0	0	0.81
			2	sex	0.53	0	0	0.81
			3	education	0.51	0	0	0.67
			3	native	0.84	0	0	0.67
			3	parity	0.45	0	0	0.67
			4	birth_weight	0.28	0	0	0.57
			4	iq	0.31	0	0	0.57
			4	behaviour	0.28	0	0	0.57
			4	bmi	0.69	0	0	0.57
			6	cohort	0.21	0	0	0.58
			7	age	0.32	0	0	0.59
VaDE	VaDE	GMM	2	asthma	0.89	0	0	0.42
			2	sex	0.53	0	0	0.42
			3	education	0.51	0	0	0.27
			3	native	0.84	0	0	0.27
			3	parity	0.45	0	0	0.27
			4	birth_weight	0.28	0	0	0.2
			4	iq	0.3	0.01	0	0.2
			4	behaviour	0.27	0	0	0.2
			4	bmi	0.69	0	0	0.2
			6	cohort	0.22	0	0	0.08
			7	age	0.32	0	0	0.39

Taula A.10: Resultats obtinguts sobre el conjunt de dades Exposome Data Challenge Event (subconjunt exposoma, dades corregides, tècniques de *deep clustering*).

A.2 Heatmaps

A continuació es mostren les comparacions de les assignacions de clústers per cada mostra, agrupades en funció del conjunt de dades sobre el que s'han aplicat els mètodes de *clustering*.



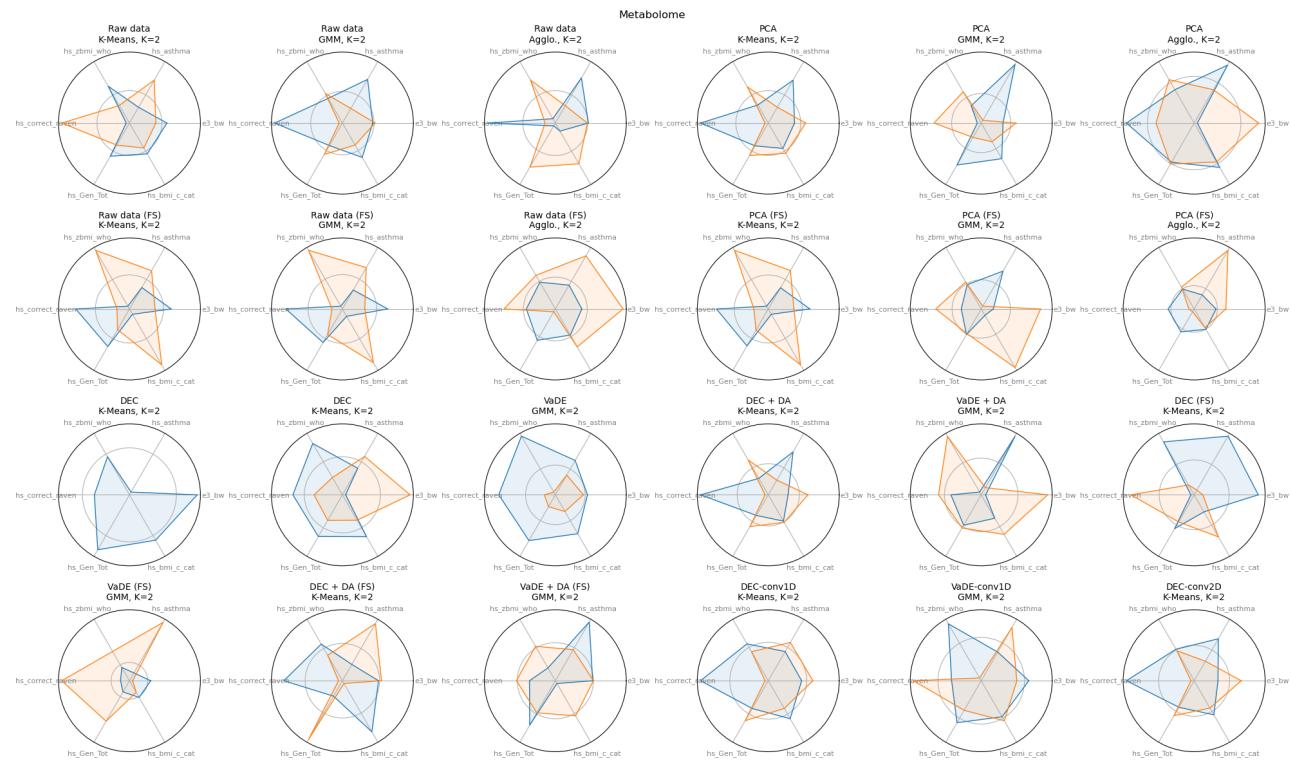


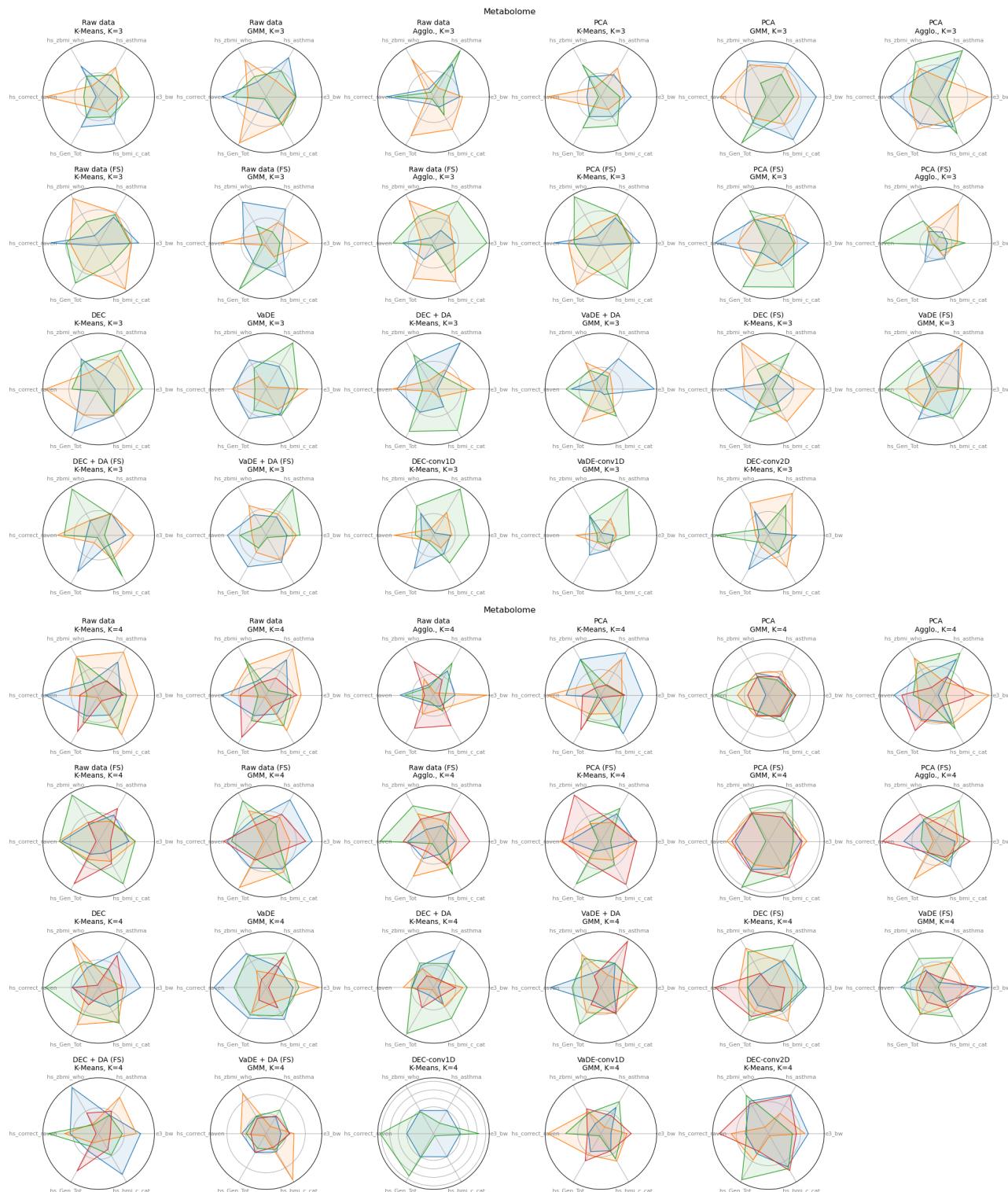
A.3 Gràfiques radials

A continuació es mostren les gràfiques radials que representen la distribució multivariant de les dades del fenotip i covariables, en segons dels clústers trobats. S'han ordenat en funció del conjunt de dades sobre el que s'han aplicat els mètodes de *clustering*.

A.3.1 Conjunt de dades: metaboloma

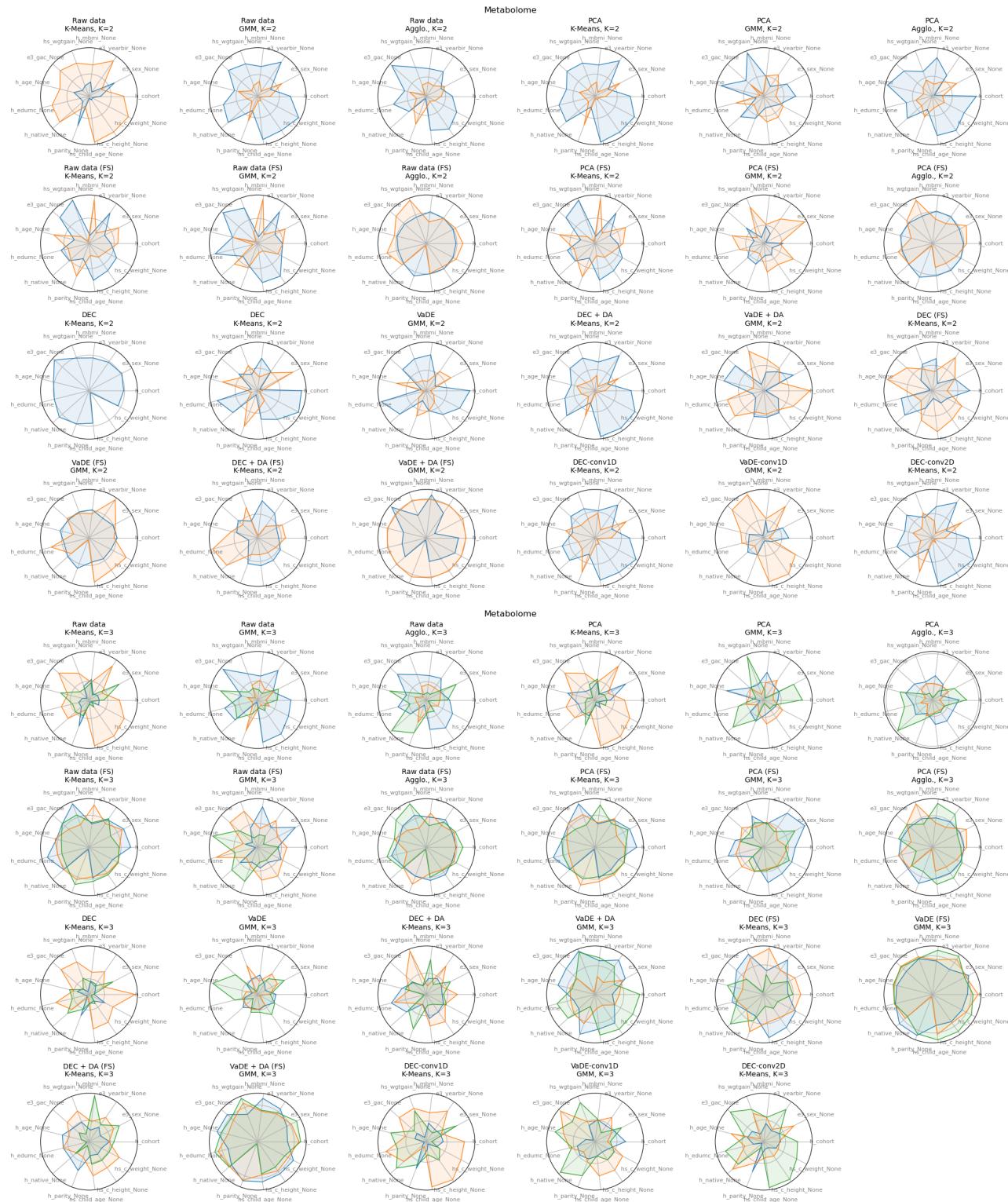
Dades fenotip: distribució multivariant en funció dels clústers.

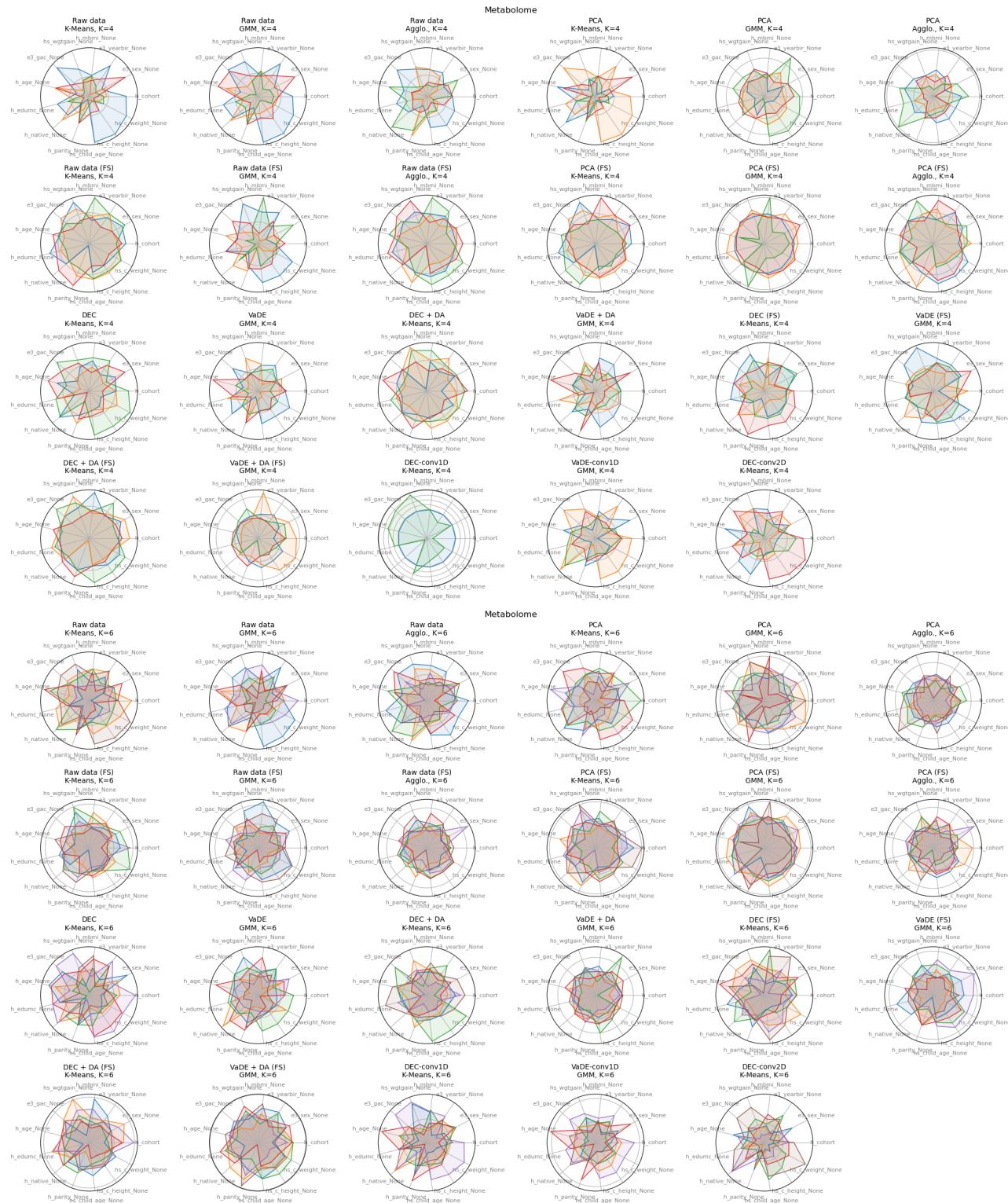


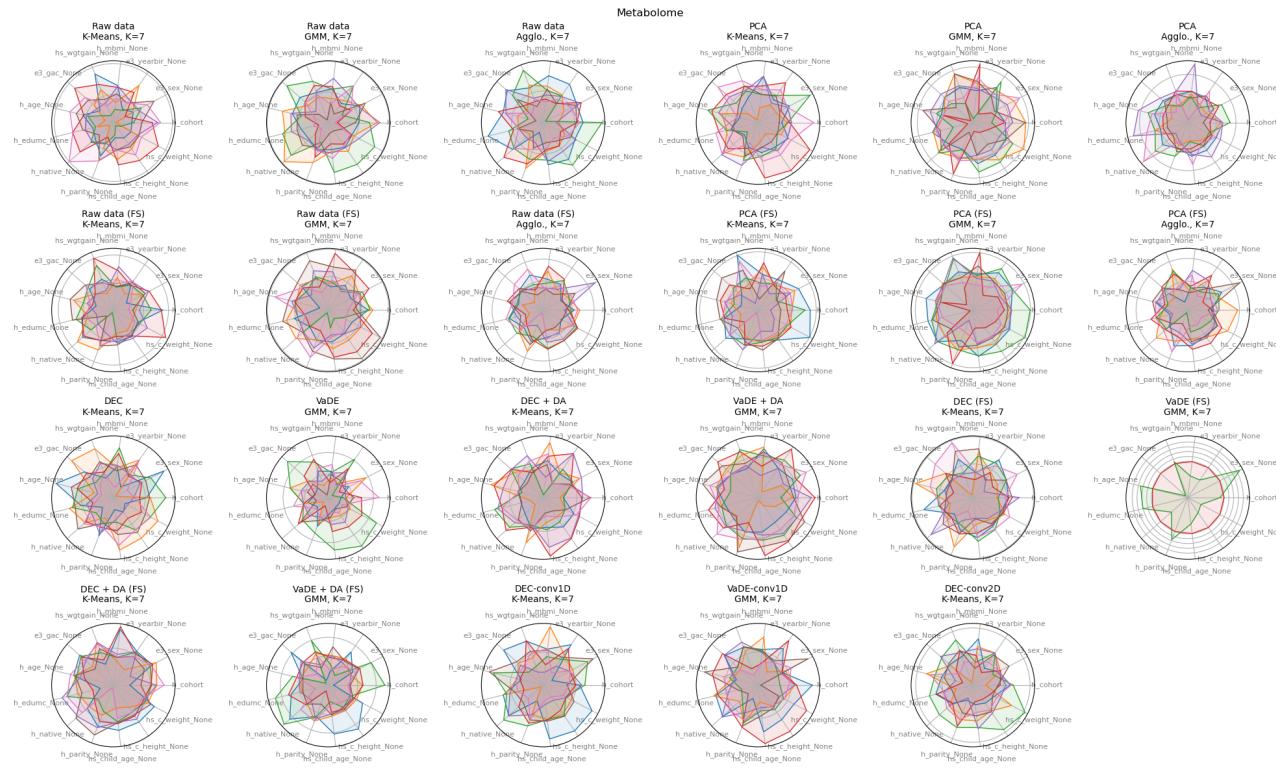




Dades covariables: distribució multivariant en funció dels clústers.



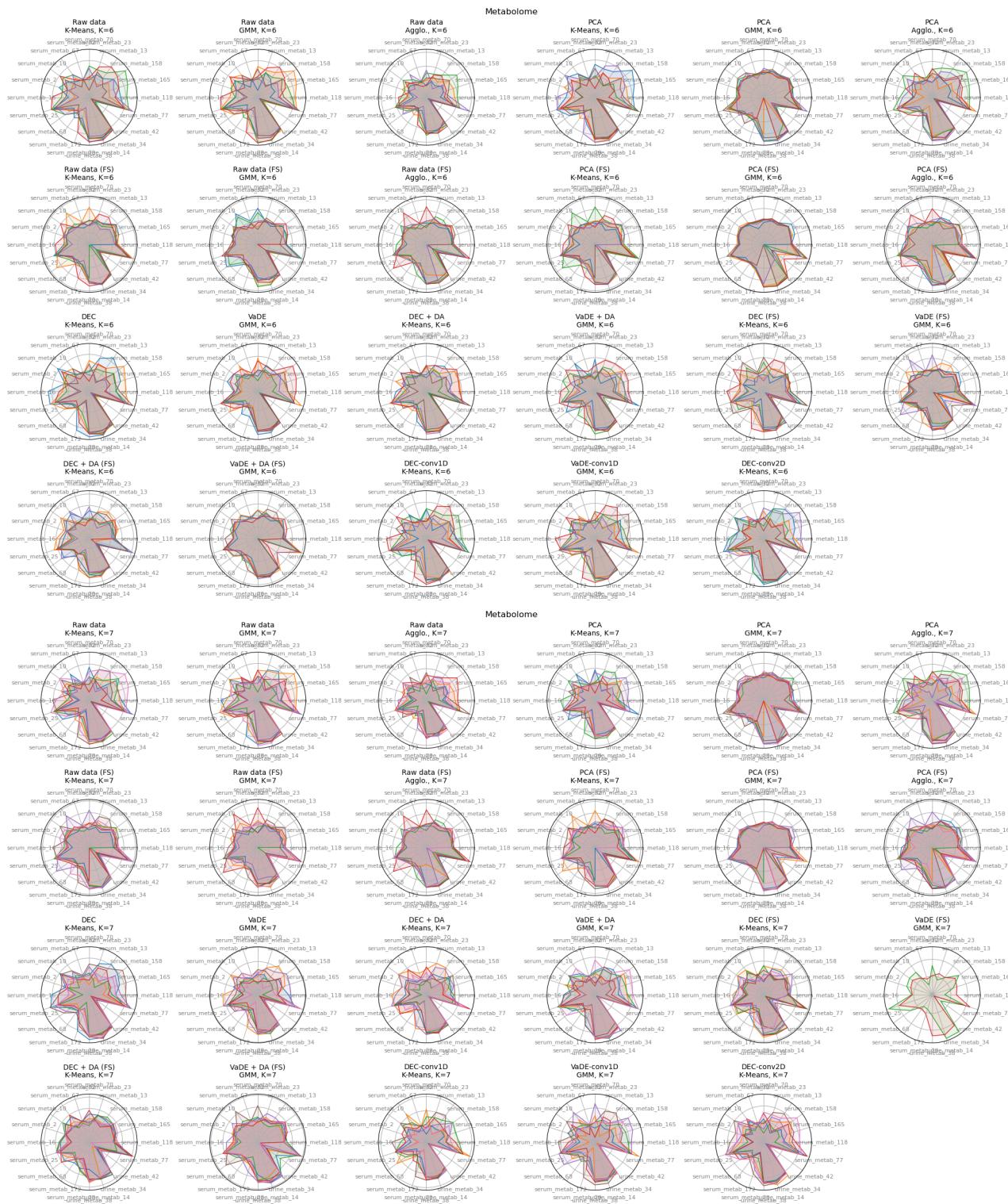




Dades metaboloma: distribució multivariant en funció dels clústers (20 variables amb major variabilitat).

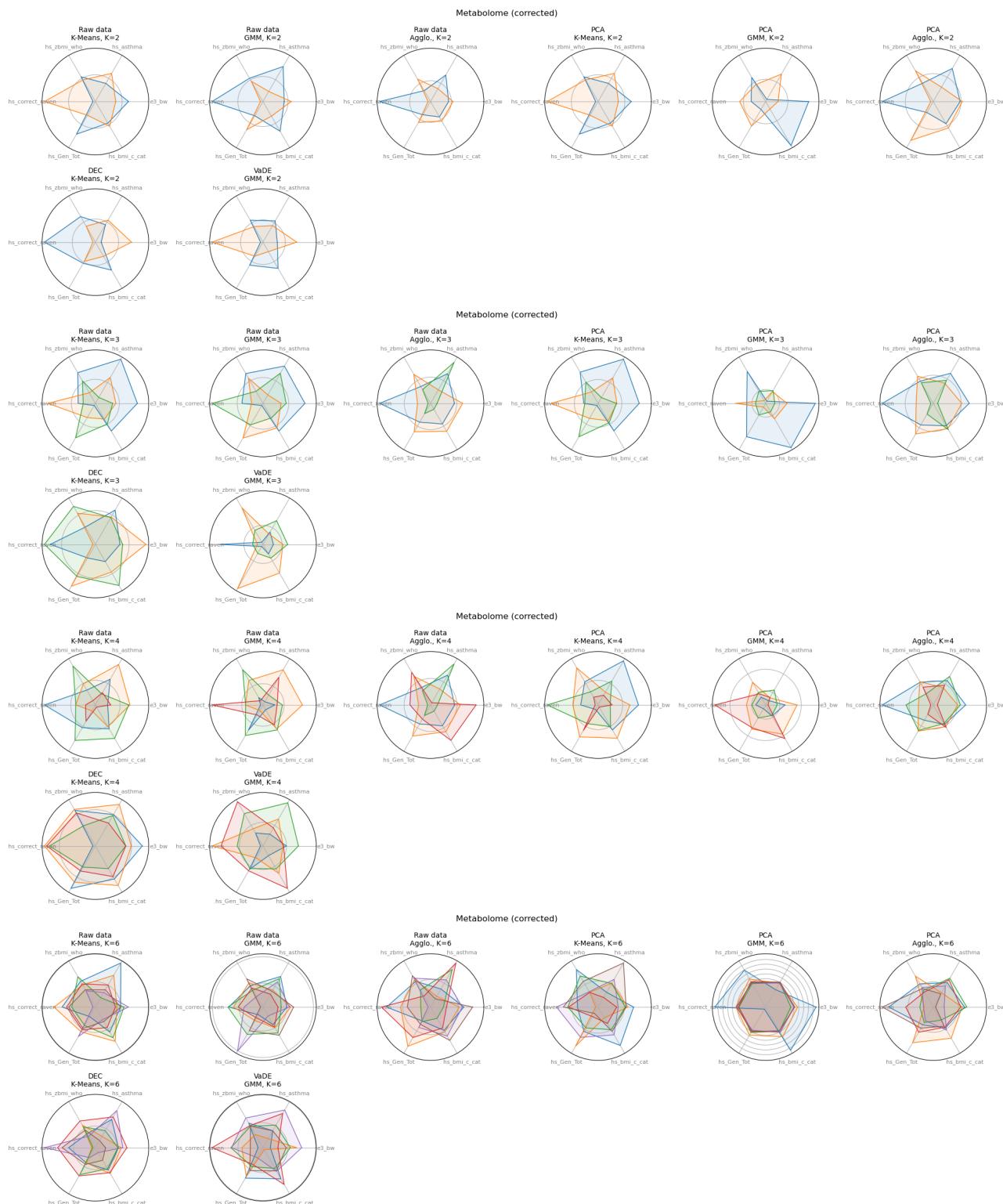


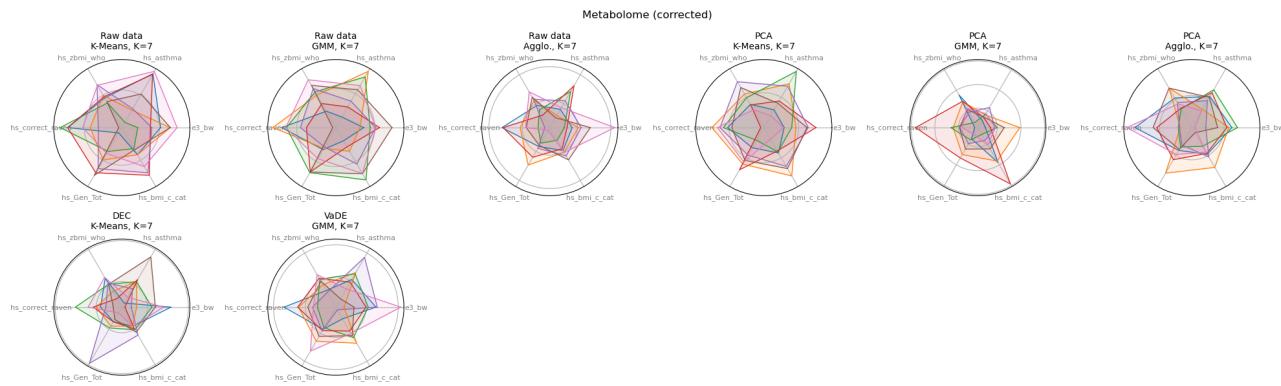




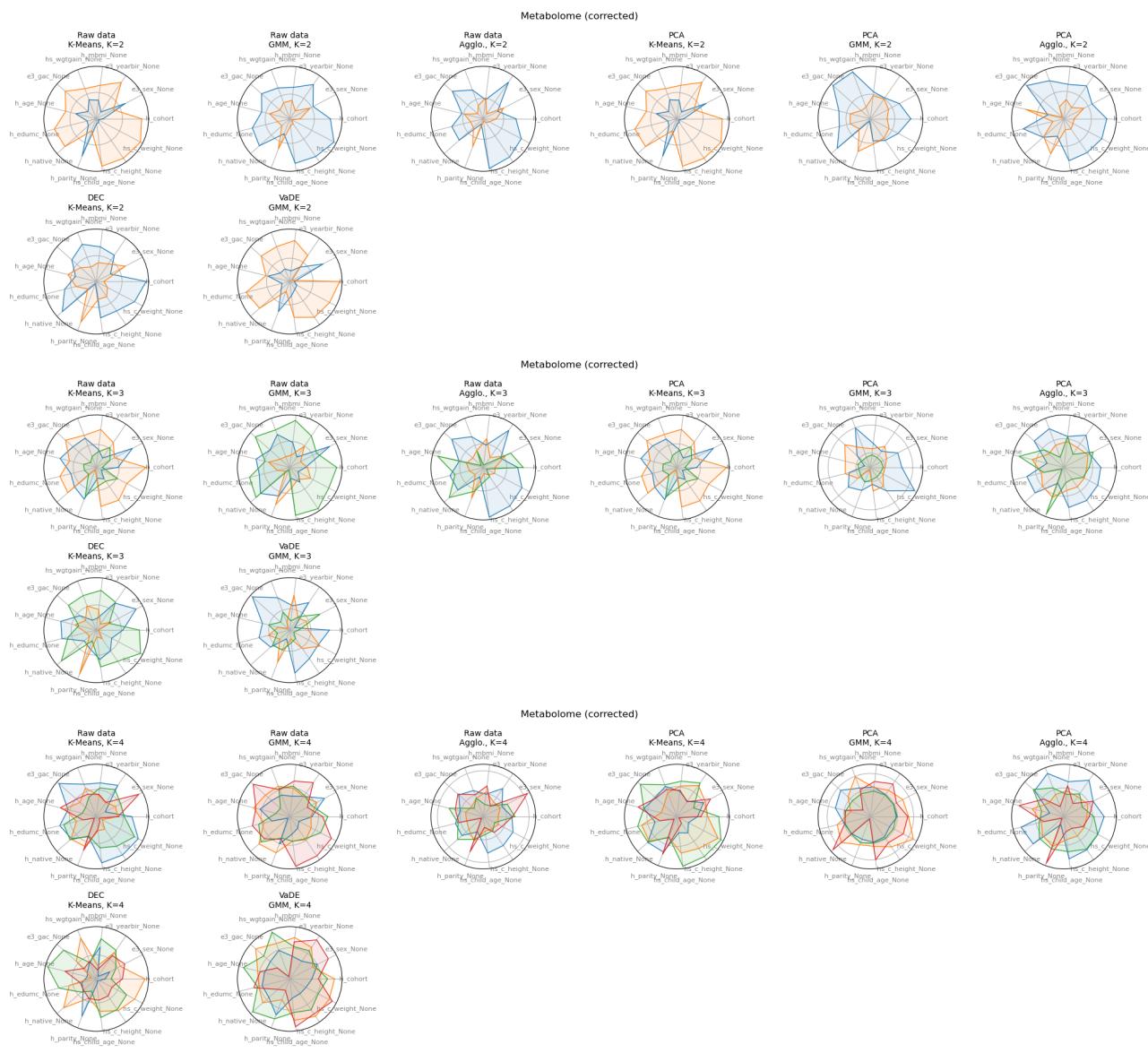
A.3.2 Conjunt de dades: metaboloma (corregit per l'efecte de log)

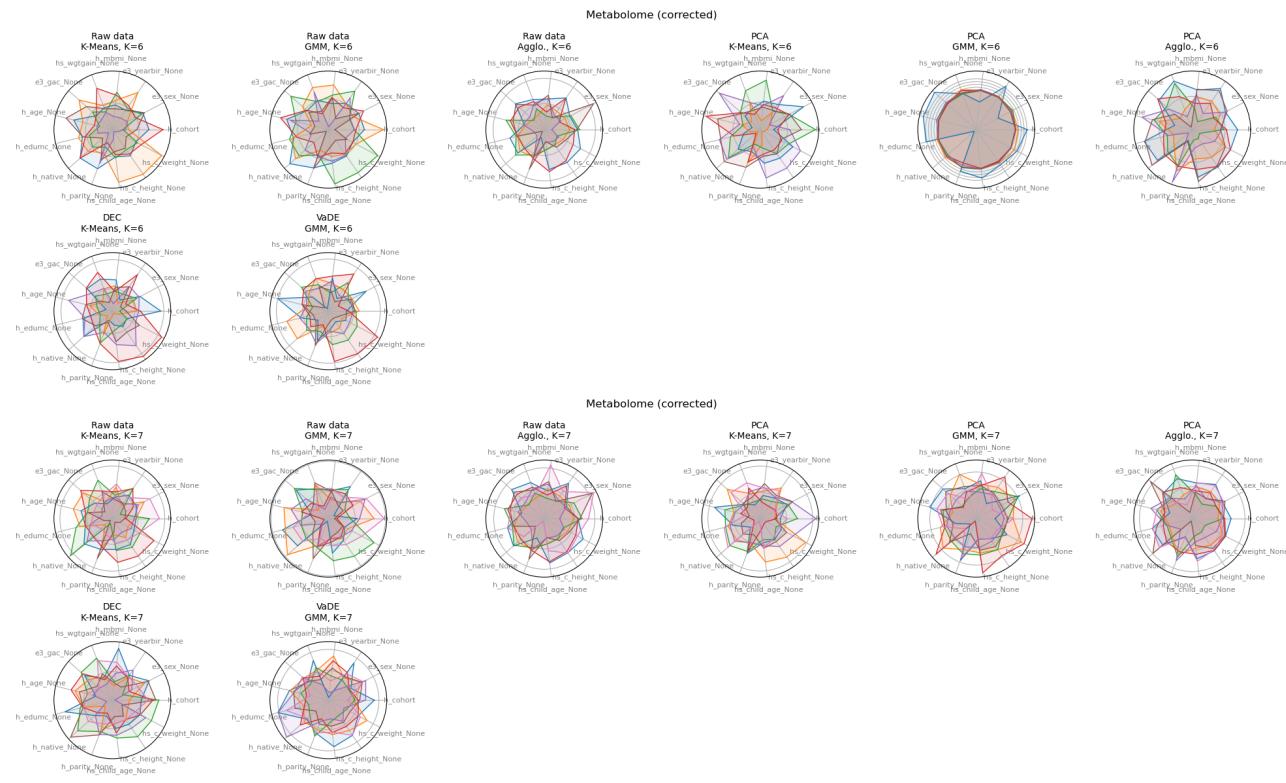
Dades fenotip: distribució multivariant en funció dels clústers.



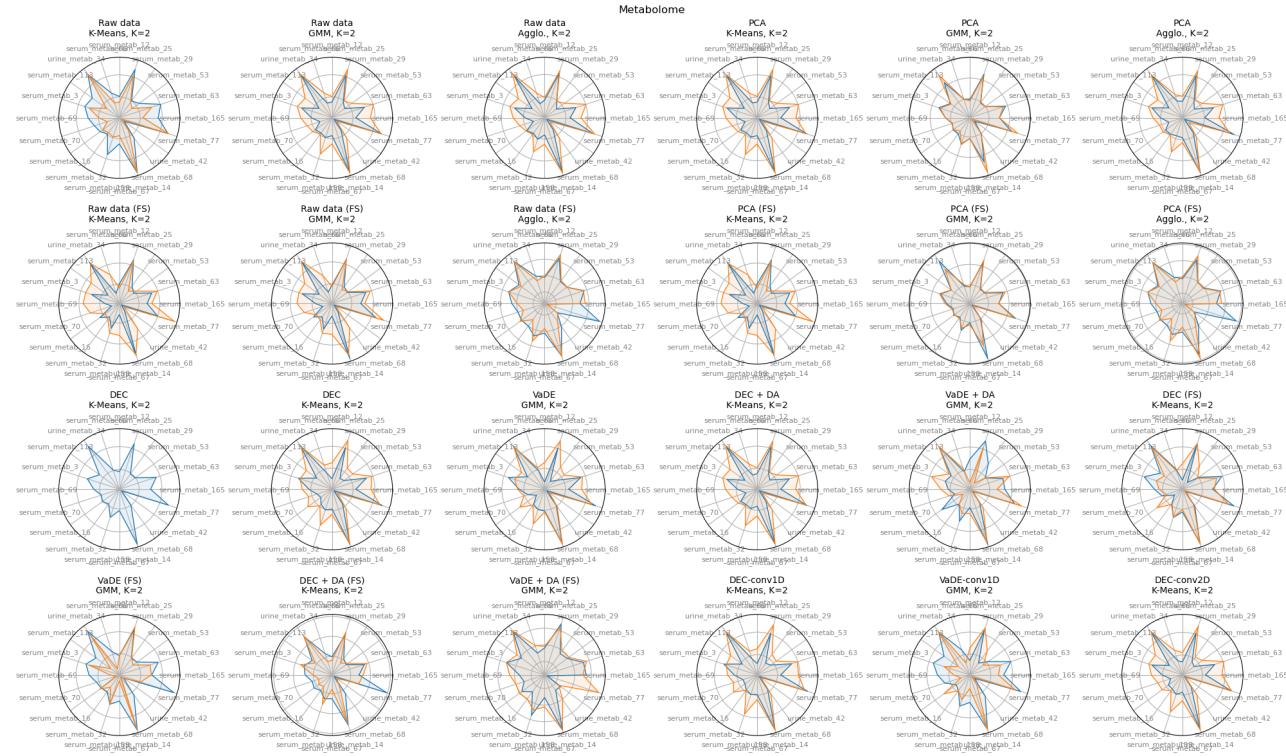


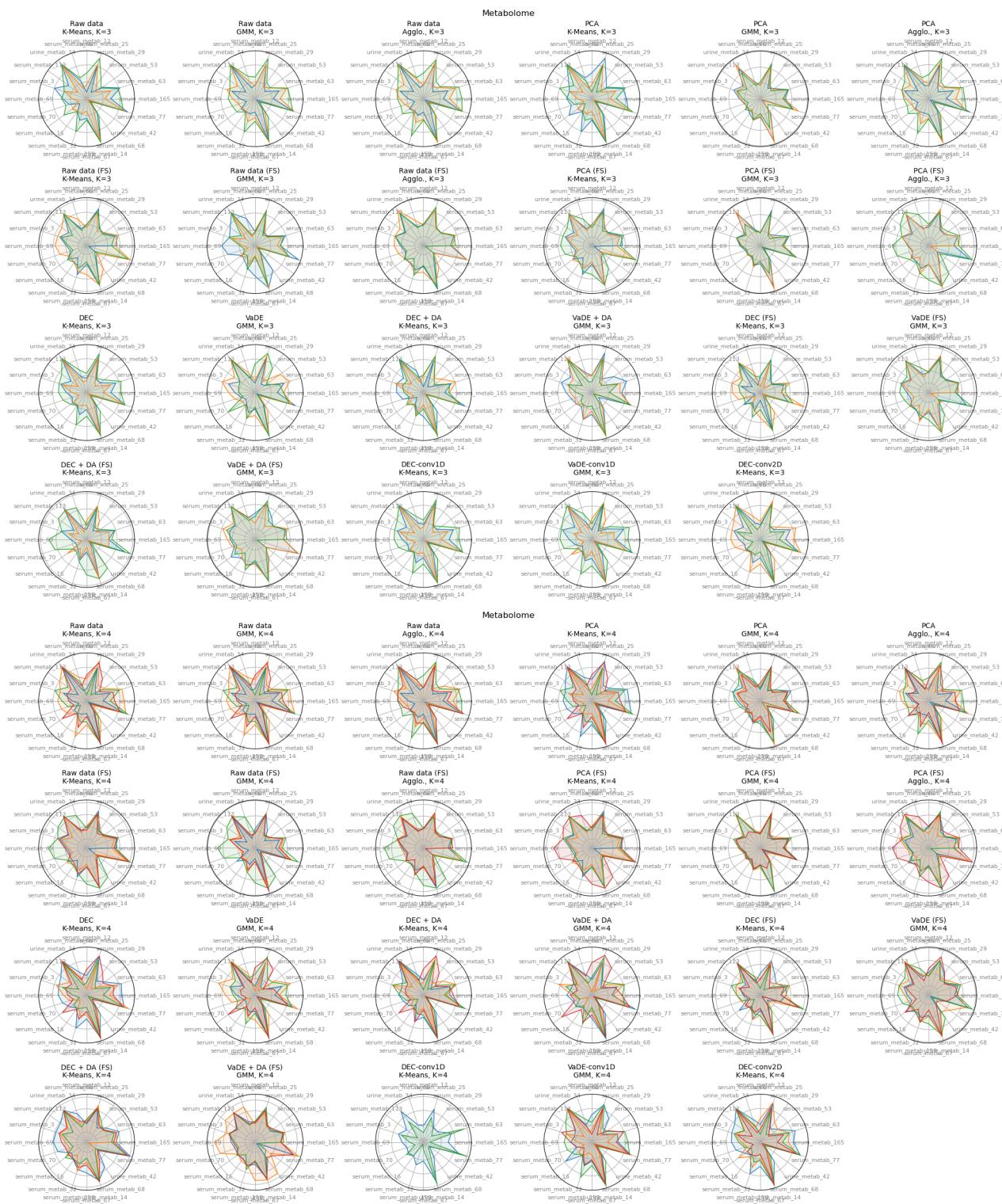
Dades covariables: distribució multivariant en funció dels clústers.





Dades metaboloma: distribució multivariant en funció dels clústers (20 variables amb major variabilitat).

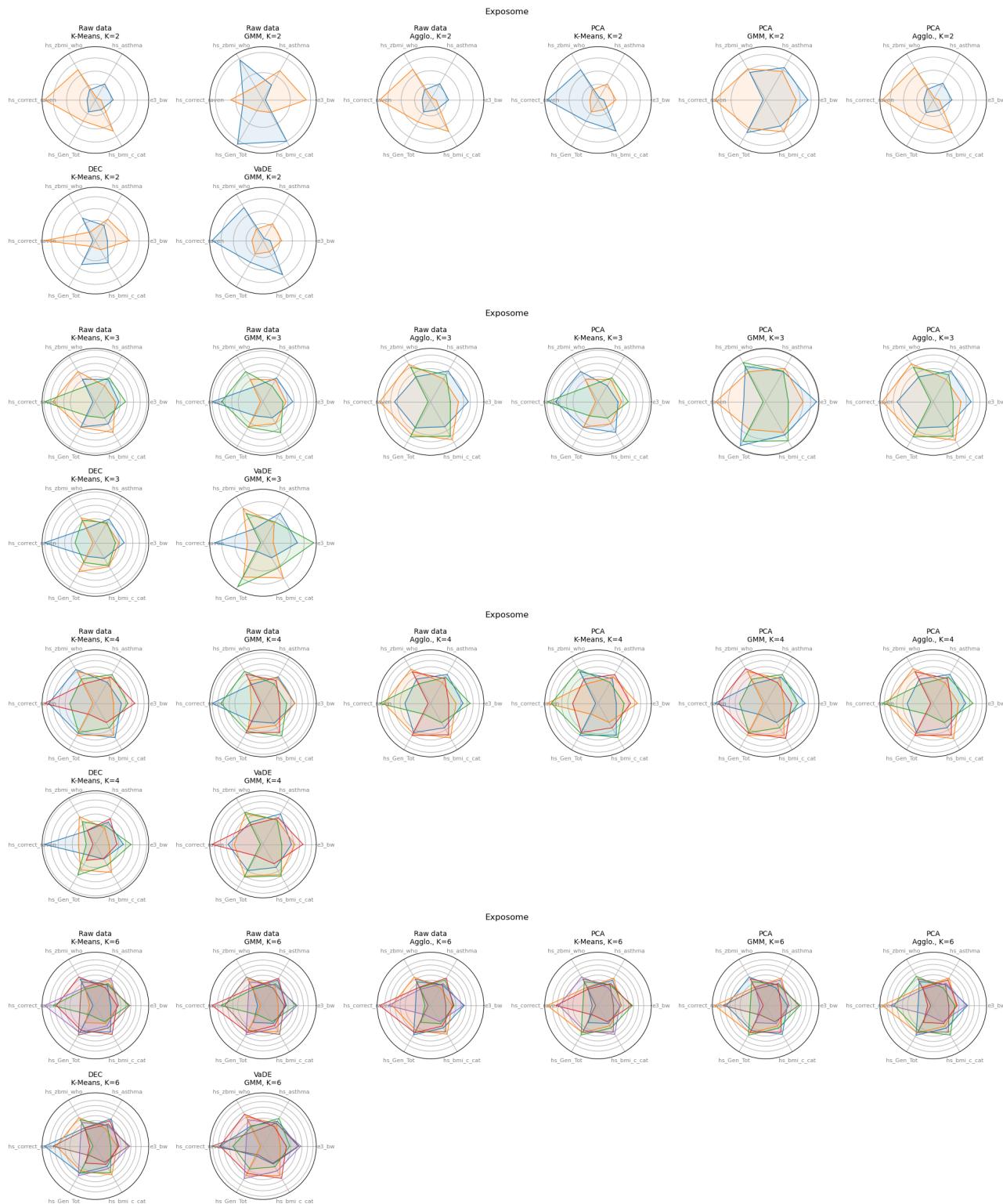


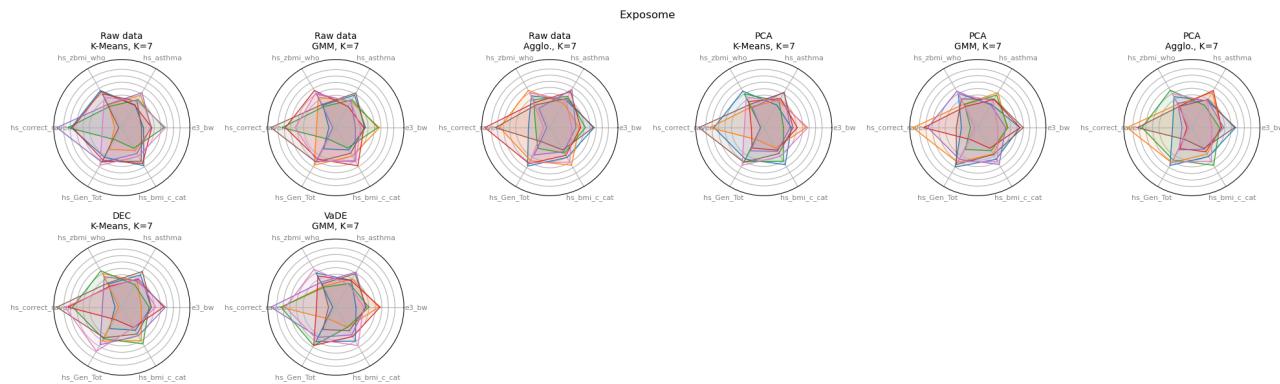




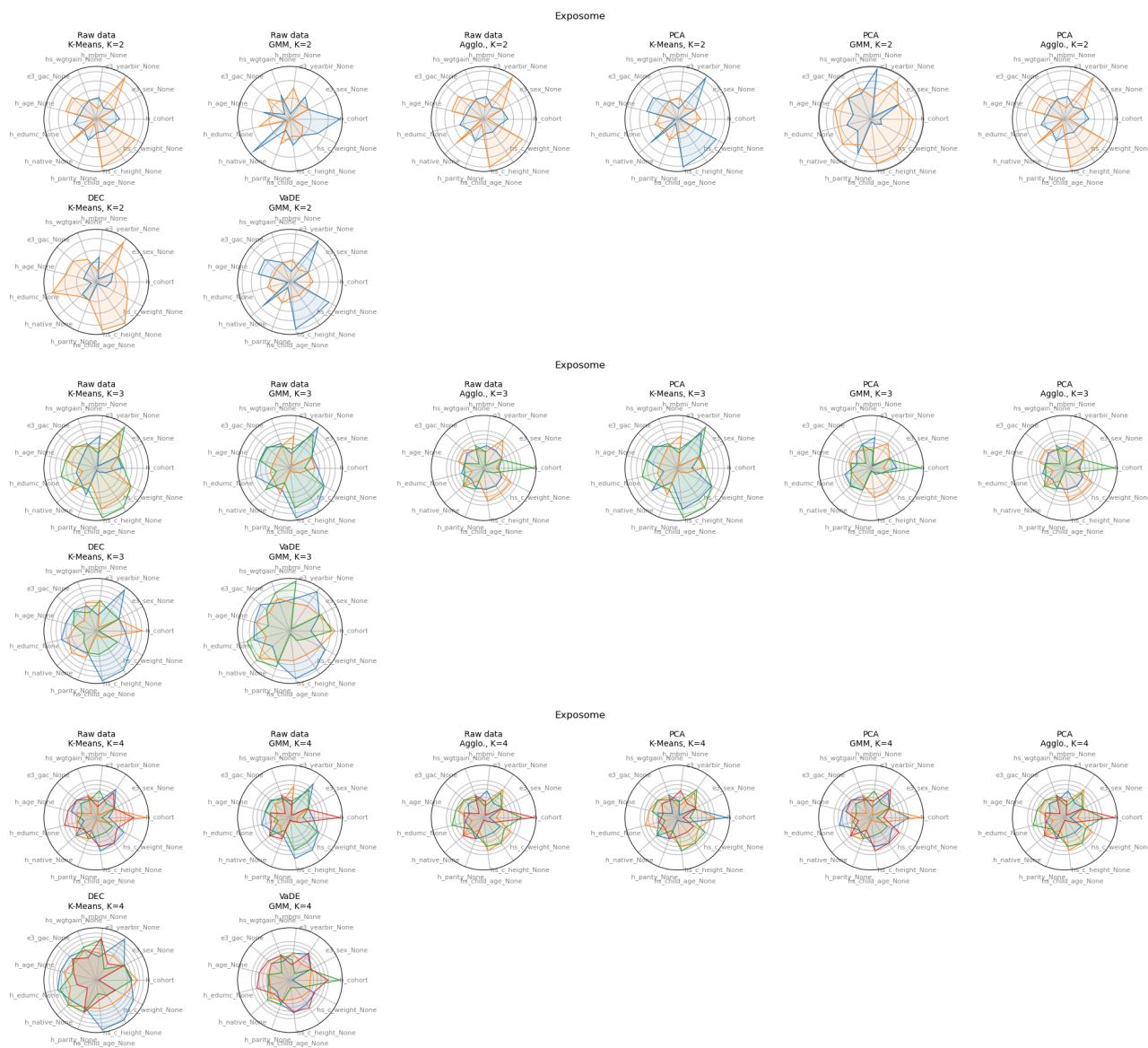
A.3.3 Conjunt de dades: exposoma

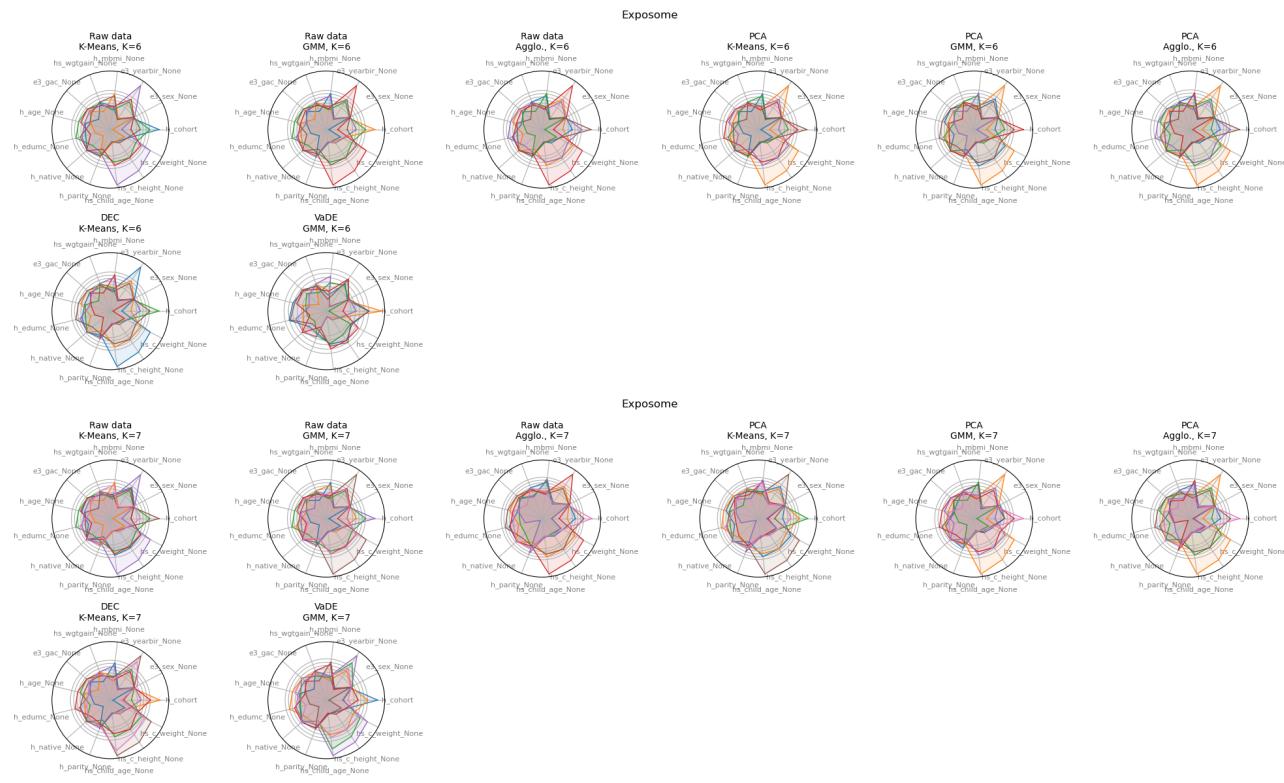
Dades fenotip: distribució multivariant en funció dels clústers.



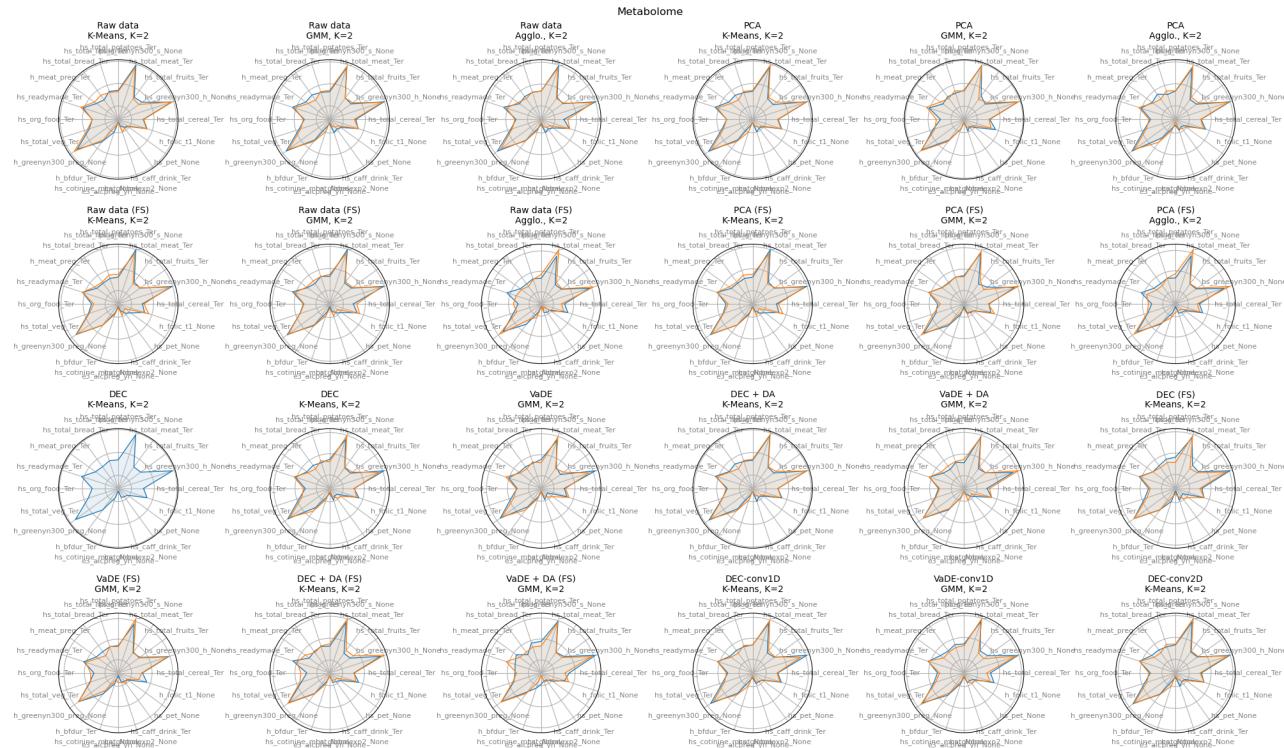


Dades covariables: distribució multivariant en funció dels clústers.

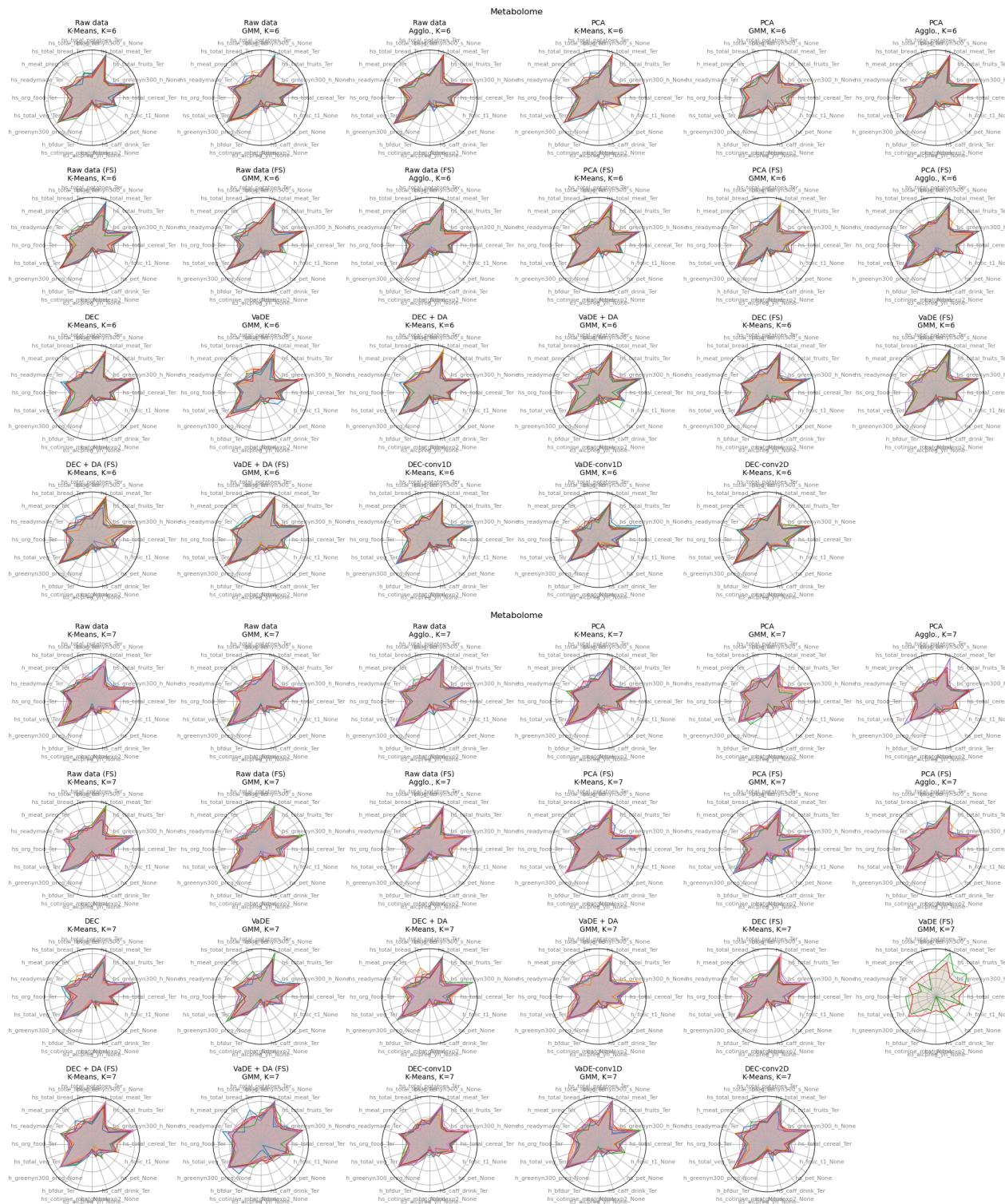




Dades metaboloma: distribució multivariant en funció dels clústers (20 variables amb major variabilitat).

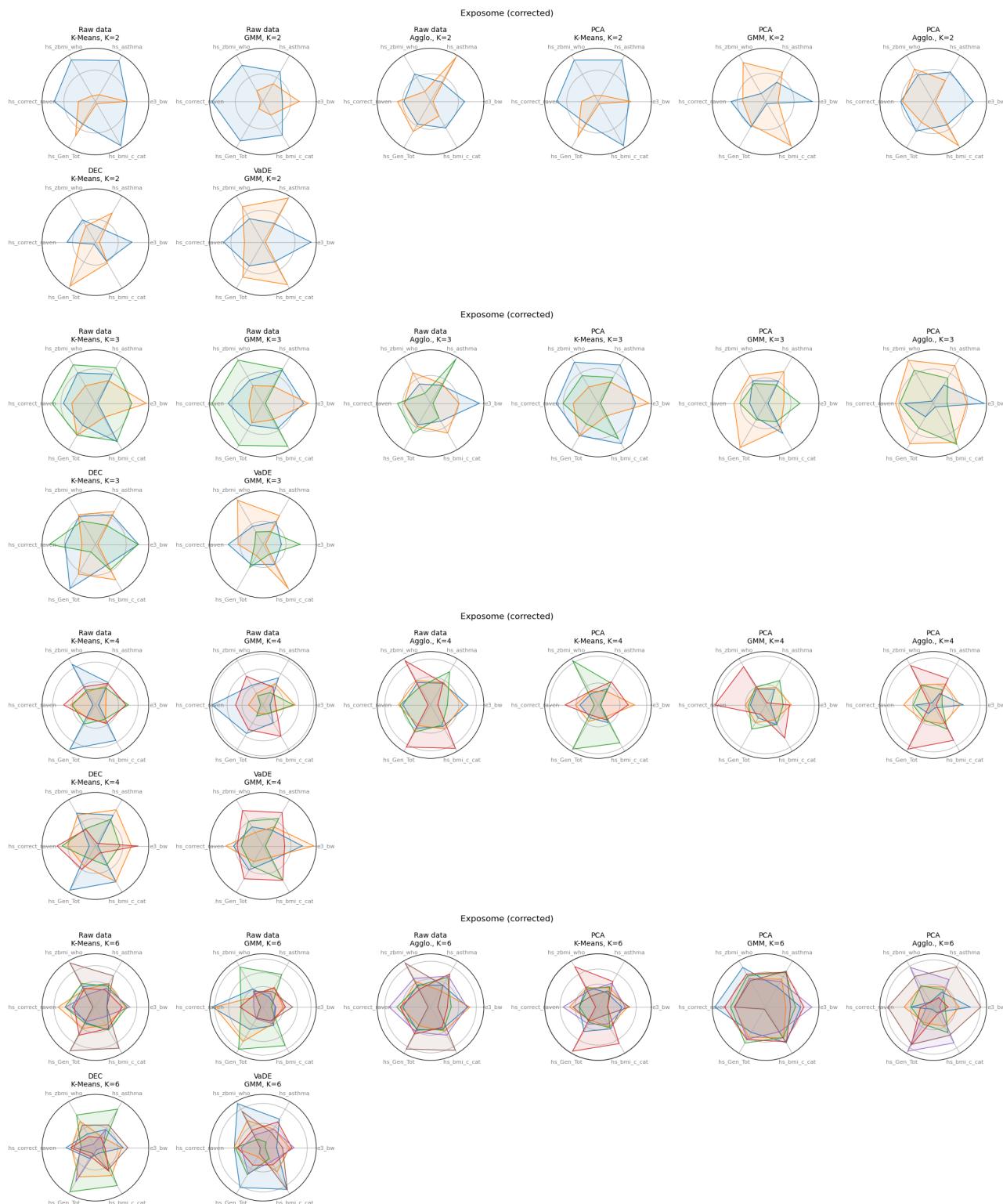


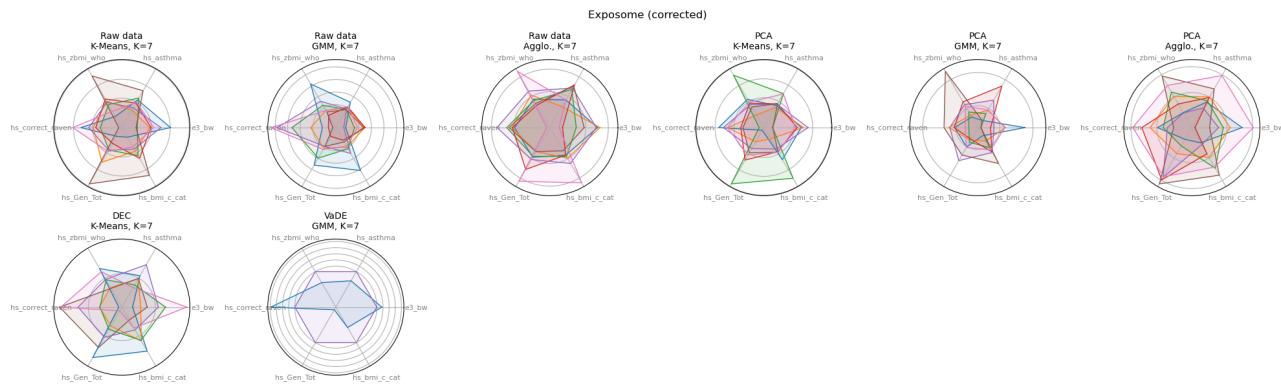




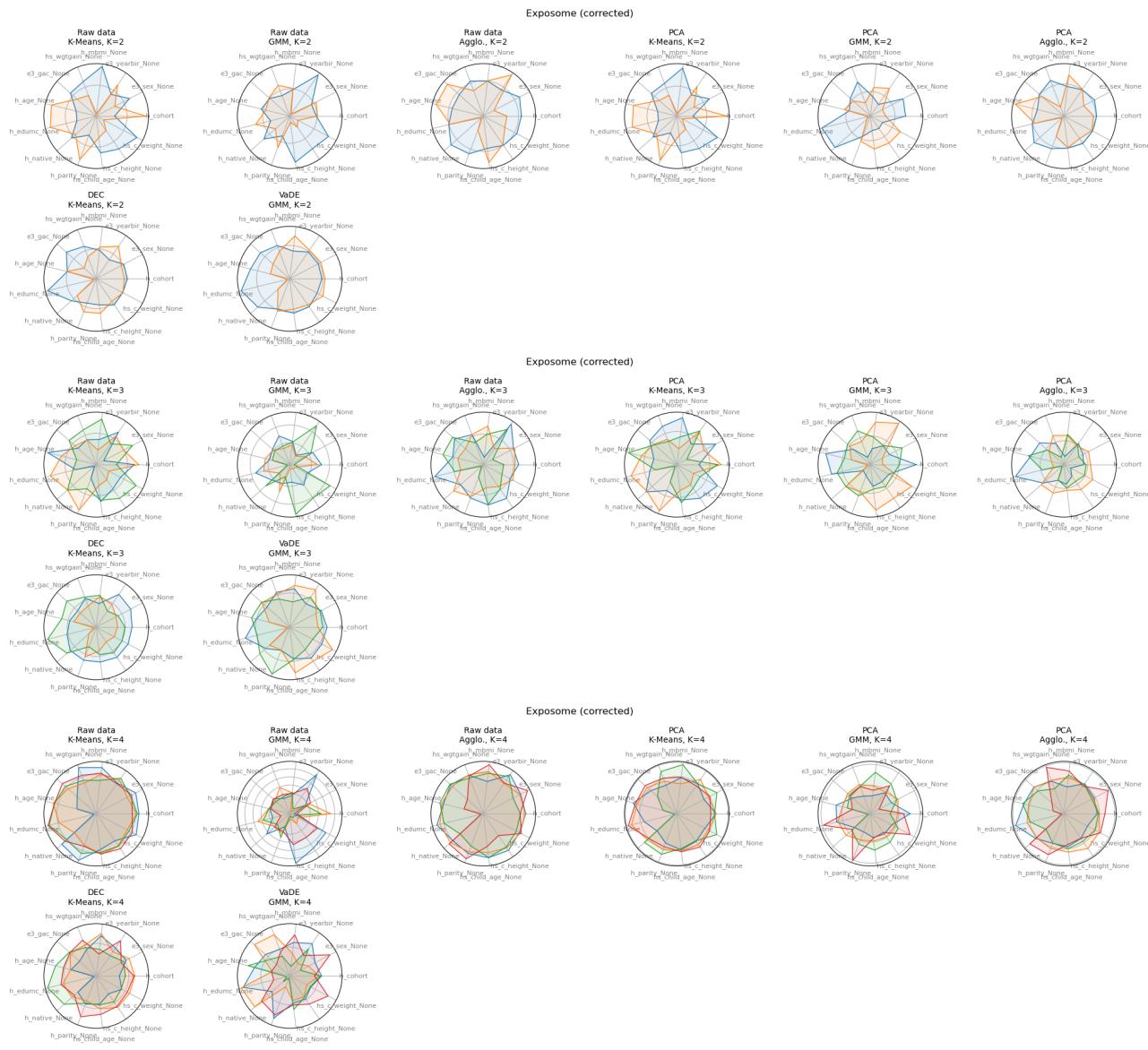
A.3.4 Conjunt de dades: exposoma (corregit per l'efecte de log)

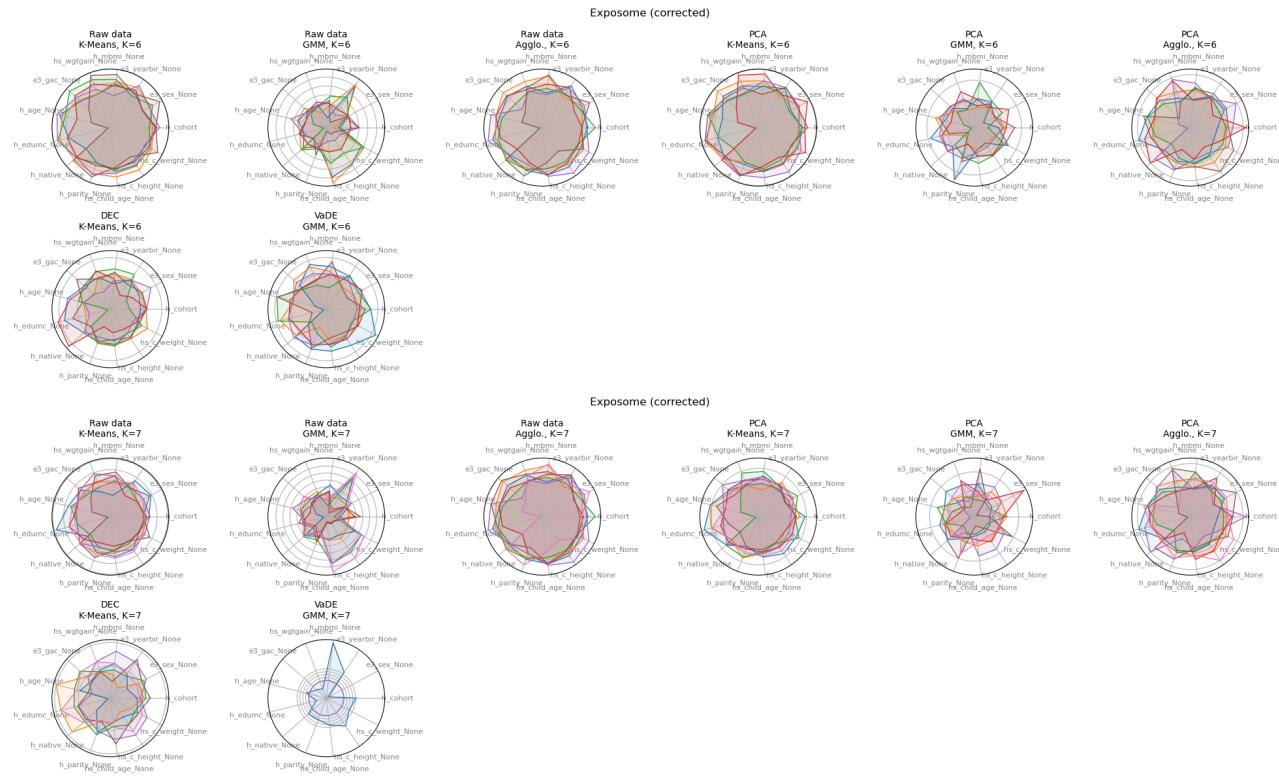
Dades fenotip: distribució multivariant en funció dels clústers.





Dades covariables: distribució multivariant en funció dels clústers.





Dades metaboloma: distribució multivariant en funció dels clústers (20 variables amb major variabilitat).



