Stability and selection of the number of groups in unsupervised clustering : application to the classification of triple negative breast cancers

Martina Sundqvist

université
PARIS-SACLAY

Stability and selection of the number of groups in unsupervised clustering: application to the classification of triple negative breast cancers

Thèse de doctorat de l'Université Paris-Saclay

Ecole Doctorale de Mathématique Hadamard (EDMH) n° 574
Spécialité de doctorat: Mathématiques appliquées
Unité de recherche: UMR 518 AgroParisTech,
INRAE, Université Paris-Saclay
Référent: Faculté des sciences d'Orsay

Thèse présentée et soutenue en visioconférence totale, le 18 décembre 2020, par

Martina Sundqvist

Composition du jury:

Christophe Ambroise Professeur Université d'Évry Val d'Essonne, Université Paris-Saclay	Président
Chloé-Agathe Azencott Maîtresse-Assistante, MINES ParisTech	Examinatrice
Avner Bar-Hen, Professeur, Conservatoire national des arts et métiers	Rapporteur & Examinateur
Anne-Laure Boulesteix Professeur, Ludwig Maximilian University of Munich	Rapporteure & Examinatrice
Max Chaffanet Responsable d'équipe en Oncogénomique Moléculaire, Institut Paoli calmettes	Examinateur
Julien Chiquet Directeur de recherche, INRAE, Université Paris Saclay	Directeur
Thierry Dubois Responsable du groupe « Biologie du Cancer du Sein », Institut Curie, Centre de Recherche, Paris, France	Codirecteur invité
Guillem Rigaill Charge de recherche, INRAE, Université Paris Saclay	Codirecteur invité

Thèse de doctorat

NNT: 2020UPASM026


institutCurie


INRAE


îledeFrance

À ma grand-mère Kerstin et à toutes les filles
qui, comme elle, n'ont pas pu faire les études qu'elles voulaient
pour la simple raison qu'elles étaient filles.

Un grand merci

D'abord, je tiens à remercier le jury d'avoir accepté d'examiner ma thèse. Merci pour le temps que vous y avez investi, vos suggestions, remarques et vos questions pertinentes. Merci à Anne-Laure Boulesteix et à Avner Bar-Hen d'avoir lu ma thèse en détail. Merci de vos remarques pertinentes et de vos commentaires gentils.

Puis je tiens évidemment à remercier mes trois encadrants, Julien Chiquet, Guillem Rigaill et Thierry Dubois, sans lesquels cette thèse n'aurait pas pu avoir lieu. Merci d'abord de m'avoir pris en stage puis de m'avoir poussé à continuer en thèse. Merci pour ces quasi quartes années que nous avons partagées ensemble. Cela fut vraiment un plaisir à travailler avec chacun de vous et j'ai appris tellement de choses. Merci d'avoir partagé avec moi l'expertise de vos respectifs domaines scientifiques devant lesquels je reste toujours aussi impressionnée. Aussi, merci de votre bienveillance, de votre patience, de votre écoute et de toujours avoir "été là" pour moi. Surtout, merci, d'avoir cru en moi avant que je le faisais moi-même et de m'avoir appris à me surpasser et à toujours aller plus loin.
Thierry, merci de m'avoir accueilli dans ton laboratoire comme si j'étais une parmi vous. Travailler côte à côte avec des biologists a été très formateur. Merci de ta patience, ton ouverture envers les maths et ta rigoureusité scientifique qui m'a beaucoup marqué. Guillem merci de m'avoir introduit à la statistique computationnelle. Merci pour toutes ces heures de calcules (et de vérification de calcules) combinatoires passées ensemble. J'ai beaucoup aimé construire mon sujet de thèse autour du cluster stabilité et le Rand Index donc merci, de m'avoir introduit à ces sujets. Surtout, merci de toujours m'avoir écouté et m'avoir étayé. Julien, merci pour toutes ces heures passées dans ton bureau en mode de travail, avec des éclats de rire jamais très loin. Merci d'avoir partagé ta passion pour des codes "bien propres et efficaces". Cela m'a beaucoup fait progresser. Merci, d'avoir été mon directeur de thèse et de toute aide administrative cela a impliqué. Surtout, merci pour ton amitié.

Je veux également remercier mes deux laboratoires d'accueil, le laboratoire de biologie du cancer du sein (BCBG) à l'Institut Curie et MIA-Paris situé

à AgroParisTech. D'abord merci à tous les membres de BCBG de m'avoir accueilli (en tant que statisticien-magicienne) parmi vous. Merci de tout ces gâteaux, escape games, wednesday pubs ainsi que toute votre aide pour que je comprenne l'enjeu biologique de ma thèse. Merci notamment, Sam(ycute), Ramon, Virginie, Olivier, Rania, Mathilde, Clarisse et Amélie.

Merci aux membres de MIA-Paris de m'avoir accueilli parmi vous. C'est vraie de ce que l'on dit, Agro, c'est comme une famille et je vous remercie tous de m'avoir permi d'en faire partie. Merci pour tous ces moments de joix partagées ensemble et de tout ce que vous m'avez appris sur la culture français. Surtout, merci à mes sœurs de thèse : Annarosa, Rana, Raph et Marie, sans vous, ma thèse n'aurait pas été la même. Merci pour tous nos rires, nos pleurs et nos aventures.
Merci aux doctorants+: Félix, Saint-Clair, Joe, Thimothée, Sema, PAM, Yann, Béwéntaoré, Mattieu, Paul, Mounia, Claire, Jade, Audrey, Gabriel, Pierre (B et G), Sarah, Laure, Tristan. Entre les week-ends à la campagne, les soirées chez quelqu'un, les verres à la belle mine, les pauses de thèse interminables (Oui bizarrement, elles étaient toujours plus longues quand tu étais là Yann ;)), les karaokés, les confs, je n'ai pas vu le temps passer. Cela a été desannées magnifiques.

Puis, je tiens à remercier ma famille et mes amis. Tack mamma och pappa för att ni alltid finns där för mig. Tack Paui för all kärlek och goda råd, tack Joel för alla galna och härliga upplevelser vi alltid har tillsammans. Tack Sigrid och Ester för att ni är dom mest underbara männsikorna som någonsin funnits på denna jord och för all kärlek och lycka som ni sprider runt omkring er. Puis, merci à ma tante chérie sœur Maria.

Tack Maria bästis för ditt stöd de senaste åren, det har verkligen varit så underbart att växa tillsammans. Tack min bästaste Mosi för allt och alla härliga minnen vi delar. Tack till alla Solna brudar (ni vet vilka ni är ;)) för att ni aldrig har slutat att finnas där. Tack till "kören tjejerna", Anna-Karin, Gabriella, Linda och Ivana, mer (kör)häng snart! Merci aux filles du master en biostaistiques, Oriane, Alex et Aurélie. Surtout merci Oriane pour ton aide avec la relecture de la thèse. Merci à Katia, Milica et Luis pour votre amitié. Merci la communauté dansant de Paris, surtout merci à mes danseuses préférées Estelle <3, Sousou, Pamela et Cyrielle.

Merci l'équipe de SBE de l'IGR de m'avoir accueilli parmi vous. J'ai hâte d'entamer cette nouvelle aventure avec vous.

Bref, merci à toute personne qui m'a accompagné pendant ma thèse et je m'excuse d'avance pour ceux que j'ai aurais oublié de mentionner. Sachez que, si m'a tête a pu oublier, cela n'est pas le cas pour mon cœur.

Je tiens également à remercier également à la région d'Île de France d'avoir financé ma thèse thèse. Également, merci à l'EDMH et ses deux directeurs Frédéric Paulin et Stéphane Nonemacher de m'avoir permis a effectuer cette thèse et pour tout soutien durant ces années.

Finalement, je tiens à remercier l'enseignement supérieur en France. En effet, j'obtiens ma thèse 210 ans après Sophie Germain a dû se faire passer pour un homme afin de pouvoir défendre ses travaux en mathématiques. Puis seulement 146 ans après que la première femme dans le monde, Sofia Kovalevskaya, a obtenu une thèse en mathématiques et 132 ans après que la première femme, Louise-Amélie Leblois, a obtenu une thèse en science en France. Aussi, aujourd'hui, 14% des enfants (filles comme garçons) entre 7 et 17 ne sont pas toujours pas scolarisés (chiffre selon UNESCO). Alors, j'ai donc été née au bon endroit, et au bon moment afin d'avoir eu cette chance de réaliser un doctorat. Je tiens donc à remercier, avec tout mon cœur, tout mouvement et toute personne qui s'est battue pour la démocratisation de l'enseignement supérieur. J'espère dans ma carrière pouvoir rendre ce qui m'a été donné en continuant cette bataille (qui ne devrait jamais cesser) afin d'un jour rendre l'éducation un droit pour tous et ne plus un privilège.

Preface

You cannot relay on the brakes to climb a hill
- Swedish saying (from Sally), freely translated by me

Je souhaite commencer cette thèse en expliquant le contexte dans lequel elle s'est déroulée. Cette partie est la seule partie (remerciements exclue) qui sera entièrement rédigée en français.

Tout d'abord, cette thèse a été une très belle aventure scientifique, elle m'a permis de m'insérer dans le domaine des biostatistiques, mais aussi de m'ancrer un peu plus en France, mon pays d'adoption depuis maintenant plus de 10 ans. Cette thèse s'inscrit dans un parcours universitaire diversifié qui a commencé par une licence en psychologie, puis un master en neurosciences cognitives et finalement un master en Recherche en Santé Publique option biostatistiques.

C'est au cours du stage de ce dernier master que j'ai rencontré mes encadrants, Julien Chiquet et Guillem Rigaill (deux statisticiens gentils et très forts) ainsi que de Thierry Dubois (biologiste, aussi gentil et aussi fort). Ils m'ont alors permis de réaliser un premier travail sur la classification des cancers du sein triple négatifs (Triple Negative Breast Cancer - TNBC). Ce stage a été le point de départ de ma thèse, pour laquelle j'ai obtenu une allocation doctorale ARDoC, de la région d'Ile de France, priorité santé en cancérologie, ce qui m'a permis de m'inscrire à l'EDMH pour une thèse en mathématiques appliquées. Pendant le stage, comme pendant ma thèse, j'ai partagé mon temps entre l'UMR MIA, à AgroParisTech (un laboratoire de mathématiques appliquées), et le laboratoire de Thierry Dubois (Breast Cancer Biology Lab), du département de recherche translationnelle à l'Institut Curie (un laboratoire de biologie cellulaire). Cette thèse est donc (comme moi), le fruit d'une rencontre entre deux cultures (scientifiques) qui se comprennent souvent... mais pas toujours. Je tiens ainsi, à appuyer l'importance que ce cadre interdisciplinaire a pu avoir sur ma thèse, à la fois sur l'élaboration du sujet que sur les contributions scientifiques produites. Avec ce manuscrit, j'espère pouvoir montrer la richesse qu'une telle collaboration, mathématiques-biologie, peut apporter à un sujet précis.

Mon sujet de thèse portait à la base sur l'intégration de protéines dans la classification des TNBC (souvent basée sur des données génomiques), dans le but de trouver des nouvelles pistes thérapeutiques. L'idée était d'utiliser, entre autre, des modèles graphiques gaussiens multi-attributs pouvant lier des réseaux de gènes et des réseaux de proteines associés aux différents groupes TNBC. Cependant, suite à la difficulté de répliquer des résultats basés sur l'analyse des protéines, le sujet de ma thèse a dévié. En effet, dans le Chapitre 2, j'expose un travail où j'ai classifié des tumeurs de TNBC en utilisant des méthodes habituellement employées dans la littérature pour cette tâche. J'ai ainsi pu trouver des résultats convaincants, mais que je n'ai pas réussi à valider par la suite sur un autre jeu de données. À ceci, s'ajoute le fait qu'il n'existe pas une classification des TNBC, mais plusieurs qui diffèrent toutes entre elles. Il y a beaucoup de raisons, tant biologiques que méthodologiques, qui peuvent expliquer pourquoi ces résultats n'ont pas pu être validés, ou pourquoi ces classifications diffèrent. Mais une des raisons qui m'a semblé particulièrement préoccupante est l'utilisation parfois « à la légère », de certains outils statistiques.

En conséquence, plutôt que proposer un nouveau modèle (encore plus complexe) afin de mieux comprendre la classification des TNBC, j'ai préféré me focaliser sur la compréhension des méthodes ayant été utilisées pour proposer ces classifications. C'est ainsi que je me suis intéressée à la stabilité des clusters pour la sélection du nombre de classes en clustering. Cette méthode est une méthode pratique et facilement utilisable, mais pour laquelle il reste encore des zones d'ombre.

L'idée est que plus un clustering (ou classification) est stable, plus il est probable qu'il représente la vraie partition du jeu de données. Le nombre de groupes d'un clustering peut ainsi être choisi comme celui donnant lieu au clustering le plus stable. Pour estimer la stabilité, les trois étapes suivantes sont employées :

1. Générer des jeu de données perturbés à partir du jeu de donnée initial en faisant par exemple du sous-échantillonnage.

2. Classifier les jeux de données perturbés en utilisant un algorithme de clustering.

3. Comparer ces clusterings en les comparant avec un score de similarité ou de distance.

Cette méthode fait entrer en jeu un grand nombre de paramètres dont on ne connaît pas encore l'impact. Aussi, plusieurs versions ont été implémentées avec une promesse de nouveauté, mais qui en réalité, revient souvent à un paramétrage spécifique de ces trois étapes. Pour cette raison, j'ai implémenté un package R, `clustRstab`, qui permet de facilement changer et tester différents paramétrages afin de voir leur impact sur un jeu de données. Grâce à ce package, j'ai pu réaliser une étude de simulation et une étude d'application testant comment et quand cette méthode fonctionne. Ces études ainsi que le package sont exposés dans le Chapitre 4.

Puis, en cherchant à mieux comprendre cette méthode, je me suis orientée vers les scores de comparaison de clusterings (étape 3) et notamment vers le Rand Index (RI). Ce score, et sa version ajustée, sont sans doute les scores les plus populaires de comparaison de clustering. Cependant, ils ont tous les deux été dérivés de manière ad-hoc. La version ajustée (ARI) a été déduite d'une hypothèse de distribution hypergéométrique, qui, d'un point de vue de modélisation statistique, est insuffisante, car elle ne permet pas de prendre en compte le cas de dépendance entre clusterings ni de faire d'inférence statistique. J'ai ainsi dédié une grande partie de ma thèse à la correction de ce score. Ce que j'ai fait en (1) redéfinissant le RI pour être plus interprétable et (2) le posant dans un cadre statistiques bien défini et le basant sur une hypothèse de distribution multinomial. Grâce à ce travail, j'ai pu proposer une nouvelle version de l'ARI, le « Modified Adjusted Rand Index », $MARI$, que j'ai par la suite implémenté dans le package R `aricode`. Ceci est la contribution méthodologique la plus importante de ma thèse et elle est exposée dans le Chapitre 3.

Finalement, ces deux méthodes sont appliquées à une grande cohorte de patientes atteintes de TNBC. Les résultats de ces analyses sont comparés aux résultats obtenus par la méthode de classification TNBC proposée par Lehmann et al. (2011) et sont présentés dans le Chapitre 5.

Étant donné la diversité des sujets traités dans cette thèse, mon premier chapitre fera l'objet à la fois d'une introduction et d'un état de l'art. J'espère ainsi pouvoir donner les bases nécessaires à la compréhension de la suite de ma thèse.

Contents

Chapter 1

State of the art

1.1. Introduction

In this thesis, I treat the topic of classifying Triple Negative Breast Cancer Tumors (TNBC) from a statistical point of view. To do so, I mainly focus on clustering and its validations techniques. More precisely, I focus on the use of cluster stability for selecting the number of groups in unsupervised clustering. Indeed, this is the method generally employed when classifying TNBC. This method aims to propose a stable clustering and to do so, clusterings obtained upon the same, but perturbed, dataset are compared. However, despite the popularity of this method, little is still known about how or under which conditions it works. In order to improve the interpretability of this method, I studied its behavior in different settings. To do so, I implemented an R package `clustRstab` that easily computes the stability in different parameter settings. Since cluster comparison is crucial for estimating the stability of a clustering, I also focused on the Rand Index Rand (1971), one of the most popular scores for cluster comparison. These methods are then illustrated on three large TNBC datasets, whereof one is presented in this first chapter (state of the art). The aim of this thesis is therefore two fold, with (1) getting a better understanding of the TNBC classifications and (2) doing so by getting a better understanding of the use of cluster stability as a criterion for selecting the numbers of groups in unsupervised clustering. Since I treat both statistical methods and their application to biological data, inevitably, a substantial part of this thesis is dedicated to data processing and data normalisation.

In order for the reader to get the biological and methodological bases to understand these topics, this first chapter is both an introduction to the topic of my thesis as well as a state of the art. The chapter is divided in three parts. The first part treats the biological fundamentals necessary to understand the biological stakes and challenges for classifying TNBC, covering omic data and its measurement, an overview of earlier TNBC classifications as well as a first

reflexion of different methodological issues linked to classifying TNBC. The second part presents the RATHER consortium that is used as an illustration set. The third part introduces the reader to unsupervised clustering methods and clustering validation criteria. A special attention is given to cluster stability and cluster comparison scores.

1.1.1. Introduction (français)

Dans cette thèse, je traite le sujet de la classification des tumeurs cancéreuses du sein triple négatif (TNBC) d'un point de vue statistique. Pour ce faire, je me concentre principalement sur le clustering et ses techniques de validation. Plus précisément, je me concentre sur l'utilisation de la stabilité des clusters pour sélectionner le nombre de groupes dans le clustering non supervisées. En effet, c'est la méthode généralement utilisée lors de la classification TNBC. Cette méthode vise à proposer un clustering stable et pour ce faire, les clusterings obtenues sur le même jeu de données, mais perturbés, sont comparées. Cependant, malgré la popularité de cette méthode, on sait encore peu de choses sur la façon dont elle fonctionne. Afin d'améliorer l'interprétabilité de cette méthode, j'ai étudié son comportement dans différents contextes. Pour ce faire, j'ai implémenté un package R qui calcule facilement la stabilité dans différents paramètres. Comme la comparaison de clusterings est cruciale pour estimer la stabilité d'un clustering, je me suis également concentré sur l'indice de Rand Rand (1971), l'un des scores les plus populaires pour la comparaison de clusterings. Dans cette thèse, je propose une une version modifiée de l'indice de Rand. Ces méthodes sont ensuite illustrées sur trois grands ensembles de données TNBC, dont un est présenté dans ce premier chapitre (état de l'art). L'objectif de cette thèse est donc double, avec (1) une meilleure compréhension des classifications TNBC et (2) une meilleure compréhension de l'utilisation de la stabilité des clusters comme critère de sélection du nombre de groupes dans le clustering. Comme je traite à la fois des méthodes statistiques et de leur application aux données biologiques, inévitablement, une partie importante de cette thèse est consacrée au traitement et à la normalisation des données.

Afin que le lecteur puisse disposer des bases biologiques et méthodologiques pour comprendre ces sujets, ce premier chapitre est à la fois une introduction au sujet de ma thèse et un état de l'art. Le chapitre est divisé en trois parties. La première partie traite des bases biologiques nécessaires à

la compréhension des enjeux et défis biologiques de la classification du TNBC, couvrant les données omiques et leur mesure, un aperçu des classifications antérieures du TNBC ainsi qu'une première réflexion sur les différentes questions méthodologiques liées à la classification du TNBC. La deuxième partie présente le consortium RATHER qui sert de jeu de données d'illustrations. La troisième partie présente au lecteur les méthodes de classification non supervisées et les critères de validation des classifications. Une attention particulière est accordée à la stabilité de clustes et aux scores de comparaison des clusters.

1.1.2. Plan

In order for the reader to easily navigate in between these three topics (statistical methods, biological application and data normalisation), here comes a small plan of the thesis (data cohorts are indicated in bold):

- **Chapter 1**: State of the art and Introduction
  - Introduction to the subtyping of TNBC
  - Presentation of the **RATHER consortium**
  - Introduction to unsupervised clustering methods
- **Chapter 2:** Proteomic classification of TNBC
  - Proteomic based classification of two large cohorts of TNBC tumors: the **RATHER consortium** and the **Curie dataset**
- **Chapter 3:** Adjusting the adjusted Rand Index
  - A methodological contribution presenting the Modified version of the Adjusted Rand Index ($MARI$) and its implementation in `aricode`. The $MARI$ is also compared to the ARI of Hubert and Arabie (1985).
- **Chapter 4:** Cluster Stability for class discovery.
  - Part 1: Presentation of the R package `clustRstab`that I implemented
  - Part 2: Presentation of a simulation study and an application study on the **NCI60 dataset** (Ross et al., 2000) investigating when and under which conditions cluster stability can be used as a method for selecting the number of groups in unsupervised clustering.

- **Chapter 4:** TNBC classification of the TCGA dataset
  - A TNBC classification study using the cluster stability methods that I developed in my thesis and comparing the results to the Lehmann et al. (2011) classification using the TNBCtype tool Chen et al. (2012) and the **TCGA dataset**.
- **Appendix** TTKi for molecular drug discovery in TNBC
  - A supplementary project investigating the inhibitor of the kinase TTK as a potential treatment for TNBC by analyzing transcriptomic and proteomic expression in **TNBC cell lines**.

1.2. Breast Cancer

Breast cancers represent the most common cancers in women with around 50 000 new cases each year in France. These cancers are heterogeneous, and the different subtypes differ to such an extent that they are considered as different pathologies (Harbeck and Gnant, 2017). There are three main subtypes of breast cancer, (1) hormone-positive cancer, cancer over expressing estrogen receptors (ER) and/or progesterone receptors (PR), (2) HER2-positive cancer, cancer over expressing epidermal growth factor receptor 2 (HER2) receptors and (3) Triple Negative Breast Cancers (TNBC).

1.2.1. Triple Negative Breast Cancers (TNBC)

TNBC subtype is an aggressive cancer that represents 12-17% of all breast cancers (Severson et al., 2015). It is characterized by a low/lack of expression of ER and PR receptors and by a lack of HER2 over-expression. TNBC is associated with a high risk of recurrence and early metastatic dissemination, especially in lung and brain (Foulkes et al., 2010).

In contrast with the two other main subtypes of breast cancer, TNBC cannot be treated with endocrine therapy or therapies targeting HER2 (Foulkes et al., 2010). The main treatment for TNBC is still chemotherapy and prognosis remains poor (Linn and Van't Veer, 2009). Thus, the identification of molecular based therapeutic targets in order to increase the survival rate of TN patients is a priority in oncology. The difficulty of this task is mainly related to the high heterogeneity among TNBC tumors. Research is therefore conducted in order to determine homogeneous subgroups in TNBC samples.

Indeed, finding such subgroups would be the first step in order to identify targeting treatments adapted for each group. A usual manner to tackle this task is to use unsupervised classification methods and classify patients based on their tumor extracted omic data. This subtyping strategy has already been successful for TNBC patients with a BRCA1/BRCA2 mutation at a genomic level, the hereditary version of breast cancer, whom are treated with PARP-inhibitors.

Several classifications based on transcriptomic data for TNBC samples have been proposed during the last ten years, (Bonsang-Kitzis et al., 2016; Burstein et al., 2014; Jiang et al., 2019; Lehmann et al., 2011, 2016). Hence, these classifications ignore the analysis of tumors at the protein level. Yet, proteins play a major role in cellular functions by regulating signaling pathways[1]. At the beginning of my thesis, our hope was to recover a more meaningful subtyping of TNBC using proteomic data. Indeed, knowing the level of activation of the signaling pathways in the different TNBC subgroups could be a key indication to understand the biological mechanisms involved, and to identify some therapeutic targets.

1.3. Triple Negative Breast Cancer subtyping

1.3.1. Measuring the transcriptome and proteins for TNBC subtyping

In order to measure the transcriptome or the protein expression in breast cancer, it exists different techniques. These techniques will be briefly introduced in the next section.

Transcriptomics

The transcriptome (or RNA) expresses the information content of an organism, encoded from the DNA into different transcripts. A simplified schema for the gene-transcriptomic link is represented in Figure 1.1. Transcriptomics is a

[1] A signaling pathway describes a group of molecules in a cell that work together to control one or more cell functions, such as cell division or cell death. After the first molecule in a pathway receives a signal, it activates another molecule. This process is repeated until the last molecule is activated and the cell function is carried out.

commonly used technique for studying an organism's transcriptome, that is, the sum of all of its 20 000 − 30 000 RNA transcripts (Glaves and Tugwood, 2011). Different techniques exist for this task whereof micro-array RNA and RNA-seq are the most common. Both these methods capture a 'snapshot' in time of the total transcripts present in a bulk of cells or in a sample.

Transcriptomic-based classifications of breast cancer has allowed to identify gene signatures enriching the (immuno-histochemistry based) diagnosis for the TNBC, hormone-positive and HER2-positive subtypes (Liu et al., 2016; Perou et al., 2000). As variations within TNBC are rather subtle compare to differences between TNBC and hormone-positive or HER2-positive tumors, studies analyzing all breast tumors simultaneously have not been able to identify TNBC subgroups. Hence, several groups have, in the last decade, proposed TNBC classifications, based on only TNBC tumors and using mainly transcriptomic data, see Section 1.3.2.

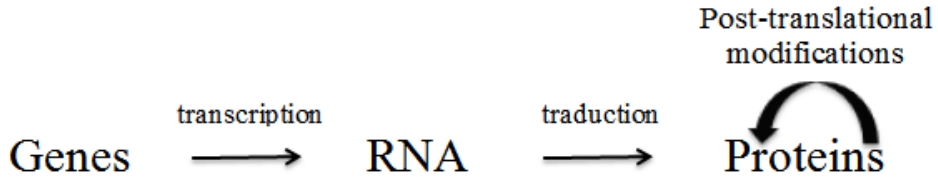


Figure 1.1 – Central dogma of molecular biology, simplified approach for the cell machinery, gene - transcriptom - protein link.

Proteins

Proteins are large biomolecules consisting of one or more long chains of amino acid residues. They are synthesized from the transcriptome by the process of translation in the cell ribosome. However, the level of proteins is not proportional to the level of RNA. Indeed, proteins are regulated by post-translational modifications, such as phosphorylation[2], which cannot be seen at the transcriptomic level (Johnson, 2009). As a consequence of these post-translational modifications, it is not always possible to predict protein levels using transcriptome. Again, see Figure 1.1 for a simplified schematic representation of the links between genes, transcriptom and proteins.

[2]For instance, a phosphoryl group (phosphate) is transferred from ATP by kinases onto particular amino-acids (serine, threonine, tyrosine) of specific proteins then regulating their activity.

Proteins regulate intra-cellular pathway signaling. It has been shown that deregulation and abnormal activation of signaling pathways contribute to tumorigenesis (Giancotti, 2014). The deregulation of ER, PR and HER2 pathways is well known for breast cancer (Song et al., 2014) and the PI3K/AKT and the Wnt pathways are deregulated in the TNBC subtype (Maubant et al., 2015; Timperi et al., 2020).

Measuring the expression of proteins is more difficult to implement than the measure of transcriptome expression. For quantitative studies, two main techniques exist which are: mass spectrometry and Reverse Phase Protein Arrays (RPPA). Whereas mass spectrometry allows an automatic analysis detecting a large scale of proteins in a single sample, the RPPA is in principle more adapted for classification studies since it allows to measure the protein expression for a large number of samples simultaneously. This technique has been used by the lab of Thierry Dubois in two large TNBC projects, the RATHER consortium and the Curie project that I will use in this thesis.

Reverse Phase Protein Arrays (RPPA). The RPPA technique allows to study protein expression levels and the activity status of proteins, by analyzing their phosphorylated state, in a large number of samples simultaneously (Akbani et al., 2014). In the RPPA analysis, a dedicated arrayer prints 1 ng of proteins extracted from tissues or cell lines, onto nitrocellulose covered microscope slides (arrays). Samples are printed in five serial dilutions and each dilution in several replicates. Highly specific primary antibodies, recognizing specific protein and their phosphorylated form, are then applied to quantify protein expression and activation. For more information see the RPPA webpage and Figure 1.2 for a schematic representation.

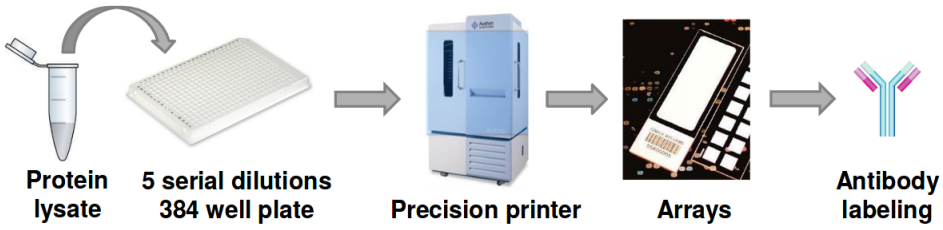


Figure 1.2 – Scheme of RPPA analyses.

The RPPA technology is more sensitive to low-signal proteins than mass spectometry technologies (Boellner and Becker, 2015), this makes it suitable

for analysing tumor samples for which only small quantities of biological material is available. However, the RPPA technology is also more sensitive to experimental procedures, batch and spotting effects.

Moreover, whereas, intraplatform RPPA results have been shown to be consistent and robust to parameters such as the temperature (Hennessy et al., 2010), interplatform reproducibility still needs to be proven. Indeed, some studies show high interplatform variability for RPPA results (Neeley et al., 2012; Troncale et al., 2012), for more information see Byron (2019).

In a recent preprinted bioRxiv study conducted by Byron et al. (2019), the authors showed that proteomic analyses of cancer cell lines using three different RPPA platforms can identify concordant profiles of response to pharmacological inhibition, even when using different antibodies to measure the same target antigenes. The authors argue that these results highlight the robustness and the reproducibility of RPPA technology. Nevertheless, it should be noticed that only well-known proteins were studied and cell-lines are much easier to study than clinical samples, containing far less material. It is important to keep this in mind when comparing results obtained from different RPPA data platforms. Indeed, it seems that the use of this technique can induce some artificial variations to the dataset. I will come back to this in Chapter 2.

Also, compared to RNA analysis, it should be noticed that the proteins in the RPPA technique have to be selected a priori. As a consequence, the analysed proteins differ from different studies, making the comparison of classifications based on different cohorts difficult. Finally, the RPPA community is much smaller than the RNAseq/transcriptome community implying that tools to analyze and normalize RPPA data are not as mature and robust.

1.3.2. Transcriptomic and proteomic based classification of TNBC

Transcriptomic classifications of TNBC

The first study to propose a classification of the TNBC was conducted by Lehmann et al. (2011). Based on a large number of samples, extracted from 21 public datasets (14 for the training set and 7 for the validation set) and

the most variant genes (SD[3] > 0.8), they proposed a classification with 7 subgroups. A molecular signature was found for 6 of these subgroups which they named: basal-like 1 (BL1); basal-like 2 (BL2); immunomodulatory (IM); mesenchymal (M); mesenchymal stem–like (MSL); luminal androgen receptor (LAR); and unstable (UNS). The basal-like subgroups are related to an increased proliferation and have as well been found in more recent classifications. Based on the most differentially expressed genes among these subgroups, they propose a transcriptomic signature[4] of $p = 2188$ genes, that can be used for classifying other TNBC datasets. Also, these authors proposed a tool, TNBCtype allowing to subtype an input TNBC dataset into these 7 subtypes (Chen et al., 2012). This is done by conducting correlations between the input samples and the centroids of the initial 6 Lehmann subtypes. Later, Lehmann et al. (2016) refined their classification excluding the groups IM and MSL which they argued were the result of infiltrated non-tumoral cells. They showed that the IM subgroup is correlated with lymphocytes (white blood cells in the immune system) and that the MSL group is correlated to tumor-associated stromal cells. The patients in these groups were re-assessed to their second highest correlated centroid subtype.

The Lehmann et al. (2011, 2016) studies were criticized for performing the normalization of their data without taking into account any center effect. Bonsang-Kitzis et al. (2016) therefore proposed a classification based on a large set of samples extracted from 7 public datasets which they normalized center by center. Bonsang-Kitzis et al. (2016) also argued that the gene signature proposed by Lehmann et al. (2011, 2016) consists of a too large number of genes $p = 2188$, which might induce instability. Bonsang-Kitzis et al. (2016) therefore searched for a smaller and more refined group of genes upon which they based their classification. To do so, they classified the 830 most variant genes (SD > 0.80) in order to find a gene signature allowing to identify TNBC subgroups. They identified 4 gene subgroups and conducted biological-network analysis for each group in order to select the most important genes. They kept the 167 most important genes and decided, based on biological intuition, to split two of the groups into two, displaying minor differences. Their six gene groups were named: Immunity1, Immunity2, Proliferation/DNA damage, AR-like, Matrix/Invasion1 and Matrix2. They then used these genes to classify the

[3] *standard deviation* (SD)

[4] With an abuse of notation, this transcriptomic signature will sometimes be referred to as a gene signature.

TNBC samples. Their classification was statistically different from the one of Lehmann et al. (2011), however, some of the groups are overlapping. Indeed, the samples with high expression of Matrix/Invasion 1 and Matrix 2 genes, respectively, tended to be classified as the M or MSL Lehmann et al. (2011) groups, the samples with a strong expression of Immunity2 genes tended to be classified as the IM Lehmann et al. (2011) group and the samples with strong expression of AR-like genes tended to be classified as of the LAR Lehmann et al. (2011) group.

Burstein et al. (2014) proposed a classification of TNBC in four subgroups. This classification was based on a smaller number of patients coming from two different centers. However, their transcriptomic data were extracted together, avoiding any bias linked to the site of transcriptomic extraction. They found four subgroups that they labeled: the luminal androgen receptor (LAR), mesenchymal (MES), basal-like immunosuppressed (BLIS), and basal-like immune-activated (BLIA) groups. These authors showed that the prognosis was worse for BLIS tumors than for BLIA tumors. When compared to the classification of Lehmann et al. (2011) they found that the LAR and MES groups of the two studies are overlapping but that their other 2 subtypes (BLIS and BLIA) contained a mixture of the other 4 Lehmann subgroups. They did not manage to replicate the classification of Lehmann et al. (2011) when they used their $p = 2188$ gene signature.

Proteomic classifications of TNBC

Finally, Masuda et al. (2017) proposed a classification of TNBC based on 108 full proteins and 46 phosphorylated proteins obtained by RPPA. They found two groups characterized by different protein signatures. In order to compare to the classifications of Lehmann et al. (2011, 2016) they used another classification method and obtained 5 groups that were significantly related to the Lehmann et al. (2011) classification (6 groups) but not to the one of Lehmann et al. (2016) (4 groups). Even though these results are promising, they are quite difficult to interpret since the criterion for the group selection was not clearly stated (allowing the authors to present two different classifications upon the same samples), and they did not validate the classification upon another TNBC cohort nor did they search a genomic signature for the different groups. I will come back to this classification in Chapter 2.

TNBC classifications: conclusion. All these TNBC classifications find TNBC subgroups with molecular signatures. These signatures seem to be relatively related to each other. This is promising since it indicates that there is a robust biological signal present in the TNBC tumor RNA expression. Still there are some important differences between the different classifications in terms of the number and types of groups. This confronts us with the following questions:

1. How can we explain these differences?
2. If we take into account all these classifications for a given patient, what would the clinical interpretation be?

Some of these differences might surely come from the fact that the different studies used different patient cohorts, with patients coming from different centers and with the data extracted from different platforms. Another reason is certainly that the used analysis pipelines were different, different normalization and different gene-selection procedure *etc.*. These issues will be discussed in the next section.

1.3.3. Methodological procedures for TNBC subtyping

When studying these different TNBC classification studies one realises that several methodological decisions have to be made at different steps of the classification procedure. These decisions ranges from the choice of samples to include and the choice of data normalization procedures, to the classification method to be employed. These questions are of different nature and might concern biological or methodological issues. The impact of the choices has never really been investigated but should not be neglected. For example, as pointed out by Lehmann et al. (2011), if the data for the TNBC tumors are normalized with the data of the tumors of other subtypes of breast cancer, a bias can be induced. Also, Bonsang-Kitzis et al. (2016) pointed out the importance of normalizing data center by center in order to not induce a center-related effect in the data. These two examples concerns only the normalization procedures of the data but similar choices have to be done for the classification method to be used or the data to be included in the study. For example, all the cited TNBC classifications are based on different patient cohorts and different genes (or proteins). A question that arises is then, if too many of these biological, statistical, computational and implementation choices differ

between studies, are we still able to compare their results? And if yes, what are we really comparing? The hope is that it is the biological signals in the datasets that are being compared, however, taking into account any methodologically induce effect is important for an enlightened interpretation of the results. As a start, I present a non-exhaustive list of such methodological choices that have to be considered when conducting TN classification.

- **Which patient cohort to use?**

  The classification could be based on a smaller cohort with patients coming from one center to avoid center effects or be based on several public datasets in order to increase the number of samples. Another alternative is to properly model the center effect. It could also be based on cell lines instead of patients.

- **What inclusion criteria to use for the TNBC patients?**

  The inclusion criteria could for example be based on TNBC diagnostic, the transcriptomic signature or immunohistochemestry of ER / PR / HER2, a combination of all or other. This implies that two studies based upon the same cohort will not include the same patients if they do not share the same inclusion criteria.

- **What omic technology should be used to extract the data? And which genes or proteins should be included?**

  So far, most classification studies of TNBC patients are based on transcriptomic data. As a consequence of the large number of RNA transcripts, most authors conduct a gene selection procedure, in most cases by basing the classification on the most variant genes, see Section 1.3.2. Since the set of most variant genes differs from different cohorts, the genes used to classify the TNBC tumors also differ from a study to another. Other gene selection criteria could also be considered, for example, considering bimodal looking genes.

  Recent improvements of proteomic expression technologies (see 1.3.1), have made it possible to base the TNBC tumor classification study on

proteomic data using the RPPA technique. However, a selection of proteins is made on available anti-bodies and this selection will differ for different cohorts.

No matter what kind of data that is used, a tumor sample is extracted from the patient. This sample contains breast cancer cells but also tumor-associated cells such as stromal and blood cells. The biological analysis of transcriptome or proteins will therefore take into account all these types of cells, which can, to some extent, bias the interpretation of the analysis based on the transcriptomic/proteomic data. An example of such bias is presented in Lehmann et al. (2016) described in Section 1.3.2. Indeed, these authors found that two of their TNBC subtypes were linked to non-tumoral cells. Signel cell could be an alternative to this, however this method is much more expensive.

- **What statistical method to use in order to classify the tumor samples?**

  The classification of TNBC tumors is an inherently unsupervised task where we look for groups that characterize the samples and that could correspond to different cancer subtypes. A very popular method to resolve this task in the TNBC community is the *Concensus Clustering* based on the stability of clusters and proposed by Monti et al. (2003). However, a multitude of clustering methods, with different properties, exist, and some of them are presented in Section 1.5.

- **How to select the number of TNBC groups?**
  Selecting the number of groups for the TNBC subtypes is biologically and statistically difficult. This is probably linked to the heterogeneity of the tumors. Also, there is not necessarily a partition structure with clearly defined groups able to explain the structure of a TNBC dataset.

  By consequence, the number of groups is often selected by hand and changed between studies, as in (Lehmann et al., 2011, 2016) (passing from 7 to 5 TNBC groups) or within studies as in (Masuda et al., 2017) (passing from 2 to 5 TNBC groups) and Bonsang-Kitzis et al. (2016) (passing from 4 to 6 gene subgroups). Inducing too much subjectivity in the choice of the number of groups raises the question of scientific reproducibility.

All the presented TNBC classification studies have used the stability of clusterings to select the number of groups (see Section 1.5.3). The stability criterion is appealing in terms of interpretation. Indeed, it is clear that an unstable classification should not be considered. It is, however, not clear (as discussed later) that one wants the most stable classification or in fact whether it is indeed possible to measure the stability. These issues will be investigated in Chapter 4. Also, many other methods exist for selecting the number of groups in clustering and some of them are presented in Section 1.5.2.

- **What statistical methods to use in order to validate the classification?**

  What is usually done in unsupervised classification studies is to cluster a training set and then validate this clustering using a supervised validation procedure on a different validation set. However, in most TNBC classification studies the method for clustering the training set is the same method used to cluster the validation set. Hence, the acquired knowledge from the training set is not used in order to classify the validation set and the two clusterings can only be compared at a descriptive level. This makes the validation procedure unnecessarily unclear.

During this thesis I tried to perform the TNBC classification analysis in a statistically more rigorous manner by taking into account these different questions. For this aim, I had at my disposal the RATHER consortium dataset (described hereafter). As will be seen throughout this thesis, classifying TNBC tumors is a difficult task for which no easy or clear answer might be provided as it is today.

1.4. The RATHER consortium

The Rational Therapy for Breast Cancer consortium (RATHER), described in Michaut et al. (2016), contains proteomic and transcriptonic data from TN breast cancer patients collected from the Netherlands Cancer Institute (NKI), Amsterdam and from the Addenbrooke's Hospital, Cambridge, UK. The collected samples in the RATHER set contain at least 30% of cancer cells.

Survival data with an average of 10 years of surveillance were collected for the patients. The proteomic data were extracted together for the two cohorts at Institut Curie. This is an advantage compared to studies where the proteomic data for different cohorts have been extracted separately, hence inducing a platform-related bias.

1.4.1. TNBC inclusion criteria

258 TNBC and lobular (hormone positive) breast cancer samples were collected and analysed together with breast cancer cell lines ($n = 56$). The definition of TN breast cancer can be based on different criteria and might thus differ among studies. We based the inclusion criteria of TNBC tumors on two criteria: TN on diagnosis and ER, PR and HER2 negative on TargetPrint (RNA analysis). This resulted in $n = 99$ included samples, 67 from NKI and 32 from Cambridge dataset. To this count, samples with more than 80% of missing proteomic data ($n = 2$) were excluded from the NKI dataset. We did not use the criterion based on tissue microarray measure of immunohistochemistry since there was more than 20% of missing value for this criterion.

1.4.2. Clinical and demographic description

Demographic and clinical information about the included patients is illustrated in Table 1.1 and survival time is plotted in Figure 1.3. Clinical information was missing for four patients from the Cambridge cohort (referred from hereon as CAM) and three patients from the NKI cohort. As it can be seen in Table 1.1, the average age of the patients is 54 years with a large variance which is expected since TNBC affects younger women than other types of breast cancer Badve et al. (2011). It should also be noted in Figure 1.3 that most relapse and metastasis appear 5 years after the diagnosis which is characteristic for the TNBC.

More importantly, there is a significant difference in age, received treatment and survival rate between the two patient cohorts (CAM and NKI). Few studies show this kind of preliminary analyses before classifying the TNBC samples coming from different cohorts. However, not doing so might dissimulate "hidden" center effects.

1.4.3. RATHER-NKI: the illustration dataset

Since the two different cohorts differ in clinical aspects, only the NKI TNBC cohort is going to be used as an illustration set. I used the TNBCtype tool proposed by Chen et al. (2012) to obtain the corresponding Lehmann sub-types of the patients from this dataset. The results of this classification are shown in Figure 1.4. However, the predicted subtypes of some patients depended on the subset of the patients included in the analysis and this classification should therefore be interpreted with caution. As shown in Figure 1.3, there is no difference in survival rate between the different Lehmann sub-types of the NKI cohort.

1.5. Unsupervised classification and clustering

Since no ground truth classification of TNBC exists, the TNBCtool, like any classification of TNBC tumor samples, is based on unsupervised clustering methods, hereon referred to as clustering. Clustering is a method widely used for exploratory data analysis. In difference with unsupervised classification methods, clustering methods do not assume an underlying model on the data. They use similarity or distance functions to regroup data points that are somewhat similar. The resulting subsets are formed such as the data points within the same subset are more similar to each other than to those in other subsets (Lebarbier and Mary-Huard, 2008). Since no prior knowledge of the classification or group belonging is available, the signal of interest is unknown. In other words, we don't know what we are looking for, and once a signal is detected, we do not know whether it is of any biological interest. The ideal would be to find a group pattern that corresponds to disease response in some way, but such a signal might as well, if it exists, be drowned in noise or experimental effect such as a center effect, batch effects or other. One could also imagine that the signal of biological interest is in the low varying genes which in many cases are eliminated by the initial gene selection threshold. Without being able to compare the obtained classification, to an initial, or true classification, it is very difficult to validate it and to know whether it is of any interest or not.

Some of the most popular cluster algorithm methods are reviewed in the next section. For the interested reader, clustering will be defined in a statisti-

	CAM (N=28)	NKI (N=64)	Total (N=92)	p-value
Age at diagnosis				**0.002[1]**
Mean (SD)	60.68 (11.17)	51.42 (13.83)	54.24 (13.70)	
Range	34.00 - 78.00	26.00 - 83.00	26.00 - 83.00	
Tumor size cm				0.444[1]
N-Miss	0	3	3	
Mean (SD)	3.05 (1.85)	2.78 (1.40)	2.87 (1.55)	
Range	1.50 - 8.40	1.00 - 7.50	1.00 - 8.40	
Nb. of pos. lymph. nodes				0.238[1]
N-Miss	2	1	3	
Mean (SD)	2.00 (4.52)	1.14 (2.28)	1.39 (3.10)	
Range	0.00 - 21.00	0.00 - 11.00	0.00 - 21.00	
Pathologic m1 at diagnosis				**< 0.001[2]**
N-Miss	3	0	3	
M0	25 (100.0%)	64 (100.0%)	89 (100.0%)	
Treatment surgery				**0.083[3]**
N-Miss	1	0	1	
conserving	12 (44.4%)	41 (64.1%)	53 (58.2%)	
mastectomy	15 (55.6%)	23 (35.9%)	38 (41.8%)	
Treatment hormonal				**0.015[3]**
FALSE	27 (96.4%)	48 (75.0%)	75 (81.5%)	
TRUE	1 (3.6%)	16 (25.0%)	17 (18.5%)	
Treatm. adjuv. chemother.				0.113[3]
FALSE	9 (32.1%)	32 (50.0%)	41 (44.6%)	
TRUE	19 (67.9%)	32 (50.0%)	51 (55.4%)	
Treatment radiotherapy				0.295[3]
FALSE	9 (32.1%)	14 (21.9%)	23 (25.0%)	
TRUE	19 (67.9%)	50 (78.1%)	69 (75.0%)	
Histological grade				**0.045[1]**
N-Miss	2	7	9	
Mean (SD)	3.00 (0.00)	2.86 (0.35)	2.90 (0.30)	
Range	3.00 - 3.00	2.00 - 3.00	2.00 - 3.00	

Table 1.1 – Clinical and demographic information of the RATHER consortium. N-Miss indicate missing values. Clinical information was completely missing for 3 patients from the NKI dataset. Footnotes indicate different tests with: 1. Linear Model ANOVA, 2. Chi-squared test for given probabilities, 3. Pearson's Chi-squared test.

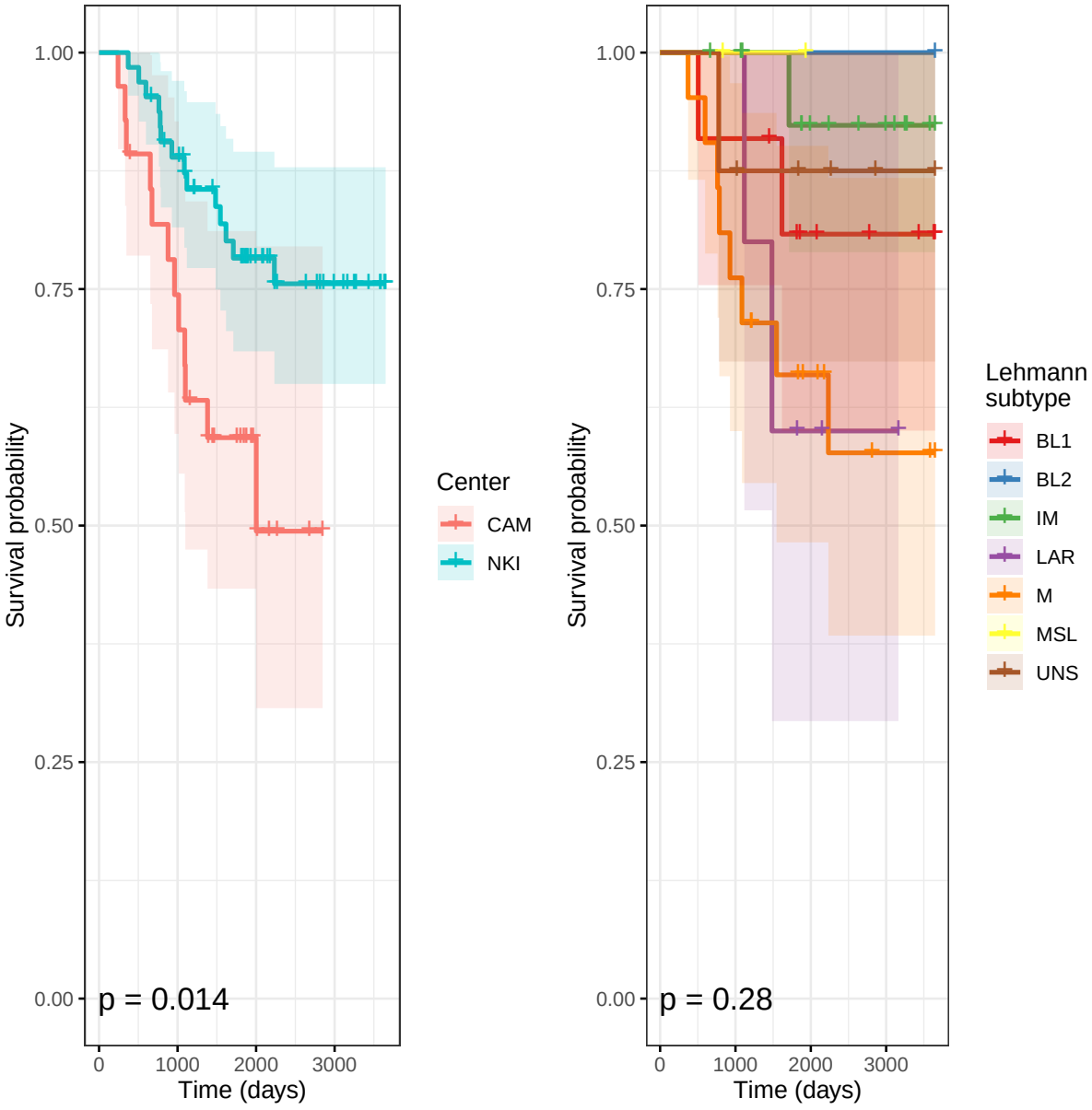


Figure 1.3 – Survival rates for RATHER patients, the survival curves are constructed by the Kaplan-Meier estimator and the p-values correspond to the Log-rank test between NKI and CAM patients (left) and the Lehmann subtypes within the NKI cohort (right).

cal context in Chapter 4. In Section 1.5.2 different methods for selecting the number of groups in clustering will also be presented. The method for selecting the number of groups based on stability will be given a particular attention. Indeed, this method has, so far, been used in almost all TNBC classification studies. This method is based on a heuristic and little is still known theoretically about this method. One of the aims of my thesis is therefore to get a better understanding for this method, in order to see whether it is a good tool for classifying TNBC tumors.

1.5.1. Clustering methods

Different kinds of clustering algorithms exist, many of them try to minimize an objective criterion such as the within cluster homogeneity or the correlation between datapoints. Since different cluster algorithms use different criteria, the obtained classification will depend on the clustering method used. This is the case, for the classification of the NKI cohort presented in Figure 1.4 where the NKI cohort is consistently clustered in four groups ($K = 4$), by some of the most popular clustering algorithms; k-means, hierarchical ascendant clustering (HAC) and Gaussian mixture models (GMM) and using the Lehmann et al. (2011) gene signature, it shows some clear differences between the obtained classifications.

K-means

The K-means algorithm search to minimize the distance between each observation $(x_1, \ldots, x_n)$ and its barycenter μ_{c_k}, by minimizing the euclidean L^2-norm (the *within-cluster sum of squares*). The obtained classification $C = \{c_1, \ldots, c_K\}$ of K groups is then defined as:

$$C = \underset{C \in C_{K_{all}}}{Argmin} \sum_{k=1}^{K} \sum_{i \in c_K}^{n_k} \|x_i, \mu_{C_k}\|^2, \tag{1.1}$$

where C is the obtained clustering and $C_{K_{all}}$ is the set of possible partitions of $x_1, \ldots, x_n$ in K classes. To minimize this distance, the k-means algorithm uses an iterative refinement technique. First it assigns each observation to its closest barycenter (mean) and then given all the obtained clusters, redefine the barycenters. For this reason, the initial barycenters have to be given to the algorithm which can be done by some a priori knowledge about the groups, or initialized randomly. Depending on these initialization points, the resulting

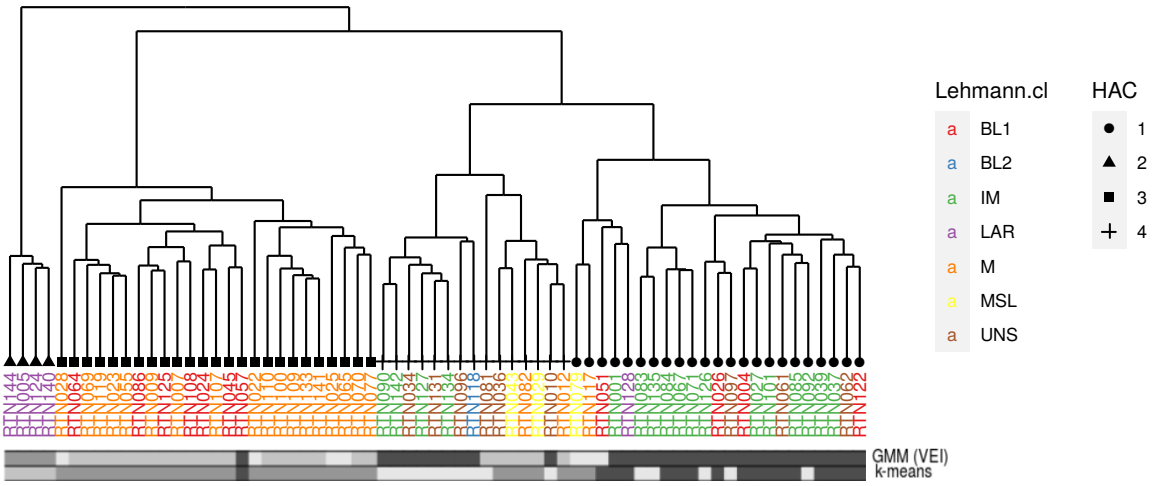


Figure 1.4 – Clustering of NKI RATHER TNBC patients; NKI patients were clustered by the HAC algorithm using the $Ward^2$ distance. The classification is based on the $p = 2188$ genes from the Lehmann gene signature Lehmann et al. (2011). The tree is cut at a level to form four groups, indicated by the leaves symbols. The patient ID is colored according to their subgroup of the Lehmann et al. (2011) classification. Beneath the samples ID are two color bars indicating the group belonging for the patients in the k-means classification (lower color bar) and the GMM classification (upper color bar) for the VEI covariant model, that is diagonal, varying volume and equal shape.

classifications will differ and the algorithm has a tendency to converge to local minimums. To resolve this problem, the algorithm can be run several times with different initialization, and the one giving the smallest *within cluster sum of squares* will then be chosen. A number of variations and improvements of the algorithm have been proposed, for example *kmeans++* (Arthur and Vassilvitskii, 2006) for the initialization.

Hierarchical Ascendant Clustering (HAC)

In the HAC algorithm each observation is considered as a distinct class. The algorithm will then regroup the two closest classes and repeat this until all observations constitute one class. There are different methods to compute the distance between classes. For example, the correlation or absolute correlation by taking $1-r$ respective $1-|r|$ where r is Spearman or Pearson correlation coefficient, or the L^2-norm between barycenters. A commonly used distance is the Ward^2 distance (Murtagh and Legendre, 2011). Ward distance between two classes (c_k and c_ℓ) with barycenters μ_{c_k} and μ_{c_ℓ} respectively, is computed as:

$$D^2(c_k,c_\ell) = \frac{n_k n_\ell}{n_k + n_\ell}\|x_{c_k} - x_{c_\ell}\|^2, \tag{1.2}$$

where n_k, n_ℓ are the number of observations of each class. When using the Ward^2 distance, the fusion of classes is done so that the intra class inertia is minimized at each step. This implies, as for the k-means algorithm that global optimality is not guaranteed. However, the hope is that, by doing local optimal fusion at each step, the obtained partition will be close to an optimal partition (Lebarbier and Mary-Huard, 2008). The obtained "tree", can then be cut at different heights with different numbers of groups as a result. The dendrogram of the RATHER NKI data is illustrated in Figure 1.4. The dendrogram was obtained by computing the Ward^2 distance and was then cut at the level for $K = 4$, the colors of the labels correspond to the Lehamnn TNBC subtypes.

Gaussian mixture models

The Gaussian mixture models (GMM) is an unsupervised classification method. Its principle is to fit K Gaussian distributions upon the n observations $(x_1,x_2,\ldots,x_n)$.

The probability to observe x is defined as,

$$\mathbb{P}(x) = \sum_{k}^{K} \pi_k \mathcal{N}(X|\mu_k, \sigma_k), \tag{1.3}$$

where π_k is the probability of group $k \in 1, \ldots, K$, with $\sum_{k}^{K} \pi_k = 1$, μ_k the mean of the group k and σ_k is the covariance (Friedman et al., 2001).

To find the most probable set of distributions, the likelihood is maximized. For this, let the $x_1, \ldots, x_n$ observations be drawn from n random variables, noted $X_1, \ldots, X_n$, assumed to be derived from the population distribution to which the observation belongs. The logarithm of the likelihood is then defined as:

$$\log \mathcal{L}(X,Z;\phi) = \log \left\{ \prod_{t=1}^{n} \prod_{k=1}^{K} [\pi_k f(X_t; \mu_k, \Sigma_k)]^{Z_{tk}} \right\} \tag{1.4}$$

where

$$Z_{tk} = \begin{cases} 1, & \text{if individual t belongs to population k} \\ 0, & \text{otherwise.} \end{cases}$$

Since Z is not observed, $\log \mathcal{L}(X,Z;\phi)$ is estimated by the Expectation Maximization (EM) algorithm (Dempster et al., 1977). This algorithm is illustrated in Algorithm 1 (below).

Algorithm 1 Expectation Maximization

1: Initialize values for model parameters
2: **E (Expectation) step**: Given the observed data and current estimate of model parameters, compute the expected value of $log\mathcal{L}(X;\phi)$
3: **M (Maximization) step**: compute the parameters which maximize the current $\log \mathcal{L}(X;\phi)$, if there is no convergence, return to the E-step
4: Stop when there is convergence

The obtained groups can be modeled with different covariance models, assuming the shape and size of the groups. For a more detailed description of the Gaussian mixture models see for example Picard (2007).

Clustering TNBC tumors, the RATHER-NKI dataset

As it can be seen in Figure 1.4, there are some similarities but also some differences between the different classification even though they are all based

on the same samples and variables. Hence, before interpreting the groups of a clustering, it is important to notice that the obtained groups depend on the chosen algorithm and its parameters. In practice the choice of clustering method is rarely discussed in applied papers. It would be too tedious to test them all, but testing a few would seem like a good idea. Also, here I set the number of groups to $K = 4$ however, it could as well be set to $K = 1, \ldots, n$ and each algorithm will be able to propose a classification with as many groups as indicated. Without any idea of the true number of groups, the choice of K turns out to be difficult to make. This might be one of the reasons why, in TNBC application studies, the number of groups vary, from one classification to the next (Lehmann et al., 2011, 2016), or within the same study (Masuda et al., 2017) (see Section 1.3.2 for more details).

1.5.2. Selecting the number of groups in clustering

How to select the number of groups in unsupervised classification is an open question in statistics. This number is unknown and has to be estimated. How to do so is a question that is inherently difficult to answer, and for this reason, several heuristics have been developed. Today these heuristics are largely used and accepted within the statistical community. Many of these methods aim to optimize a criterion ($Crit$) of the classification, for example the overall *within sum of squares* or the stability of the clustering. Clustering is then performed for the same dataset for different values of K and the value optimizing $Crit(C_K)$ is then selected.

I am now going to present some of the most popular and standardized methods for selecting the number of groups. For a more detailed and comprehensive description of these methods, see Kassambara (2017).

The '*elbow*' method

The aim of the '*elbow*' method is to optimize the *within sum of squares* WSS (see the k-means algorithm) of the classification. The problem is that WSS decrease with the number of groups increasing. Hence, the criterion cannot be minimized since this implies selecting the largest number of groups ($K = n$) which is not of much interest interpretation wise. The aim is therefore to select K for which there is an observed change of steepness in the WSS curve. That

is, the difference of WSS is large for the number of groups inferior to K and then becomes smaller for the number of groups superior to K. This kind of judgment makes the '*elbow*' method a rather subjective criterion.

Average Silhouette

The average silhouette (Rousseeuw and Kaufman, 1990) computes the average distance for each observation x_i and (1) the observations belonging to the same cluster $a(i)$ and (2) the observations belonging to the nearest cluster $b(i)$. It takes the following form :

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

The larger the $s(i)$ the better x_i is clustered, hence, the value K can be selected as the one maximizing the average silhouette for $i = 1, \ldots, n$ (Fischer, 2011).

Information criterion

One can also optimize an information criterion. This is only possible for model-based cluster algorithm such as the GMM, which is their big advantage. K can then be estimated by maximizing the Bayesian Information Criterion (BIC), defined as:

$$Crit(C_K) = \log \mathcal{L}(X; \phi_K) - \frac{\log(n)}{2} \times \text{params}, \tag{1.5}$$

where params correspond to the number of parameters of the model with K populations (Lebarbier and Mary-Huard, 2008). K is then estimated as the value maximizing the BIC,

$$\widehat{K} = \underset{K=1,\ldots,n}{Argmax}\ Crit(K). \tag{1.6}$$

The BIC is more adapted to density estimation and might not be ideal when components are poorly separated; in this later case the Integrated Complete-data Likelihood (ICL, Biernacki et al. (2000)) might be more appropriate. It

is a classification version of the BIC taking into account the group belonging of the observations. The ICL is difficult to compute and several approaches to approximate it have been proposed, see for example Bertoletti et al. (2015).

The Gap statistic

Selecting the number of groups can also be based on statistical testing. Tibshirani et al. (2001) proposed the gap statistic in order to formalize the '*elbow*' heuristic. The gap statistic compares the total *within intra-cluster dispersion*, W_K (distance between all observations within a cluster), with what would have been expected under the null hypothesis. The null reference distribution of the data is obtained by generating a large number B of datasets with random uniform distribution. The *within intra-cluster dispersion* then is computed for each of these datasets, noted W_{Kb}. This is done for several K and the larger the "gap" the further the observed clustering is from what would have been expected under the null distribution. The gap takes form as:

$$Gap(K) = \frac{1}{B}\sum_{b=1}^{B} log(W_{Kb}^{*}) - log(W_K).$$

The number of groups is then selected via

$$\widehat{K} = \text{smallest } K \text{ such that } Gap(k) \geq Gap(K+1) - SD_{K+1},$$

where SD_{K+1} is the estimated standard deviation of the statistics $Gap(K+1)$.

Cluster Stability

Cluster stability is based on the idea that the more stable a clustering is the more likely it is that it reveals the true structure of the data. Hence, clustering will be performed for different subsets of the data and for different K. K is then selected as the value giving rise to the most stable clustering. Cluster stability is going to be presented in detail in Section 1.5.3.

In the next section I will apply these clustering methods and criteria to the RATHER-NKI cohort in order to estimate the number of groups present in this dataset.

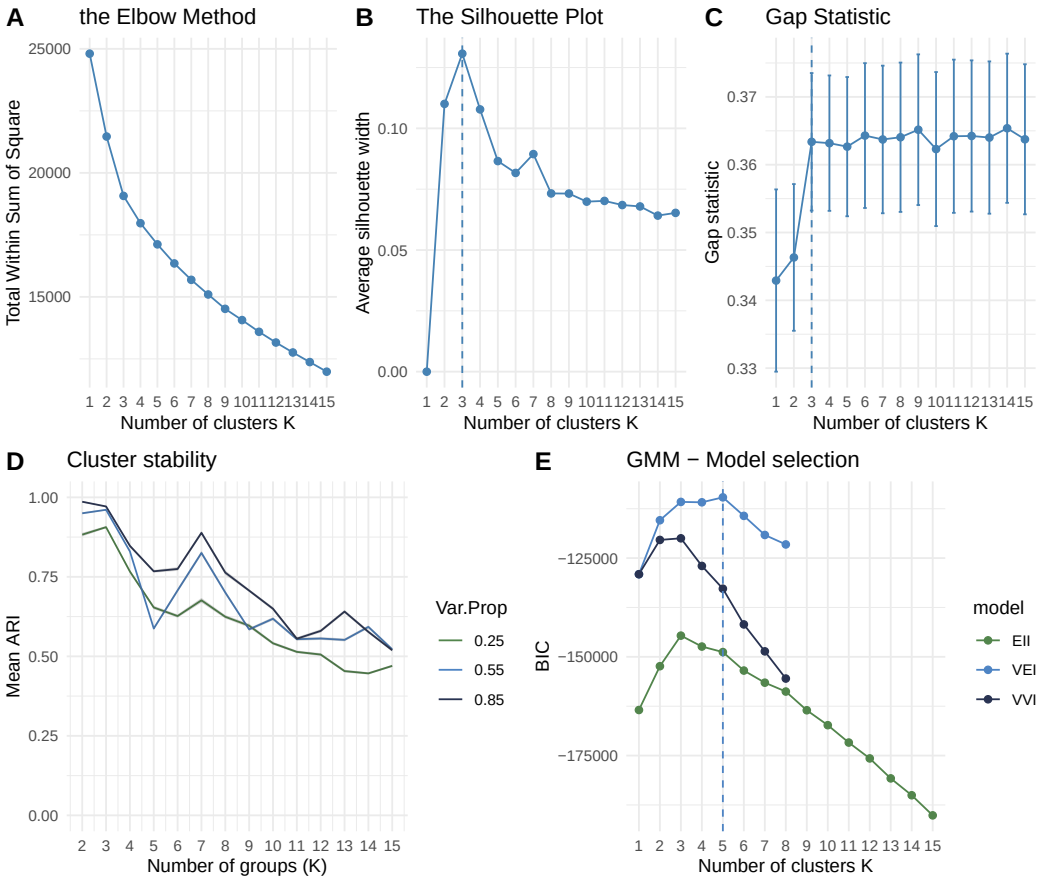


Figure 1.5 – Different methods for selecting the number of groups in clustering. The clusterings are based on the NKI-RATHER cohort using the Lehmann $p = 2188$ gene signature Lehmann et al. (2011). For all methods except for the GMM, the k-means algorithm was used with 30 initializing points. The GMM was obtained using the `mclust` R package (Scrucca et al., 2016a) and the covariance models are: "EII", spherical, equal volume, "VEI", diagonal, varying volume, equal shape and "VVI" diagonal, varying volume and shape. Plot A, B, C and E were obtained using the `factoextra` R package, proposed by `factoextra`. Plot D was construction by the `clustRstab` R package (Sundqvist et al., 2020b), presented in Chapter 4. It uses the ARI score presented in Section 1.5.4, the higher the ARI, the more stable the classification.

Selecting the number of groups of TNBC tumors, the RATHER set

In Figure 1.5 different methods for selecting the number of groups are shown. The '*elbow*' method, the silhouette statistic as well as the Gap statistic seem to propose a cluster structure of three groups, $\widehat{K} = 3$. The best model selected by the GMM (Figure 1.5E), is VEI with a cluster structure of five groups. In coherence with (Figure 1.5A, B and C) the GMM EII (spherical with equal variance) covariance model, *i.e.* the same method as for k-means, maximizes the BIC value for $K = 3$. Figure 1.5D) illustrates the estimated stability for different numbers of groups. The aim of this method is to select the number of groups that yields the most stable clustering (here by a mean ARI close to 1). The stability is assessed using subsampling and I varied the proportion of genes that were subsampled. As it can be seen in the figure, the most stable clustering depends on this subsampled proportion.

As a conclusion, the "best" K for the RATHER-NKI cohort depends on the criterion/method used. It might therefore become problematic if a user only base its choice of K on one of the criteria, without taking into account the others. This is, however, often done in TNBC classification studies where the choice of number of clusters is selected as the one giving the most stable classification without taking into account other criteria. This is a shame since this leaves out important information and impact the resulting clustering.

1.5.3. Cluster stability for selecting the numbers of groups in clustering

The use of cluster stability for selecting the number of groups has become very popular in oncology and RNA based clustering studies. It is particularly popular in the classification studies of TNBC. The philosophy of cluster stability is that a clustering structure on a data set should be stable, see Von Luxburg et al. (2010); Ben-Hur et al. (2001), that is, two clusterings upon the same data, or upon subsampled proportions of the original dataset, should be similar. To measure the stability of a clustering, ideally, one would like to have access to a large number of datasets upon the same n observations and p variables that one could cluster and then compare. However, since the data acquisition procedures are very long, tedious and expensive, especially in biology, in practice we will only have access to one dataset. As a consequence, we can only try to estimate this quantity. However, the estimation is not trivial and involves

a complex parameter setting. The estimation follows, in general, the three following steps:

1. Generate a large number of perturbed datasets from the initial dataset by, for example, subsampling variables or observations.
2. Cluster each perturbed dataset using a given cluster algorithm.
3. Compare the obtained clusterings using:
   (a) A given comparison strategy, for example comparing all or some of the obtained clusterings
   (b) A given cluster comparison score *e.g* the ARI of Hubert and Arabie (1985)

The stability is then computed as the arithmetic mean of the different comparisons. The stability is computed and compared for different numbers of groups and the K is selected as the one yielding the most stable clustering. Several variants of cluster stability have been implemented, for example, the R packages `clv` (Nieweglowski, 2020), `clusterStab` (MacDonald et al., 2018), `ClusterStability` (Lord et al., 2016), `ConsensusClusterPlus` (Wilkerson et al., 2010) `fcp` (Hennig, 2020). Even though they often present them self as "new cluster stability methods" they mostly correspond to a specific parameter setting for the three indicated algorithmic steps. These three steps as well as the different R packages will be presented in detail in Chapter 4. I will now present some of the most popular clustering comparison scores.

1.5.4. Cluster Comparison scores

The aim of a score for clustering comparison is to measure the similarity or dependency between two clusterings. As already described, it is one of the most important steps for estimating the stability of a clustering. Several different scores exist, and some of the most popular include the Jaccard index, the Hamming distance, the Normalized Information Distance (NID) (Vinh et al., 2010), the Rand Index (RI) (Rand, 1971) and the Adjusted version of the RI (ARI) (Hubert and Arabie, 1985). They do all have their different properties but they have in common that they all count, in some manner, the agreement or disagreement of pairs or point of pairs (see Meilă (2003); Vinh et al. (2010)). In this section I will present the ARI and the NID since they are the

two scores that I will use in my thesis. These scores are computed from the contingency table of the two compared classifications. The contingency table for the classifications C^1 and C^2, containing K respective L groups, is shown in Table 1.2.

$C^1 \backslash C^2$	c_1^2	$\cdots$	c_ℓ^2	$\cdots$	c_L^2	Sums
c_1^1	n_{11}	$\cdots$	$n_{1\ell}$	$\cdots$	n_{1L}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k^1	n_{k1}	$\cdots$	$n_{k\ell}$	$\cdots$	n_{2L}	$n_{k.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_K^1	n_{K1}	$\cdots$	$n_{K\ell}$	$\cdots$	n_{KL}	$n_{K.}$
Sums	$n_{.1}$	$\cdots$	$n_{.\ell}$	$\cdots$	$n_{.L}$	$\sum_{k\ell} n_{k\ell} = n$

Table 1.2 – Contingency Table between clusterings C^1 and C^2; each entry $n_{k\ell}$ corresponds to the number of observations in group k in C^1 and group ℓ in C^2.

The (Adjusted) Rand Index

The Rand Index (RI) was proposed by Rand (1971) and counts the pairs of observations that are clustered in the same manner in the first and the second clusterings. That is, either those that are clustered in the same group in both classification or in different groups in both classifications. The first of these types of pairs are referred to as consistent by similarity (a) and the second as consistent by difference (b). a and b can be computed from Table 1.2 as

$$a = \sum_{k,\ell}^{K,L} \binom{n_{k\ell}}{2}, \qquad b = \binom{n}{2} + \sum_{k,\ell}^{K,L} \binom{n_{k\ell}}{2} - \sum_{k}^{K} \binom{n_{k.}}{2} - \sum_{\ell}^{L} \binom{n_{.\ell}}{2},$$

from which the RI is computed as

$$RI(C^1,C^2) = \frac{a+b}{\binom{n}{2}}, \tag{1.7}$$

where $\binom{n}{2}$ corresponds to the cardinal of unordered pairs among the n observations.

The Rand index can therefore be seen as a probability for two observations to be consistent with $RI = 1$ for two identical classifications and $RI = 0$

when one classification only contain one group ($K = 1$) and the other has as many groups as there are observations ($K = n$). The RI has been shown to depend on the number of groups Morey and Agresti (1984). Therefore, different manners to correct the RI for chance have been proposed. Brennan and Light (1974) proposed to correct the RI by its expected value under the null. They proposed to base the expected value on a hypergeometric hypothesis, assuming fixed cluster sizes. This correction was then incorporated into a *Kappa* like score proposed by Hubert and Arabie (1985). They referred to this score as the Adjusted version of the RI (ARI) and is defined as follow:

$$ARI(C^1, C^2) = \frac{RI(C^1, C^2) - \mathbb{E}(RI(C^1, C^2))}{1 - \mathbb{E}(RI(C^1, C^2))},$$

where $\mathbb{E}(RI(C^1, C^2)) = 1 + 2\sum_{k,\ell}^{K,L} \Big/ \binom{n}{2}^2 - \Big[\sum_k^K \binom{n_{k\cdot}}{2} - \sum_\ell^L \binom{n_{\cdot\ell}}{2}\Big] \Big/ \binom{n}{2}$. Morey and Agresti (1984) proposed the same score but instead of basing the expecting value on a hypergeometric hypothesis, they based it on a multinomial hypothesis, which has the advantage to not consider the cluster sizes as fixed. However, they made an error in their calculation of the expected value and proposed the estimator as a plugin. I will finalize their work in Chapter 3.

The Normalized Information Criterion

The Normalized Information Distance (NID, Vinh et al. (2010)), an information based measure with the advantage of having both a metric and normalization proprieties. Basically, it measures to which extent knowing the clustering C^1 would help to predict the clustering C^2 and vice versa. The entropy's $H(C^1)$ and $H(C^2)$ and the conditional entropy $H(C^1|C^2)$ between C^1 and C^2 are necessary to compute the NID which is defined as,

$$NID(C^1,C^2) = 1 - \frac{H(C^1) - H(C^1|C^2)}{max\{H(C^1),H(C^2)\}}, \tag{1.8}$$

where

$$
\begin{aligned}
H(C^1) &= -\sum_{k}^{K} \frac{n_{k.}}{n} log \frac{n_{k.}}{n}, \\
H(C^2) &= -\sum_{\ell}^{L} \frac{n_{.\ell}}{n} log \frac{n_{.\ell}}{n}, \\
H(C^1|C^2) &= -\sum_{k,\ell}^{K,L} \frac{n_{k\ell}}{n} log \frac{n_{k\ell}/n}{n_{.\ell}/n}.
\end{aligned}
\tag{1.9}
$$

It should be noticed that $H(C^1) - H(C^1|C^2) = H(C^2) - H(C^2|C^1)$. The numerator quantifies to which extent the knowledge of C^1 reduces the information, or bits, needed to encode C^2. The higher it is, the more useful the information in C^2 is to predict labels in C^1 and vice-versa. The NID is a distance function and a metric (satisfying the positive definiteness, symmetry and triangle inequality) and is normalized within the range $NID \in [0,1]$, equaling 0 when the two clustering are identical, and 1 when they are independent, *e.i.*, sharing no information with each other. In contrary to the ARI, the NID is not corrected for chance.

Cluster comparison scores depend on the number of groups

An issue with cluster comparison is that the scores depend on the number of groups and therefore needs to be corrected. Indeed, the larger the number of groups is, the more information (or the more similar) will the two clusterings share (be) (Morey and Agresti, 1984; Vinh et al., 2010; Von Luxburg and Ben-David, 2005). This is particularly crucial for cluster stability since (1) the stability of a clustering is estimated as the arithmetic mean of cluster comparison scores and (2) this mean is compared for different number of groups. If these scores are not corrected for chance, the resulting selection of the number of groups can be biased, favoring, by chance, a larger number of groups. Some scores such as the ARI or the $MARI$ (presented in Chapter 3) are intrinsically corrected for chance, whereas others, such as the NID are not.

To illustrate how these scores depend on the number of groups, I conducted a simulation where I compared a clustering with a perturbed version of itself. That is, when a given proportion of the cluster labels have been permuted among the observations. This perturbation procedure was repeated a large

number of times and for different numbers of groups. The similarity of the obtained partitions was measured by the ARI respectively the NID score. The experiment is described in details in Algorithm 2 and the results are to be found in Figure 1.6. As it can be seen in this figure, the mean-value of the NID but not the ARI depend on the number of groups. This is expected since the latter, but not the first, is corrected for chance. However, the standard deviation of both scores depends on the number of groups, and a smaller number of groups seem to generate a higher variation when the perturbation (permutation) level is low, whereas opposite is observed for higher perturbation levels.

Algorithm 2 Cluster comparison score with permutation

1: **for** $K = 2, \ldots, K_{max}$ **do**
2: **Sample** n observations with the probability $1/K$ to belong to a given class $\{1, \ldots, K\}$ to get the initial clustering C
3: **for** *proportion* = 0,05, 0.10, . . . , 1 **do**
4: **for** d = 1,. . . , $nsim$ **do**
5: **Perturb** the obtained classification by permuting a certain *proportion* of the cluster labels to get the perturbed clustering C'_d
6: **end for**
7: **Compare** the obtained perturbed clusters $Sc(C'_d, C'_{d'})$ for $d < d'$
8: **Compute** mean value and standard deviation (SD) of the obtained *scores*:

$$mean(score) = \frac{1}{\binom{nsim}{2}} \sum_{d<d'}^{nsim} \widehat{Sc}(C'_d, C'_{d'})$$

$$SD(score) = \sqrt{\frac{1}{\binom{nsim}{2} - 1} \sum_{d<d'}^{nsim} Sc(C'_d, C'_{d'}) - mean(score)}$$

9: **end for**
10: **end for**
Experimental setting: $n = 100$, $K_{max} = 10$ $nsim = 100$, $score \in$ {ARI, NID}

The result of this experiment stresses out two things, (1) the cluster comparison scores need to be corrected for chance and (2) the variance (or standard deviation) needs to be taken into account when interpreting the stability of a clustering.

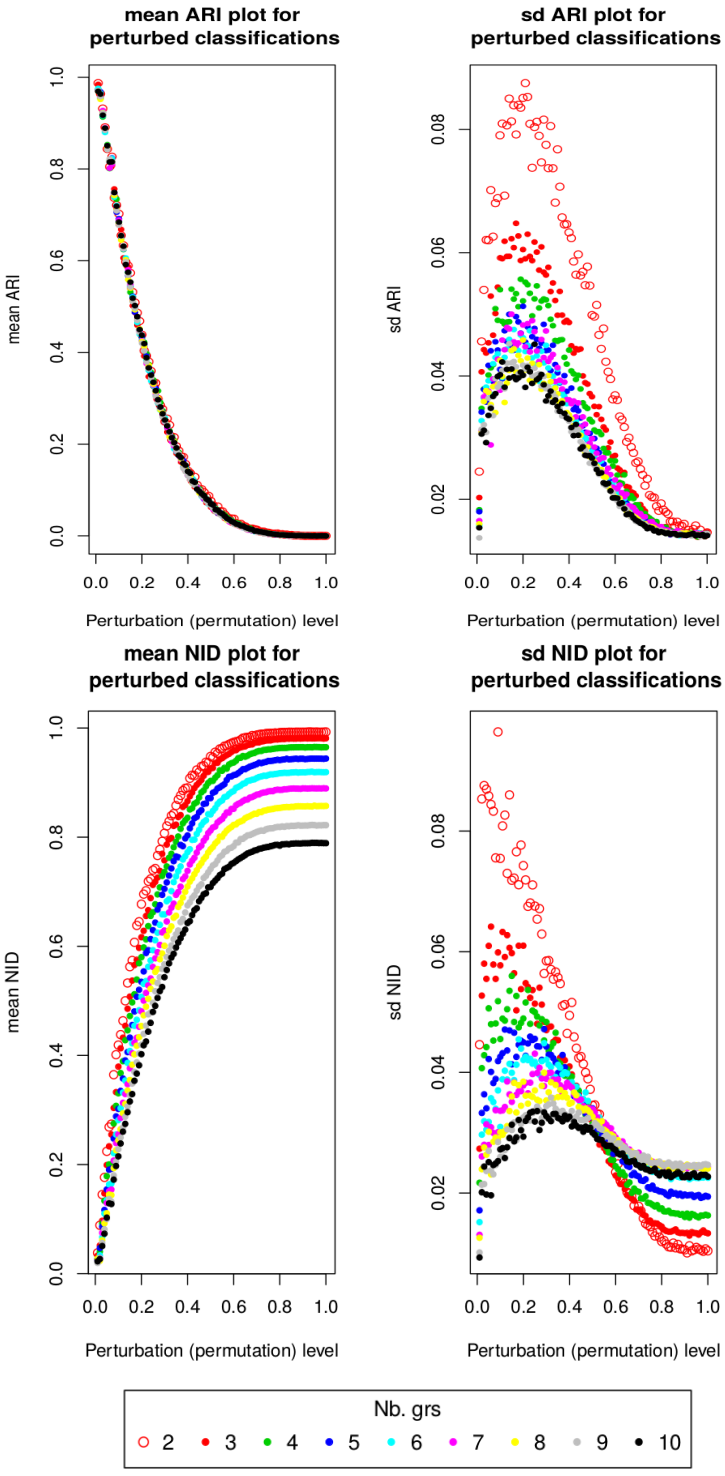


Figure 1.6 – The mean ARI (top left) and NID (bottom left) for different K (color code) and different proportions of permuted labels in one of the clusterings. Top right: standard deviation for ARI, bottom right: standard deviation for NID.

1.6. Problematic and contributions

Classifying TNBC based on proteomic or genomic data is an important topic in oncology. Indeed, identifying subgroups with specific omic signatures is a crucial step for the development of molecular treatments for this pathology. In order to propose a TNBC classification, unsupervised clustering methods are employed. However, as shown in this chapter, clustering is a data exploration method that is not always easy to use and the results for a TNBC cluster analysis will depend on several parameters such as:

- biological: *e.g.* the number and the inclusion criterion of TNBC patients and the selected genes or proteins,
- methodological: *e.g.* the method of preprocessing or data normalization, the choice of clustering algorithm and measured distance between the data observations, the criterion for selecting the number of groups and the validation method

The fact that several TNBC classifications have been proposed in the literature is illustrating how difficult this task is. An explanation that is often forgotten when discussing these results is the impact and the use of different statistical methods. For this reason, I will, in this thesis, try get a better understanding of the TNBC classification by focalizing on the statistical methods used for classifying this pathology. The aim of this thesis is therefore two-fold, with (1) getting a better understanding of the TNBC classifications and (2) doing so by getting a better understanding of the use of cluster stability as a criterion for selecting the numbers of groups in unsupervised clustering. The contributions will therefore be both biological and methodological.

1.6.1. Problématique (français)

Classificatier les tumeurs de TNBC basé sur des données protéomiques ou génomiques est un sujet important en oncologie. En effet, l'identification de sous-groupes de TNBC présentant des signatures omiques spécifiques est une étape cruciale pour le développement de traitements moléculaires de cette pathologie. Afin de proposer une classification TNBC, des méthodes de clustering non supervisées sont utilisées. Cependant, comme le montre ce chapitre, le clustering est une méthode d'exploration des données qui n'est pas toujours

facile à utiliser et les résultats d'une analyse de clustering de TNBC dépendront de plusieurs paramètres tels que

- Biologique : *e.g.* le nombre et le critère d'inclusion des patients atteints de TNBC et les gènes ou protéines sélectionnés,
- Méthodologique : *e.g.* la méthode de prétraitement ou de normalisation des données utilisée, le choix de l'algorithme de clustering et la distance mesurée entre les observations de données, le critère de sélection du nombre de groupes et la méthode de validation

Le fait que plusieurs classifications TNBC ont été proposées dans la littérature illustre la difficulté de cette tâche. Une explication souvent oubliée lors de la discussion de ces résultats est l'impact et l'utilisation de différentes méthodes statistiques. C'est pourquoi, dans cette thèse, je vais essayer de mieux comprendre la classification TNBC en me concentrant sur les méthodes statistique utilisées pour classer cette pathologie. L'objectif de cette thèse est donc double : (1) mieux comprendre l'utilisation de la stabilité de clusters comme critère de sélection du nombre de groupes dans le clustering afin de (2) mieux comprendre les classifications TNBC. Les contributions seront donc à la fois biologiques et méthodologiques.

1.6.2. Biological contributions

I will start and end this thesis by two different TNBC classification studies, where I classify the TNBC tumors in a more statistically rigorous manner than what is normally done. In the first study, I base the classification on proteins and use the RATHER consortium. In the second study, I base the classification on genes (RNAseq) and the publicly available TCGA dataset.

Contribution 1: Proteomic classification of TNBC

Chapter 2 is dedicated to the proteomic classification. In this study I successively investigated the following questions:

- **Clustering:** Is there any groups present in this dataset?
- **Selection of the number of groups:** How many groups are they?
- **Characterization:** Which are these groups?

- **Validation:** Are these groups found in another dataset?

By doing so, I find a stable classification of $K = 2$ groups in the RATHER-NKI dataset, that serves as training set in my study. These two groups are characterized by different proteomic expression and are found as well in the RATHER-CAM dataset, that serves as validation set in this study. However, when I used an external dataset, for which the proteomic data had not been extracted together with the training set, for the validation, these two groups were not found. Also, these two groups did not differ in survival rates in neither the RATHER-NKI nor the RATHER-CAM dataset, and no differential gene expression characterizing the groups was found. Important batch effects were also observed for the proteomic data. Taken together, these results question the existence of the classification in this dataset and show the importance of (1) correctly normalizing data in order to take into account any batch effect and (2) conducting rigorous validation procedures. In this proteomic classification study I had promising results for the first 3 questions but the answer to the 4^{th} questioned their validity and robustness.

Contribution 4: TNBC classification of the TCGA dataset

In (the last) Chapter 5, I classify a large TNBC dataset from the TCGA cohort using the clustering validation methods that I developed during my thesis. I compare the results to those obtained by the TNBCtype tool (Chen et al., 2012) corresponding to the classification of Lehmann et al. (2011). These results show that the Lehmann et al. (2011) classification is important for TNBC subtyping but not sufficient to take into account all the diversity that exists among the TNBC tumors.

Supplementary contribution: TTKi for molecular drug discovery in TNBC

In Appendix A I present a supplementary project. This project was conducted with the aim of finding molecules that could be used in combination with a TTK inhibitor (TTKi) to treat TN patients. It was a collaboration with an industrial, and the data are therefore confidential. In this report I discuss the analysis pipeline, I put in place and the linear model I defined in order to conduct these analyses.

1.6.3. Methodological contributions

As, illustrated by my proteomic classification study of TNBC, it is possible to find a stable TNBC classification in a dataset that is then difficult to validate. To get a better understanding of this, I focalized on the method used for selecting the number of groups in TNBC classifications, that is the measurement of cluster stability. Indeed, selecting the number of groups in unsupervised clustering is difficult and remains an open question in statistics. Basing this choice on the stability of the clusterings makes sense since unstable classifications are difficult to validate in other datasets, especially if there is not a perfect overlap between the studied variables. Also, the idea of subsampling is reasonable in this context since the genes are often highly correlated with each other (Ben-Hur and Guyon, 2003). However, despite the popularity of cluster stability, little is still known about how or under which conditions this method works. For this reason, I propose two important methodological contributions, increasing the usability and interpretability of this method.

Contribution 2: Adjusting the adjusted Rand Index - A multinomial story

In the first of these two methodological contributions I focalize on the ARI Hubert and Arabie (1985) score for cluster comparison. This work is submitted to the journal Computational Statistics (Sundqvist et al., 2020a).
Comparing clusterings is a crucial step for estimating the stability of a clustering. As show in Section 1.5.4, cluster comparison scores depend on the number of groups and needs to be corrected for chance. The adjustment of the ARI is based on a hypergeometric distribution assumption which is unsatisfying from a modeling perspective as (i) it is not appropriate when the two clusterings are dependent (ii) it forces the size of the clusters, and (iii) it ignores randomness of the sampling. Chapter 3 of my thesis is therefore dedicated to propose a corrected version of this score by basing it on a multinomial distribution hypothesis. The multinomial model is advantageous as it does not force the size of the clusters, properly models randomness, and is easily extended to the dependent case. I show that the ARI is biased under the multinomial model and that the difference between the ARI and $MARI$ can be large for small n but essentially vanish for large n, where n is the number of individuals. The $(M)ARI$ scores are then implemented in an efficient algorithm to compute all

these quantities in the `aricode` R-package.

Contribution 3: Cluster stability for class discovery

In the second methodological contribution, I implemented an R package `clustRstab` (Sundqvist et al., 2020b), that easily enables to estimate the stability of a clustering in different parameter settings. With this method I then conducted a simulation and an application study. In the simulation study I tested whether, (1) if we had access to a large number of datasets, the most stable clustering corresponds to the correct number of groups and (2) whether we are able to estimate this stability when we, as in practice, only have access to one dataset. In the application study, I applied this method to the NCI60 cancer dataset (Ross et al., 2000) consisting of cell lines from 9 different types of cancer.The results of these two studies show that (1) cluster stability does not correctly estimate the number of groups, and this, even in simple data settings, and (2) cluster stability does not always allow to identify interesting clusterings. Indeed, the most stable clustering of the NCI60 dataset only separated one of the cancer types from the others. These results question the use of cluster stability for clustering such complex data as TNBC tumors and the importance of combining this method with other clustering validation criteria.

Chapter 2

Proteomic Classification of Triple Negative Breast Cancer

Résumé. Le cancer du sein triple négatif (TN) est une forme agressive de cancer du sein. Actuellement, il n'existe aucun traitement ciblé pour cette pathologie et la recherche de cibles thérapeutiques potentielles reste une priorité en oncologie. La principale difficulté de cette tâche est liée à l'hétérogénéité des tumeurs TN. Pour cette raison, plusieurs classifications TN ont été proposées. Ces classifications ignorent les informations protéomiques des tumeurs. Cependant, les protéines sont des importantes actrices dans les cellules.

L'objectif de la présente étude était donc de classer les cancers TN sur la base de données protéomiques en utilisant des méthodes de classification non supervisées. Pour ce faire, nous avons étudié une approche de la stabilité des clusters (Von Luxburg et al., 2010) afin d'identifier des groupes de tumeurs TN qui étaient robustes, *i.e.*, insensibles aux variations des variables. Dans cette veine, nous avons développé et mis en œuvre un algorithme de stabilité des clusters. Puis, cette méthode a été combinée avec une méthiode de classification 'model based' et appliquées à un jeu de données d'entraînement contenant $n = 67$ patients et $p = 116$ protéines. Deux groupes de tumeurs TN robustes étaient identifiés dans ce jeu de données, montrant des patterns d'expression de protéines différents. Les deux mêmes groupes ont ensuite été identifiés dans un jeu de données de validation ($p = 116$, $n = 31$) partageant les mêmes protéines, mais pas dans un second jeu de données de validation ($n = 42$) partageant uniquement $p = 70$ protéines avec le jeu de données d'entraînement. Aucune signature transcriptomique permettant de prédire les groupes TN à partir du modèle d'entraînement n'a été trouvée. De plus, de nombreux biais expérimentaux ont été observés pour les données RPPA (protéomiques), ce qui rend l'interprétation des résultats difficile. Cette étude souligne l'importance de valider une classification sur des ensembles de données externes et la nécessité de procédures de normalisation plus robustes pour les données RPPA.

Abstract. Triple negative (TN) breast cancer is an aggressive form of breast cancer. Currently, no targeted treatment exists for this pathology and the research of potential therapeutic targets remains a priority in oncology. The main difficulty in this task relates to the heterogeneity among TN breast cancer tumors. The aim of the present study was therefore to classify TN cancer based on proteomic data by using unsupervised classification methods.To do so, we investigated an approach of cluster stability (Von Luxburg et al., 2010) in order to identify TN tumor groups that were robust, *i.e.*, insensitive to variable variation. In this vein, we developed and implemented an algorithm for cluster stability. This method was combined with model-based unsupervised classification methods setting and applied to a training set of $n = 67$ samples and $p = 116$ proteins. Two robust TN tumor groups were identified in the training set showing different protein expression pattern. The same two groups were later identified in a validation set ($p = 116$, $n = 31$) sharing the same proteins, but not in a second validation set ($n = 42$) sharing only $p = 70$ proteins. No transcriptomic signature that could predict the TN groups from the training model was found. Also, many experimental biases were observed for the RPPA (proteomic) data making the interpretation of the results difficult. This study highlight the importance of validating a classification on external datasets and the need for more robust normalization procedures for the RPPA data.

Acknowledgements. This project was conducted under the supervision of the supervisors of my thesis as well as Leanne De Koning (head of the RPPA platform, Institut Curie). I want to thank her for all the help for understanding the RPPA data.

2.1. Introduction

2.1.1. Triple Negative Breast Cancer

Triple negative (TN) breast cancer is an aggressive form of cancer with poor diagnosis. It is characterized by a low/lack of expression of estrogen and progesterone receptors and by a lack of human epidermal growth factor receptor 2 (HER2) over-expression (Foulkes et al., 2010). Therefore, this subtype cannot be treated with endocrine therapy or therapies targeting HER2 (Foulkes et al., 2010) as other subtypes of breast cancers. Currently, no targeted treatment exists for this pathology and the research of potential therapeutic targets remains a priority in oncology. The main difficulty in this task relates to the heterogeneity among TN breast cancer tumors. For this reason, several classifications based on transcriptomic data have been proposed for TN breast cancer (see for example Bianchini et al. (2016); Burstein et al. (2014); Lehmann et al. (2011, 2016)). These classifications do not take into account the analysis of tumors at the protein level. Yet, proteins play a major role in cellular functions by regulating signaling pathways. Hence, a proteomic based classification could give important clues for therapeutic development. A promising study on this topic was conducted by Masuda et al. (2017). Based on proteins, these authors classified TN tumors into two stable groups. These two groups could be split into five smaller groups which showed some similarities to the classification proposed by Lehmann et al. (2011). However, they did not validate their classification on an external dataset. The aim of the present study was therefore to (1) classify TN cancer by using proteomic data and unsupervised classification methods and (2) characterize the resulting groups at a proteomic and genomic level in order to identify potential therapeutic targets. To do so, we had at our disposal two large TN datasets that contained proteomic data, measured by the RPPA technique (for information about this technique, see Section 1.3.1).

2.1.2. Unsupervised classification of Triple Negative Breast Cancer

The most difficult tasks when conducting unsupervised classification is to select the number of groups. Indeed, this number is unknown and has to be estimated. Several methods have been developed for this aim. In oncology, it

is popular to base this choice on the stability of clustering. The idea is that, an identified clustering should be robust to changes such as variable or observation subsampling. The number of groups can then be selected as the one giving the most stable clustering. Even though this idea is a heuristic, studying the stability of a classification is important since unstable classifications are difficult to validate in other datasets, especially if there is not a perfect overlap between the studied variables.

In the study of Masuda et al. (2017), as well as in the transcriptomic based TN classification studies, the choice of the number of TN-groups was based on the *Consensus Clustering* method (Monti et al., 2003). This method allows to propose and validate a classification by measuring the stability of different number of groups and by removing unstable observations from the classification. Yet, in the simulation study conducted by Şenbabaoğlu et al. (2014), the authors found that this method is not always reliable since (1) it is able to divide randomly generated unimodal data into stable clusters for a range of different numbers of groups and (2) for data with known structure, it poorly identified the true number of groups. Other, more technical constraints of this method are going to be detailed in Section 2.2.2.

For all these reasons, it is not clear whether the correct number of groups was recovered and whether the clustering is indeed stable. To re-assess this, we implemented our own algorithm estimating the stability of a clustering which allows to correct the estimated stability for chance. This algorithm was inspired by the work of Von Luxburg et al. (2010). In order to make the selection of the number of groups more robust we combined the cluster stability criteria by an information criterion. To do so, we conducted unsupervised classification on our TN training cohort in both a non-probabilistic setting (as done in the other TN classification studies), by using the hierarchical ascendant clustering algorithm (HAC), and in a probabilistic setting, by using the Gaussian Mixture Model (GMM) algorithm. This latter allowed us to validate the identified classification by computing the posterior probability of the training model in two external validation sets. This validation procedure is statistically more stringent than what is normally done in TN classification studies. Indeed, in these latter, the validation procedure usually consists in classifying the validation set using the method used for the training set. The classification is then validated by descriptively comparing the obtained results.

Another issue when working with proteomic RPPA data is that the pro-

teins, included in a study, need to be selected *a priori* and, therefore, differ between different cohorts. Making the comparison and validation of different classifications difficult. Also, there are still few TN cohorts that contain RPPA data. In order to remedy to this our idea is to find a transcriptiomic signature that can predict the identified proteomic groups from RNA data. This would allow to study the proteomic based classification in larger, publicly available, datasets for which no RPPA data is available. To implement this, we conducted two types of supervised classification methods on our TN training set; linear discriminant analysis (LDA) and penalized logistic regression.

The originality of this work was therefore:

1. The use of proteomic and not transcriptomic data for the TN classification.
2. Implementing and using a normalized clustering stability index in order to select the number of TN-groups.
3. Conducting unsupervised classification in both a non probabilistic and a probabilistic setting, allowing us to combine the stability index with an information criterion in order to select the number of groups and to validate the classification by computing the posterior probability of the training model.
4. Using supervised methods to search for a transcriptomic signature that could predict the identified TN-groups.

2.2. Methods

2.2.1. Data

Training and validation sets. Two large data sets of TN breast cancer were at our disposal.

- First, the Rational Therapy for Breast Cancer consortium (RATHER, described in Michaut et al., 2016, see `www.ratherproject.com`), containing proteomic ($p = 116$) and transcriptomic data from $n = 99$ TN breast cancer patients collected from the Netherlands Cancer Institute (NKI), Amsterdam ($n = 69$, referred to as the RATHER-NKI

dataset), and from Addenbrooke's Hospital, Cambridge, UK ($n = 32$, referred to as RATHER-Cambridge dataset). Two samples with more than 80% of missing proteomic data were excluded from the RATHER-NKI dataset resulting in $n = 67$ TN samples. One sample from the RATHER-Cambridge dataset was later excluded due to its strange proteomic expression resulting in $n = 31$ TN samples.

- Second, the Curie data set, containing proteomic ($p = 249$) and transcriptomic data from $n = 43$ TN breast cancer patients collected at Institut Curie. The Curie data set served as a validation data set and had $p = 70$ proteins in common with RATHER. One sample with missing values was excluded resulting in $n = 42$ included TN samples.

The collected samples in the RATHER cohort contain at least 30% of cancer cells, whereas the Curie samples contain at least 50% of cancer cells. Both the RATHER and the Curie data sets contain survival data with an average of 10 years of surveillance. The RATHER-Cambridge data set was later defined as a validation set, for more information see Section 2.3.1.

TN inclusion criteria. As described in Section 1.4.1, the inclusion criterion of TN breast cancer tumors for the RATHER consortium was based on two criteria: (1) TN on diagnosis and (2) ER, PR and HER2 negative on TargetPrint (RNA analysis). At first we added a supplementary criterion: ER, PR and HER2 negative on tissue microarray measure of immunohistochemistry (IHC). This resulted in the exclusion of three samples from the RATHER-NKI cohort. Given that there was more than 20% of missing value for this criterion it needs to be used with caution. We therefore checked that the inclusion and exclusion of these three patients did not drastically change the output and decided to not take into account this criterion. A schematic representation of the inclusion criteria and pre-processing for the RATHER Consortium is illustrated in Figure 2.1 (taking into account the three initial inclusion criteria).

In the Curie data base, the criterion for TN breast cancer was based on ER, PR and HER2 negative on tissue microarray measure of IHC ($< 10\%$). Aberrant and low signal proteins were excluded ($p = 15$).

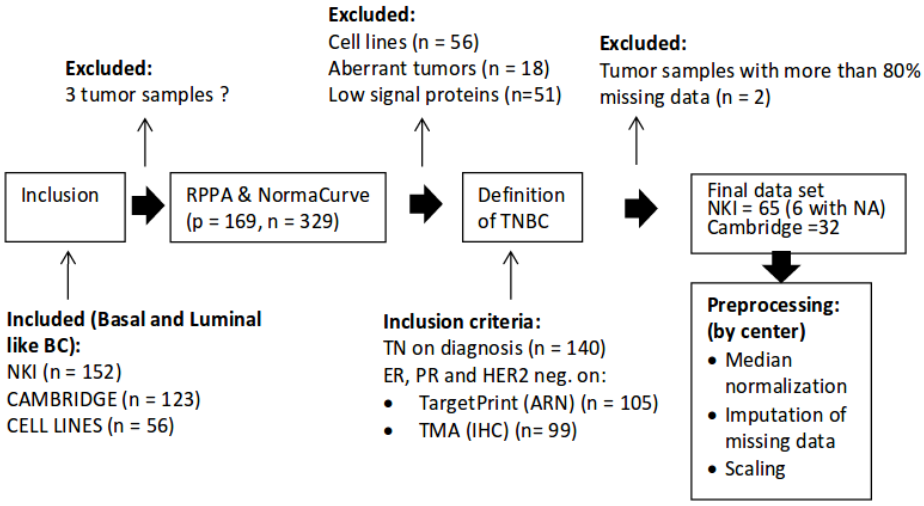


Figure 2.1 – Scheme of inclusion and pre-processing for RATHER cohort protein data.

Data analyses and pre-processing

RPPA analysis was performed at Institut Curie for both data sets. The RPPA technique is still quite new and no standardized normalization procedures exist yet. At the Institut Curie, a method called NormaCurve, proposed by Troncale et al. (2012), was used for this aim. The NormaCurve correction corrects for background noise (using a slide labeled without any primary antibody) and the total amount of printed protein (using a slide labeled with a total protein stain). Next, Normacurve draws an antibody response curve based on all samples present on the array, and then aligns each sample on this curve using the serial dilutions and replicates of that sample, in order to determine one adjusted normalized value per sample. In order to correct for sample intensity, each protein was corrected by its median expression. Results of this median correction are shown in Section 2.3.1. Missing values were imputed by the R-package `impute`, which compute a row and columns distance to predict missing values (Hastie et al., 2016). The data were then centered and scaled.

2.2.2. Unsupervised classification

Clustering methods

In order to classify the TNBC tumors, we used two clustering methods. First, a distance based unsupervised clustering method, the Hierarchic Ascendant

Clustering (HAC) with Euclidean distance and the Ward method (Murtagh and Legendre, 2011). Second, the probabilistic Gaussian Mixture Model (GMM) was fitted (Dempster et al., 1977), which allows us to estimate the group belonging of the observations. The reader is referred to the Section 1.5 in the previous chapter for more information about these methods. To implement the HAC algorithm, we used the `hclust` function from the R-package `stats` (R Core Team, 2016). To implement the GMM and the EM algorithm, we used the `Mclust` function from the R-package `mclust` (Scrucca et al., 2016b).

These unsupervised methods were used first for classifying the observations from the training set, and later for classifying the proteins that were differently expressed among the identified TN groups.

Model selection

In order to estimate the number of TN-groups present in the training dataset we used two model selection criteria: an information criterion and cluster stability.

Cluster stability. As in the previously cited TN breast cancer classification studies, we used a selection criterion based on the stability of the clustering. However, in difference with the other TN breast cancer classification studies, we did not use the *Consensus Clustering* (Monti et al., 2003) but implemented our own cluster stability algorithm based on the work of Von Luxburg et al. (2010). We did this for three reasons. First, the stability of a clustering depends on the number of groups and hence, needs, to be adjusted for this in order to be correctly interpreted. This is not possible by the *Consensus Clustering*. Second, *Consensus Clustering* is not estimating the moments (mean or variance) of the stability of a clustering, but is computing the distribution of the 'consensus matrix'. This makes the interpretation difficult and prone to subjectivity. Thirdly, and most important, the *Consensus Clustering* merges the procedures for proposing and validating a clustering. Indeed, this method proposes a clustering that is tailored to be stable by removing the observations that are unstable or "difficult" to cluster. By doing so, the stability of the clustering is artificially enhanced. Thus, validating it by its stability do no longer make sense. Separating the clustering and validation procedures also allows to use supplementary clustering validation criteria. The clustering stability algorithm that I implemented is presented in Section 2.2.2. This algorithm

will be detailed in Chapter 4 of the thesis.

Bayesian Information Criterion. The advantage of using a probabilistic setting for the clustering, as with the GMM, is that an information criterion can be used for the model selection. In this study we are going to use the Bayesian Information Criterion (BIC) as defined in Equation 1.5 in the previous chapter. This criterion is balancing the maximization of the likelihood of the model and the number of estimated parameters (for example the number of groups). To compute the BIC, we used the `Mclust` function implemented in the `mclust` R-package.

Normalized Clustering Stability. The stability of a clustering can be estimated in several manners, and this is going to be discussed in Chapter 4.2. Most of the time, it is estimated by comparing clusterings that are obtained from perturbed versions of the same dataset. In order to obtain perturbed datasets, we are going to subsample the variables (proteins) by different proportions *prop*. To compare the clusterings we are going to use the Normalized Information Distance[1](NID; Vinh et al. (2010)). Since the NID is a distance-based score, it can be viewed as an instability index. To do so, we will estimate the instability index $\widehat{Instab}_{K_{prop}}$ of a clustering, where K is the number of groups.

An issue in clustering comparison is to set a proper baseline, that is, the expected value of shared information between two clusterings, independently sampled at random. Indeed, this amount of shared information increases with K (Morey and Agresti, 1984; Vinh et al., 2010; Von Luxburg et al., 2010). In order to correct for this, we implemented a permutation procedure which allows to correct each cluster comparison by a "baseline -permuted- score". That is, the score obtained between two clustering when the labels of one of them have been permuted. This correction procedure was implemented in an algorithm similar to the one presented in Von Luxburg et al. (2010) with the following steps:

> For a given dataset $\mathbb{X}$, of n observations $x_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, and p variables (proteins) do:

[1]The NID is a real distance and a metric, based on the entropy of the two clusterings. It measures to which extent, knowing one of the clusterings, is helpful for predicting the other. It is computed from the contingency table of the two clusterings, and its mathematical definition can be found in Definition eq:EntropiForNID in Section 1.5.4

- **for** $prop \in \{0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$

  1. Generated a large number, D, of perturbed datasets by subsampling a proportion, $prop$, of the proteins such that,

$$subsampling_{prop} : \mathbb{X} \rightarrow \mathbb{X}'_{prop_1}, \ldots, \mathbb{X}'_{prop_D}.$$

  2. **for** $K \in \{1 \ldots, 10\}$ **do:**

     (i) Cluster each subsampled dataset, $\mathbb{X}'_{prop_d}$, for $d \in \{1, \ldots, D\}$ into K groups using either the HAC-ward or the GMM clustering algorithm, C_K, such that

$$C_K : \mathbb{X}'_{prop_d} \rightarrow 1, \ldots, K.$$

     (ii) Compute pairwise comparisons using the NID score between all $\binom{D}{2}$ possible pairs of clusterings, such that,

$$NID : C_K(\mathbb{X}'_{prop_d}), C_K(\mathbb{X}'_{prop_{d'}}) \rightarrow [0,1],$$

     where $d < d'$ and
     - $NID(C_K(\mathbb{X}'_{prop_d}), C_K(\mathbb{X}'_{prop_{d'}})) = 0$ indicate that the two clusterings, $C_K(\mathbb{X}'_{prop_d})$ and $C_K(\mathbb{X}'_{prop_{d'}})$ are identical
     - $NID(C_K(\mathbb{X}'_{prop_d}), C_K(\mathbb{X}'_{prop_{d'}})) = 1$ indicate that they are independent.

     (iii) Permute the labels in $C_K(\mathbb{X}'_{prop_d})$ to obtain $C_{K perm.}(\mathbb{X}'_{prop_d})$, compute pairwise comparisons between $C_{K_{perm.}}(\mathbb{X}'_{prop_d})$ and $C_K(\mathbb{X}'_{prop_{d'}})$, repeat 10 times and compute the mean to obtain

$$NID_{perm.}(C_{K_{perm.}}(\mathbb{X}'_{prop_d}), C_K(\mathbb{X}'_{prop_{d'}}))$$

     (iv) Compute the normalized instability index $\widehat{Instab}_{K_{prop}}$ as the arithmetic mean of the cluster comparison, scaled by their permutated score:

$$\widehat{Instab}_{K_{prop}} = \frac{1}{\binom{D}{2}} \sum_{d<d'} \frac{NID(C_K(\mathbb{X}'_{prop_d}), C_K(\mathbb{X}'_{prop_{d'}}))}{NID_{perm.}(C_{K_{perm.}}(\mathbb{X}'_{prop_d}), C_K(\mathbb{X}'_{prop_{d'}}))}$$

3. end **for**

4. Estimate the true number of groups $K^\star$ as the one minimizing the instability index:

$$\widehat{K} = \underset{K}{Argmin}\ \widehat{Instab}_{K_{prop}}$$

- end **for**

2.2.3. Differential analysis

Differential analysis was performed in order to characterize the identified TN groups at proteomic and clinical level. To do this we used analysis of variance (ANOVA) fitting the following linear model:

$$y_{ki} = \alpha_k + \epsilon_{ki}, \quad \epsilon_{ki} \sim \mathcal{N}(0,\sigma^2), \tag{2.1}$$

where k_i is the i-th individual in the k-th identified TN groups with $i \in 1, ..., n$ and $k = 1, \ldots, \widehat{K}$. The alternative hypothesis, $\mathcal{H}_1$, stating that at least one pair of groups exists such that $\alpha_k \neq \alpha_{k'}$ was tested against the null, $\mathcal{H}_0$, for all k, k' $\alpha_k = \alpha_{k'}$. False discovery rate corrections (Benjamini and Hochberg, 1995) were applied to correct for multiple comparisons. T-test comparisons were conducted for variables with a p-value$<$ 0.05 associated to the ANOVA analysis.

To conduct the survival analysis, the survival function was estimated by the *Kaplan-Meier estimation* (Kaplan and Meier, 1958) and group differences were tested by the *log-rank* test (Harrington and Fleming, 1982). These analyses were implemented by using the R-package `survival` (Therneau, 2015).

2.2.4. Validation

In order to validate the obtained clustering, we predicted the groups of the TN tumor samples, in the validation set, by computing the posterior probability of the GMM training model. Being able to compute this posterior probability is another advantage of using a probabilistic setting for the unsupervised clustering of the training set. The probability for a given observation x_i, with

$i = 1, \ldots, n$, to belong to group k, with $k \in \{1, \ldots, \widehat{K}\}$, is predicted by

$$p(k|x_i) = \frac{\hat{\pi}_k \exp\left[\frac{1}{2\hat{\sigma}_k^2} \sum_{i=1}^{n}(x_i - \hat{\mu}_k)^2\right]}{\sum_{k=1}^{K} \hat{\pi}_k \exp\left[\frac{1}{2\hat{\sigma}_k^2} \sum_{i=1}^{n}(x_i - \hat{\mu}_k)^2\right]}, \ \forall \ k \ \in 1, \ldots, \widehat{K} \tag{2.2}$$

where $\hat{\sigma}^2$, $\hat{\mu}$ and $\hat{\pi}$ are the estimated variance, mean and weight parameters for the training GMM model. The cluster label of a given observation x_i, is predicted as the one maximizing this probability.

2.2.5. Searching for a transcriptomic signature

In order to search for a transcriptomic signature predicting these groups, the following strategies were used:

- Linear Discriminant Analysis (LDA) was applied to search for linear combinations of genes that best separates the identified TN groups in the training set. To do so, we used the `lda` function from the R-package `MASS`.

- Penalized logistic regression was used to predict the proteomic groups of the training set from the genes. To do so, we used the R-package `glmnet` (Simon et al., 2011).

  The reader can find more information of these methods in chapter 4 of Friedman et al. (2001).

2.3. Results

2.3.1. Classification

Selection of training set

Our first intention was to use the RATHER dataset, containing both samples from the RAHTER-NKI and from the RAHTER-Cambridge cohort, as the training set. The model selection procedure for this (global) dataset, when using the GMM and the BIC, selected the spherical, varying volume

(VII) model with three clusters consisting of respectively $n_1 = 38$, $n_2 = 36$ and $n_3 = 22$ observations. At first sight, these clusters seemed to be well balanced on both NKI and Cambridge tumors. Indeed, the two principal component analysis (PCA) plots as well as the heatmap in Figure 2.2 show no obvious effect of the center. However, the Adjuster Rand Index (ARI, Hubert and Arabie (1985)) for NKI tumors, when the center-based classification of these tumors was compared to the global classification was only $ARI(\text{NKI}_{\text{global}}, NKI_{\text{center-based}}) = 0.42$ whereas the Cambridge tumors $ARI(\text{Cambridge}_{\text{global}}, \text{Cambridge}_{\text{center-based}}) = 0.03$. Hence, there was almost no overlap between the center-based classification of Cambridge tumors and the global classification. It therefore seems that, when the tumors from the two centers were treated together, an artifact was generated. We therefore decided to separate the tumor samples from both centers and consider the RATHER-NKI samples as the training and the RAHTER-Cambridge samples as the validation set. The procedures of preprocessing were then applied to each center separately.

Median correction

We observed that the expression of the different proteins was highly correlated with the median expression of the samples. That is, the expression of a given protein depended on the quantity of the printed (spotted) biological material by the RPPA machine. This indicates that the NormaCurve preprocessing did not correct sufficiently for the quantity of printed material. An example of the relation between the expression of a given protein and the sample median is illustrated in Figure 2.3.a), and the histogram of all the 118 resulting median regression coefficients is presented in Figure 2.3.b). Since the regression coefficients are centered around one, our first intention was to correct each protein by dividing its expression level by the median of each sample. In this manner, the corrected proteomic value of a given protein p' and observation i was given as $x^c_{ip'} = \frac{x_{ip'}}{median(x_i)}$, where c indicates that the proteomic value is corrected and $p' = 1, \ldots, p$. When doing so, the correlation coefficients between the different proteins were largely decreased as can be seen in Figure 2.3c) and d). However, we also observed a strong technical batch and spotting effect of the experiment (all the proteins were not analyzed during the same day) and proteins analyzed within the same experiment are strongly correlated to each other. There

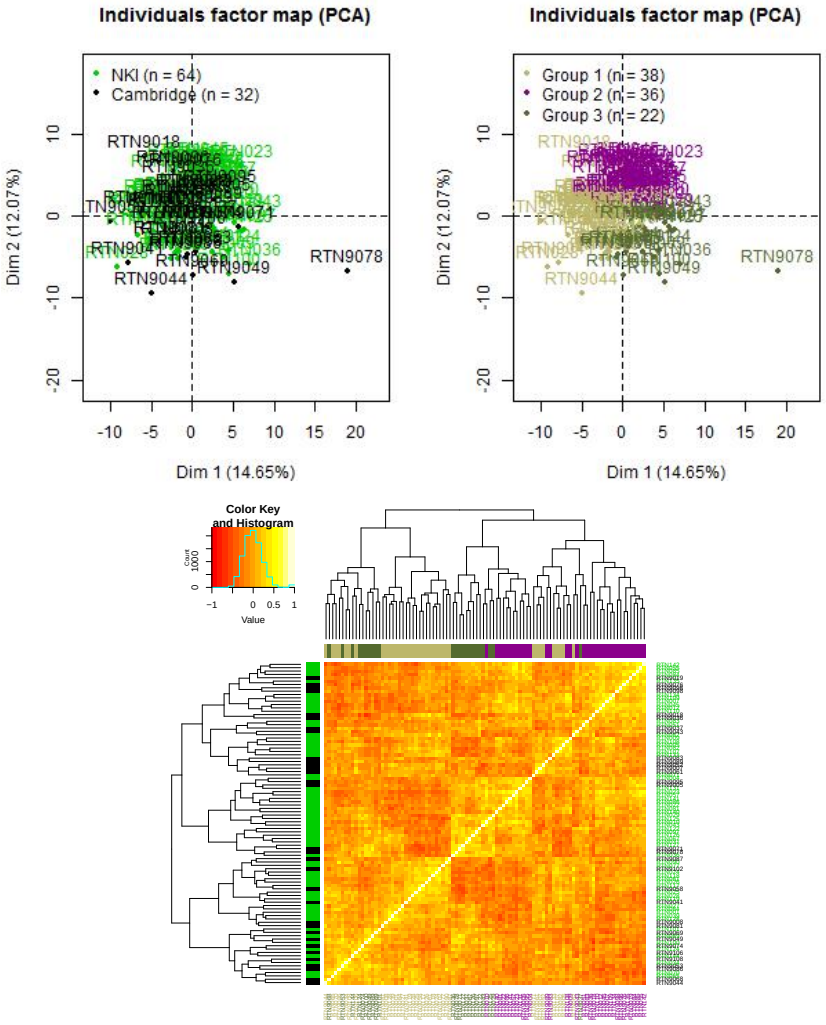


Figure 2.2 – Top: Individual plots of first two axes in PCA of RATHER RPPA data. Left panel: color code indicates the source of the tumor, green = NKI; black = Cambridge. Right panel: the color code indicates the three different groups found by the global GMM classification. Bottom: Heatmap for tumor samples ordered after HAC. Heatmap colors indicate strength of Pearson correlation coefficient. Row and column color codes are identical to the color codes of the left respectively right PCA plots.

were three batches that were each containing three spottings (experiments). The proteins of the same batch and that were spotted during the same day were highly correlated with each other. This can be seen by the "squared" pattern in Figure 2.4a), showing the Pearson correlation coefficient matrix for the RATHER-NKI proteins ordered after spotting and batch (before median correction), where each "square" correspond to a spotting or batch. In order to correct for this experimental effect, we conducted, for each spotting separately, a linear model for each protein and the sample median with:

$$\forall p' \in s,\ x_{ip'} = a + b * median_s(x_i) + e_{ip'},$$

where a is the intercept, b the regression coefficient and $e_{ip'}$ is the residual of the linear model and s indicate the spotting to which the protein belongs, with $s = 1, \ldots, 9$. The corrected value of $x_{ip'}$ is defined as the residual $x^c_{ip'} := \widehat{e}_{ip'}$. Figure 2.4b) shows the Pearson correlation coefficient matrix for $x^c_{ip'}$, the corrected proteomic values. As it can be seen, the "squared" structures now are gone, indicating that the batch and spotting effects are (at least to some extent) corrected. Two supplementary proteins were excluded since they were only two proteins in a spotting and could therefore not be used for median regression (initially, there was $p = 118$ proteins in the RATHER dataset).

Once we had corrected the RATHER dataset for these batch/spotting effects, the GMM and BIC clustering procedure did no longer find $\widehat{K} = 3$ but $\widehat{K} = 2$ groups. The 3 TN groups identified in the previous section might therefore have been a result of the three experimental batches. This highlights the importance of robust preprocessing procedures correcting for different experimental effects. Yet, since the RPPA technique is still quite "young", other effects than the batch/spotting effect, might still be unknown and go unnoticed.

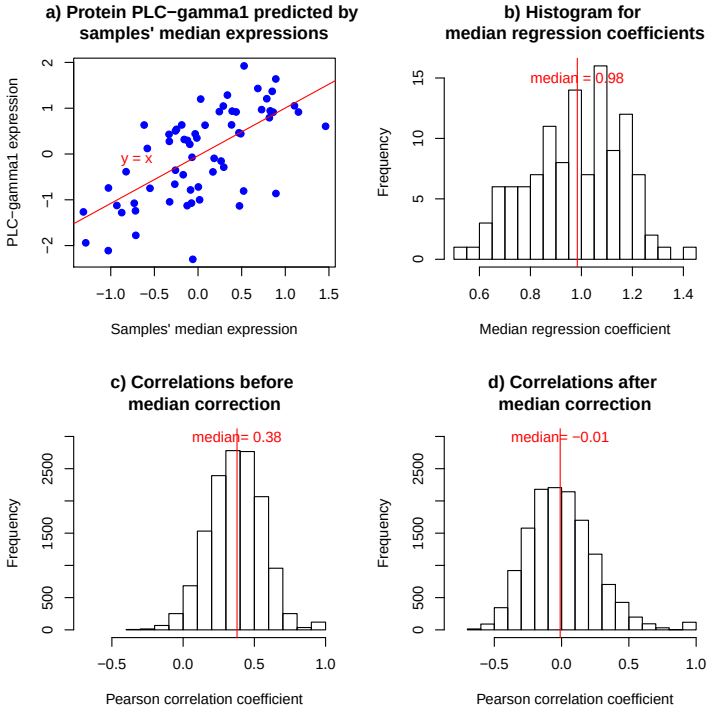


Figure 2.3 – Figure illustrating median correction of RATHER NKI RPPA data. a) Example of a scatter plot illustrating the correlation between the expression of a given protein and the samples' median expressions. The red line represents the median regression slope, $x_{ip'} = a + b * median(x_i)$, where $x_{ip'}$ is the protein expression and $median(x_i)$ is the sample median expression and $b = 1$, $a = 0$. b) Histogram of the median regression slopes (b), the red line indicates the median of the regression coefficients. c) Histogram of Pearson correlation coefficients for proteins before, to the left and d) after, to the right, applying a median correction, that is $x^c_{ip'} = \frac{x_{ip'}}{median(x_i)}$. (*Obs.* both the upper and the lower part of the correlation matrix as well as the diagonal are included)

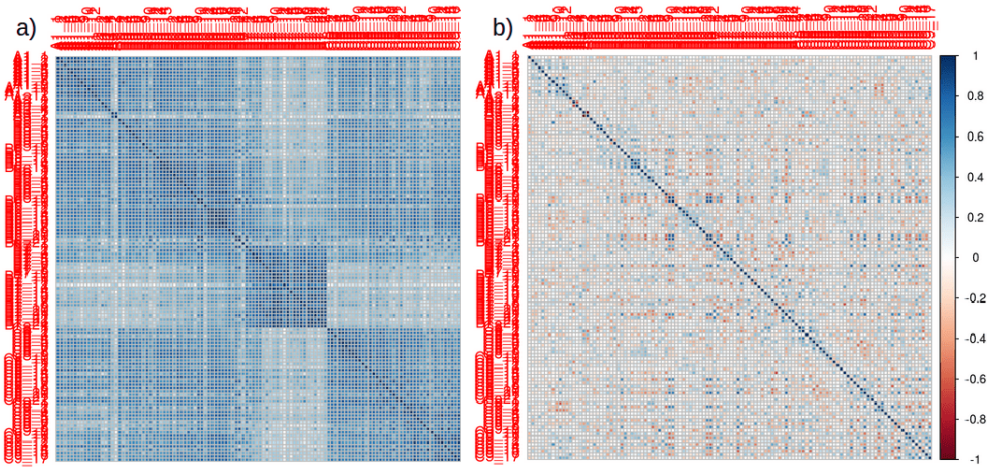
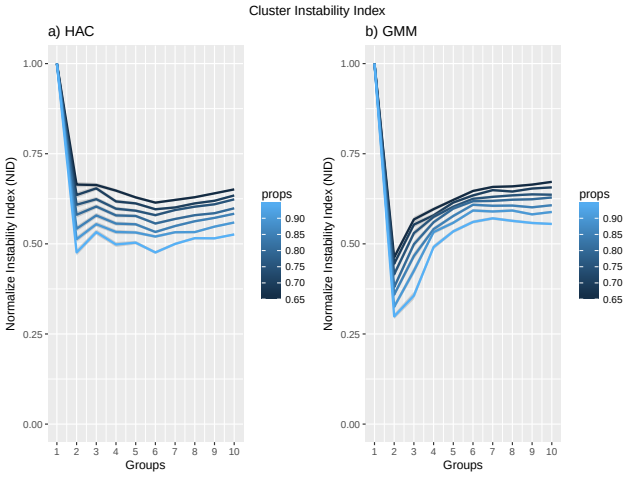


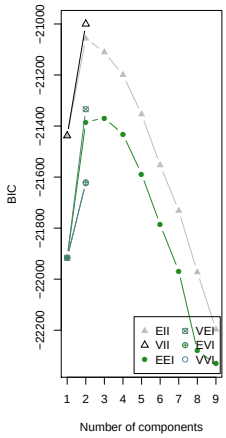
Figure 2.4 – Pearson correlation coefficients matrices for RATHER-NKI proteins ordered after spotting and batch. a) Correlation coefficients before the correction of (intra-spotting) median regression. b) Correlation coefficients after this correction. Color code indicate the strength and the sign of the correlation coefficients.

Model selection for RATHER-NKI training set

Determination of the number $\widehat{K}$ of TN groups. In order to estimate the number of groups present in the NKI dataset we computed the stability of the HAC-Ward and GMM clusterings as well as the BIC for the GMM models for $K = 1, \ldots, 10$ groups. The results are to be found in Figure 2.5. The most stable clustering was yield for $K = 2$ for both tested clustering methods. Indeed, both HAC-Ward and GMM minimized the instability index ($\widehat{Inst}_{K_{prop}}$, *i.e.* average NID) at $K = 2$, and this for all the different proportions, *prop*, of subsampled variables. Also, $K = 2$ maximized the BIC values for almost all tested covariance models. These two results both converge towards estimating the number of groups present in the RATHER-NKI training set to $\widehat{K} = 2$.



(a) Cluster Stability



(b) BIC values GMM

Figure 2.5 – **(a)** Right panel: cluster stability for a) HAC and b) GMM, $K = \{1, 2,..., 10\}$ and proportion of kept proteins $prop = \{0.65, 0.70, 0.75,...,0.95\}$. Results based on 500 simulations. When $K = 1$, the istability index is penalized with 1 and $\widehat{Instab}_{1_{prop}} = \widehat{Instab}_{1_{prop}} + 1$. **(b)** Right panel: BIC values for different GMM covariance models with $K = \{1,...,10\}$.

Selection of the clustering As can be noted in Figure 2.5a, the clusterings obtained by the GMM models are more stable than the clusterings obtained by the HAC-Ward algorithm. We therefore decided to continue with the GMM clusterings. Also, as can be seen in Figure 2.5b, the most convincing GMM model is the spherical, equal volume (EII) model, indeed its BIC values are high and it converges for all estimated K. The two identified groups consisted of $n_1 = 33$ and $n_2 = 34$ TN samples respectively.

2.3.2. Group characterization

In order to understand the underlying biological differences between these groups, we need to characterize them. This can be done at several levels; proteomic, transcriptomic, clinical and by studying the signaling pathways. The characterization of the TN groups at the protein level is one of the key elements in this study. Indeed, this would give important information concerning the molecular profiles of the different TN groups. Differential protein analysis is thus conducted in order to find the dominant biological features of each

cluster.

Differential protein expression

First, Figure 2.6 shows the two groups projected to the two first axis of the RATHER-NKI PCA space. As it can be seen, the two TN groups are well represented by the these two axis, representing 27% of the total variance. This is promising for the differential protein analysis, indicating that these two groups represent a third of the global variation of all proteins.

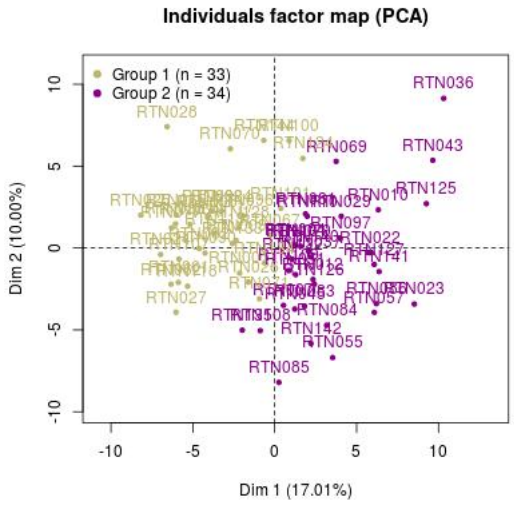


Figure 2.6 – NKI clustering projected to the two first axis of the Principal Components Analysis (PCA) space. Color code indicates the TN sample group belonging.

Second, ANOVA was conducted for each protein by fitting a linear model as described in Section 2.2.3. Since only two TN sample groups were found, the ANOVA analysis came down to conducting t-test between the two groups. Results show that 38 of the 116 proteins are differentially expressed among the two groups ($p.value < 0.05$, FDR corrected). 15 of these proteins were phosphorylated, that is 44% phosphorylated proteins in comparison to 34% phosphorylated proteins in the complete dataset. Yet, this difference was not statistically significant, $\chi^2(1) = 0.21$, $p.value = 0.65$. Moreover, considering only the 7 proteins with the smallest $p.values$ was enough to recover the initial clustering, *i.e.* the clustering based on all the $p = 116$ proteins. These 7

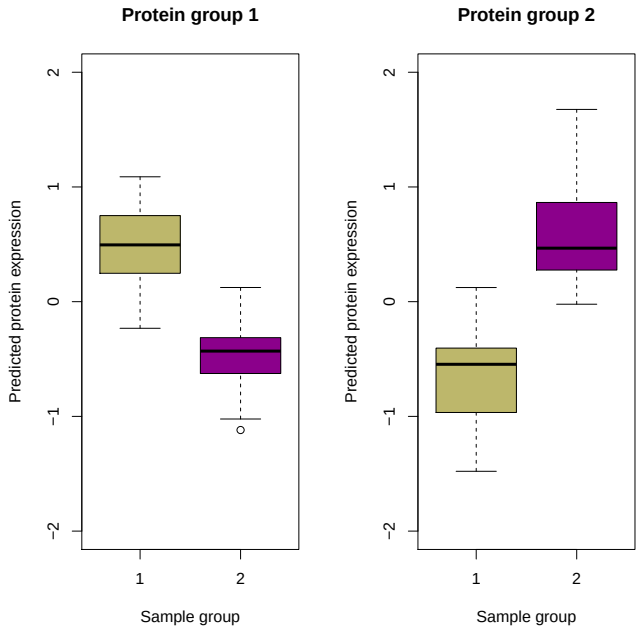


Figure 2.7 – Boxplot with the median, first and third quantile values for the two RATHER-NKI TN groups and the two groups of proteins.

proteins were: Met, IRS1, Rsk1-MAPKAPK1a-p90, CaMKI, Phospho-eEF2k-Ser366, CrkL and Ack1.

Protein classification

In order to get a better vision of how these 38 significant proteins characterize the two TN tumor groups, they were, as the samples, clustered using GMM. The highest BIC values were obtained for $K = 2$ protein groups and the EII model was retained. The two identified protein clusters regroup 19 proteins each and were statistically independent from the spottings $\chi^2(5) = 1.45$, $p.value = 0.92$. The median expression of each protein group and the two TN sample groups are shown in Figure 2.7. As it can be seen, the two TN tumor groups show opposite expression patterns for the two protein groups. Indeed, the first TN group show higher protein expression for the first than for the second proteomic group. Whereas the opposite is observed for the other TN group.

Clinical and survival analyses

Characterizing the clinical profiles of the TN groups is important since they might be related to biological processes such as the age of the patient or their survival rates. Average age, tumor size and number of positive lymph nodes for the two groups are shown in Table 2.1. Group comparison using t.tests showed no statistical difference between these two groups for any of the mentioned variables.

	All		G1		G2		T.test	
	mean	sd	mean	sd	mean	sd	stat.	p.val
Age at diagnosis	51.42	13.83	51.10	12.05	51.73	15.50	-0.18	0.86
Tumor size (cm)	2.78	1.40	2.72	1.17	2.84	1.62	-0.35	0.73
Positive Lymph Nodes	1.14	2.28	1.10	2.28	1.18	2.32	-0.14	0.89

Table 2.1 – Clinical and demographic information for RATHER-NKI samples. G1 and G2 indicate the two discovered TN sample groups

The Kaplan-Meier estimation of the survival function for the two TN groups is shown in Figure 2.8. For both TN groups, events occurred within the first five years and the proportion of censured events was high, 78%. There was no difference in survival rate between the two groups $log-rank = 0$, $p.value = 0.97$. It should be noted that almost all patients received different mixtures of treatments (surgery, chemotherapy, hormonal or radiotherapy). This would be needed to be taken into account in order to correctly estimate the survival functions, yet, it would probably not have changed the group comparison results. This stresses out the importance to find a transcriptomic signature that could predict the two TN-groups in larger cohorts where more survival events were registered.

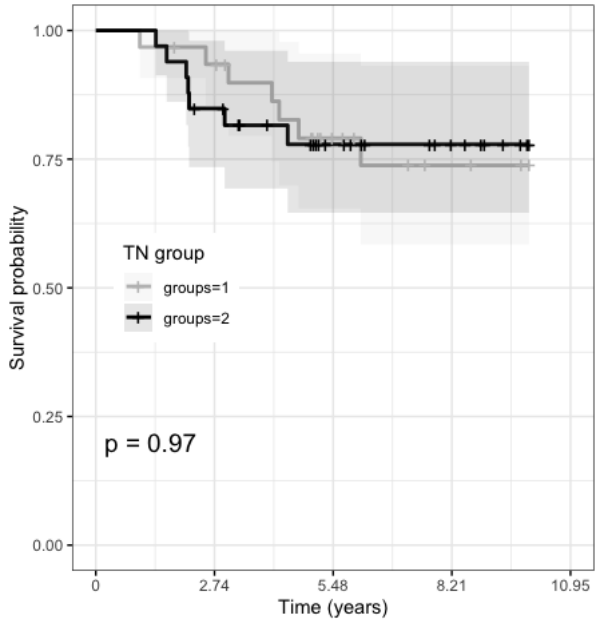


Figure 2.8 – The survival function of RATHER-NKI samples estimated by the Kaplan-Meier estimator, color code indicate the TN group. The p-value associated to the log-rank test is printed.

2.3.3. Validation

Validating the identified RATHER-NKI classification at a proteomic level is important. Indeed, this would allow to verify that the proteomic signal, generating this classification, can be generalized to other datasets. To do so, we proceeded in a statistically more robust manner than is normally done in TN classification studies by computing the posterior probability of the GMM training model. This allowed us to predict the group belonging of the TN samples in the two validation sets: RAHTER-Cambridge and Curie datasets.

RATHER-Cambridge validation set

Classification prediction. Since all the, $p = 116$, proteins were in common for the RATHER-NKI and the RATHER-Cambridge datasets, we could predict the cluster labels (group belonging) for the $n = 31$ TN Cambridge samples, by computing the posterior probability of the GMM-EII training model presented in Section 2.3.1. The two generated groups consisted of $n_1 = 16$ verses $n_2 = 15$ TN samples each. We projected these two TN groups into the two first axes of the RATHER-NKI PCA space and as can be seen in Figure 2.11a the two

groups are well represented by these two axes and largely overlapping with the RATEHR-NKI TN groups.

Group characterization. Five proteins were differently expressed among two TN groups ($p.value < 0.05$, FDR corrected): CaMKI, IRS1, Met, Phospho-HER2-ErbB2-Tyr877 and Phospho-VEGF-Receptor2-Tyr1214. These five proteins were all, as well, differently expressed among the two TN groups in the RATHER-NKI dataset. The two RAHTER-Cambridge TN groups had similar proteomic expression pattern for the two protein groups found in the RATHER-NKI dataset as the RATHER-NKI TN groups. Indeed, as can be seen in Figure 2.9, each TN group was characterized by an "higher" proteomic expression of a given protein group. As can be seen in Figure 2.10, no difference in survival rate was found for the two groups.

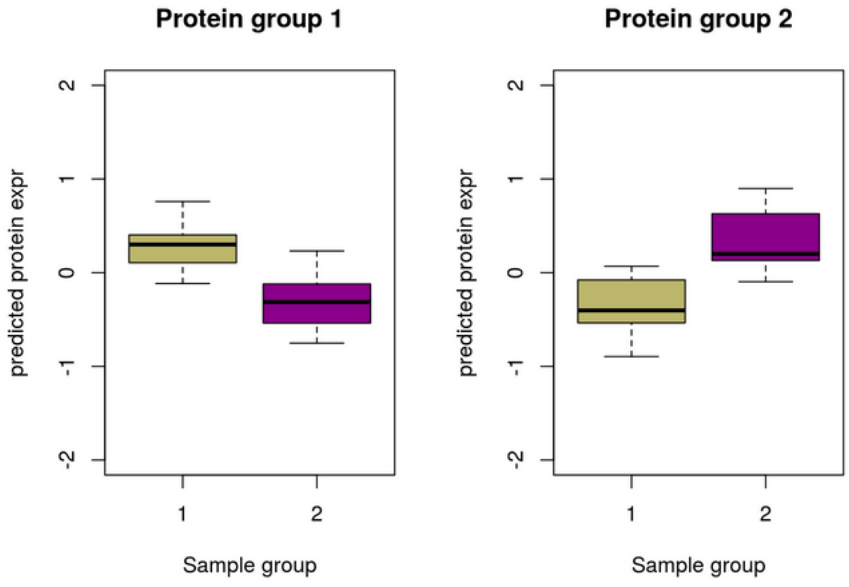


Figure 2.9 – Boxplot with median values, first and third quantile for the two RATHER-Cambridge TN groups and the two groups of proteins found in the RATHER-NKI dataset.

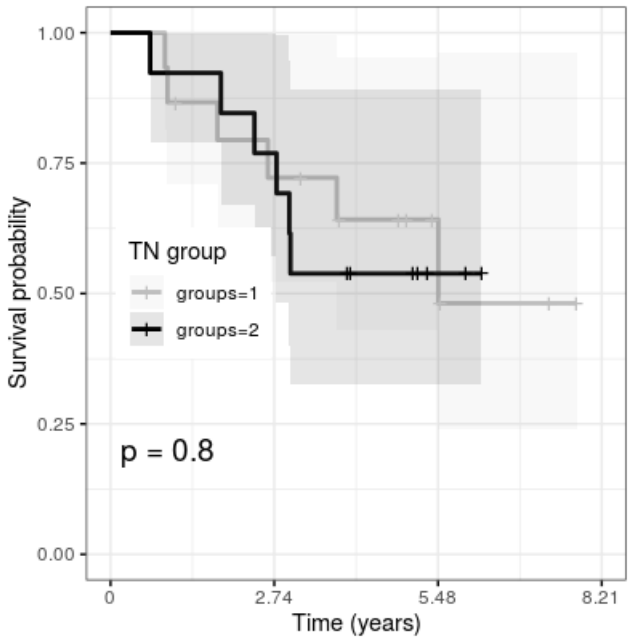


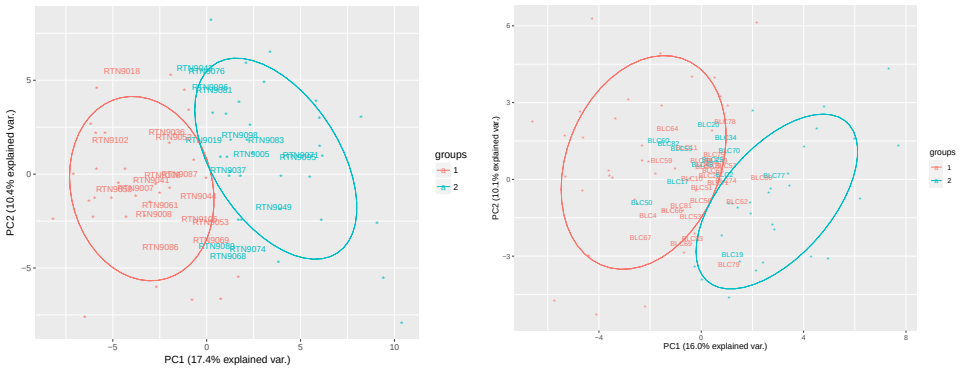
Figure 2.10 – The survival function of RATHER-Cambridge samples estimated by the Kaplan-Meier estimator, color code indicate the TN group. The p-value associated to the log-rank test is printed.

Curie validation set

The Curie validation set ($n = 42, p = 249$) had $p = 70$ proteins in common with the RATHER dataset. Hence, in order to predict the groups in this validation set we computed a GMM-EEI training model of the RATHER-NKI dataset including only these proteins. When the posterior probability of this model was computed for the Curie validation samples, two groups were identified with $n_1 = 13$ and $n_2 = 29$ observations each. No proteins were differentially expressed at a significant level among these two TN groups. The projection of the Curie TN samples in the space of the first two principal components of the RATHER-NKI set (based on the 70 included proteins) is illustrated in Figure 2.11b). As it can be seen, the Curie TN samples are regrouped in the center and do not separate between the two groups, which are superposed.

It should be noticed that a strong technical batch effect of the RPPA experiment was as well observed in this dataset, which was corrected by the median regression as described in Section 2.3.1. There was also a problem with extreme values due to sample manipulation. Moreover, no cluster struc-

ture was found in the dataset using either GMM/BIC or HAC/cluster stability procedures (results not shown).



(a) Cambridge projected TN samples (b) Curie projected TN samples

Figure 2.11 – Representing first two axis of RATHER-NKI PCA. Left panel, Curie TN samples projected to NKI PCA based on $p = 70$ proteins. Right panel, Cambridge TN samples NKI PCA based on all $p = 116$ proteins. Color code indicates the two TN samples groups. Training RAHTER-NKI samples are represented as * whereas the validation TN samples are represented by their study id.

2.3.4. Supervised classification - Searching for a gene signature

We aimed to identify a transcriptomic signature that would be able to separate samples according to the identified TN proteomic classification. This would enable us to predict the protein classification from transcriptomic data and thus do so in other datasets for which (1) no protein data, but only transcriptomic data, are available, or (2) few proteins are in common (as for the publicly available TCGA-BRCA dataset). To find such a transcriptomic signature, linear discriminant analysis was conducted as well as penalized logistic regression for the following different combinations of genes:

- All genes ($q = 21\ 508$)
- The highly varying genes *standard deviation* > 0.8 ($q = 4\ 592$)

- The genes that were significantly correlated to any of the, $p = 116$, ($p.value < 0.5$ FDR corrected) proteins ($q = 500$)
- The genes that were biologically linked to at least one protein ($q = 128$).

Linear Discriminant Analysis: No linear combination of the genes was found that could predict the proteomic based TN tumor groups.

Penalized Logistic regression: The misclassification rate for this model was at best above chance, indicating that it was not able to predict the two TN tumor groups.

2.4. Discussion

2.4.1. Main results

In the present study, we performed unsupervised clustering and classification analyses on TN breast cancer tumors. The aim was to identify subgroups having distinct molecular profiles that could be clinically helpful for the identification of new treatment targets. The novelty in this work was that:

1. the classification was conducted on proteomic data instead of transcriptomic data,
2. a normalized cluster stability index was implemented in order to determine the number of TN groups,
3. unsupervised classification was conducted in a probabilistic setting that allowed us to validate the training model by an information criterion and compute the posterior probability of this model to predict the classes in the validation sets,
4. supervised classification methods were used to search for a transcriptomic signature that could explain the identified proteomic based TN groups.

To conduct these analyses, the HAC-Ward and the GMM clustering algorithms were used to classify the RATHER-NKI training set ($n = 67, p = 116$). A stable classification was fond with two groups with $n_1 = 33$ and $n_2 = 34$ TN samples each. This classification was also validated by the BIC. The classification produced by the GMM-EEI model was retained and differential analysis

showed that there were $p = 38$ proteins differentially expressed between the two TN groups. These proteins could be classified into two groups for which the two TN groups showed different expression patterns. No difference in survival rates was found for the two TN groups. However, it should be noted that very few events were observers (only 23%) and the patients, within each group, had received different treatments. The posterior probability of the RATHER-NKI GMM-EEI training model was computed in two datasets in order to validate the obtained classification. By doing so the classification was validated in the RATHER-Cambridge dataset ($n = 31, p = 116$) where the two identified groups, with $n_1 = 16$ and $n_2 = 15$ observations each, had similar proteomic expression patterns as the RATEHR-NKI TN groups. Also, the two RATHER-Cambridge groups were well scattered among the RATHER-NKI groups in the first two dimensions of the RATHER-NKI PCA space. However, the RATHER-NKI classification was not validated in the Curie dataset ($n = 42$), containing $p = 249$ proteins whereof $p = 70$ were in common with the RATHER dataset. The two discovered groups $n_1 = 13$ and $n_2 = 29$ did not show any differentially expressed proteins and were not separable by the first two axis of the RATHER-NKI PCA space. Moreover, we did not find any transcriptomic signature that could relate to the RATHER-NKI classification.

Two important results were also found concerning the preprocessing of the RPPA data:

1. When the samples from the RATHER-NKI and RATHER-Cambridge dataset were analyzed together, an artificial classification, that was not found in neither of the datasets when analyzed separately, was generated.

2. Even though the RPPA data were corrected by the NormaCurve preprocessing procedure, proteins that were analyzed within the same experiment were strongly correlated with each other. This was observed for both the RATHER and the Curie datasets. This batch effect, before being corrected, was the most prominent signal in the RATHER dataset and generated a stable clustering of $K = 3$. Once the batch effect was corrected, this clustering was no longer found.

2.4.2. Some limits

There are many plausible reasons why we did not manage to validate the RATHER-NKI classification in the Curie dataset.

Biological reasons. First, it should be stressed out that tumors collected from different centers differ. For example, the Curie TN samples contained at least 50% of cancer cells whereas the RATHER TN samples contained only $> 30\%$. Hence, we are not comparing the same kind of tumors. Second, the protein expression observed for the TN samples is relative to the studied cohort. RPPA analysis of tissue samples from healthy controls need to be analyzed in order get a baseline. Third, the proteins, even though a large proportion were in common, differed between the RATHER and the Curie cohorts.

Also, the number of tumor samples in the Curie dataset was small. It would therefore be important to validate the two identified TN-groups on other TN proteomic data sets, preferentially including more patients. However, TN data sets with proteomic data are still rare and the studied proteins might also differ according to the different data sets. For this reason, we wanted to be able to predict the proteomic TN-groups from transcriptomic data, for which several large public TN cancer data sets are available, for example the TCGA dataset. Yet, we did not manage to find any transcriptomic signature that could predict the groups. It should also be noticed that the links between the RNA and the proteomic data were few. Indeed, only 500 of the 21 508 genes were correlated to the proteins. This is somewhat expected since a large proportion of the proteins were phosphorylated. Since protein phosphorylation can not be seen in RNA expression, this might, to some part, explain why we did not find any transcriptomic signature able to predict the groups.

Experimental reasons. In this study, we also observed some preprocessing issues and experimental effects linked to the RPPA technique. First, the fact that the analysis of RATHER tumors had to be conducted for tumors collected in the two centers separately gives an indication on how sensitive the RPPA data analysis is to experimental bias. For example, the observed differences between the two centers might be due to differences in time laps between the sectioning and the freezing of tumor samples in the two centers. In the study of Bonsang-Kitzis et al. (2016), the authors point out the importance

of normalizing RNA data coming from different centers separately. Indeed, as argued by these authors, not doing so might induce a center effect to the TN classification. Even though, we conducted the median normalization separately for the RATHER-NKI and the RATHER-Cambridge dataset the data were still processed by the NormaCurve together. This might have induced an effect that we did not manage to correct and more analyses are necessary to understand if this was the case.

Second, the experimental batch effect observed in both the RAHTER and the Curie dataset is another indication on the sensitivity of the RPPA technique. This observed batch effect is quite concerning since, when not corrected, it induces a strong signal in the dataset. Indeed, when it was not corrected a classification of $K = 3$ groups was found in the RATHER dataset that then disappeared once corrected. The NormaCurve normalization procedure should be revised in order to take into account these effects.

These observations stress out the importance of validating the training model, here the RATHER-NKI classification, in external datasets. Indeed, it would be of interest to know whether the proteomic classification proposed by Masuda et al. (2017) would have similar validation issues if it were predicted on an external dataset. This would give supplementary insight to the usability of the RPPA data for classifying TN tumors. Also, these results emphasize the importance of publishing the preprocessing procedure, particularly for data as complex as the RPPA data. Again, it would be interesting to know whether the authors in the study of Masuda et al. (2017) would have encountered similar normalization issues. However, normalization pipelines are generally ignored or described to its minimum, in TN classification studies. This is a pitty since it makes the reproducibility of results difficult. Indeed, even if some TN classification studies have been using public available data, one cannot validate their classification if the normalization pipeline is not published or shared. This highlights the importance of open science where data and code are shared in the community. It also highlights the importance for biologists and statisticians to work together on the topic of data preprocessing. These issues will be discussed in some more details in the general discussion Chapter 6.

2.4.3. Conclusion

To conclude, in this study I found a stable clustering of two groups in the RATHER-NKI training set. These two groups had different proteomic patterns but did not differ in clinical ways or in survival rates. The two groups were found in the RATHER-Cambridge cohort, which showed similar proteomic expression patterns, but not in the Curie dataset. Also, I did not find a transcriptomic signature that could predict these groups in other datasets. In sight of these results and the observed preprocessing issues, it might be that, the observed classification in the RATHER dataset is induced by another (but not yet discovered) experimental effect. Also, similar preprocessing problems of the RPPA data were found in a project we conducted in collaboration with an industrial testing the effect of an TTK inhibitor and presented in Appendix A.1. For all those reasons, I decided to no longer work with the RPPA data but concentrate on RNA data in the rest of my thesis.

Chapter 3

Adjusting the adjusted Rand Index - A multinomial story

Résumé. L'indice de Rand ajusté (Adjusted Rand Index ARI) est sans doute l'une des mesures les plus populaires pour la comparaison des clusters. L'ajustement du ARI est basé sur une hypothèse de distribution hypergéométrique qui n'est pas satisfaisante du point de vue de la modélisation, car (i) elle n'est pas appropriée lorsque les deux clusterings sont dépendants, (ii) elle force la taille des clusters (tout observation étant considéré d'être observé), et (iii) elle ignore le caractère aléatoire de l'échantillonnage. Dans ce travail, nous présentons une nouvelle version "modifiée" de l'indice. Tout d'abord, nous redéfinissons le MRI (Modified Rand Index) en ne comptant que les paires cohérentes par similarité et en ignorant les paires cohérentes par différence, ce qui augmente l'interprétabilité du score. Deuxièmement, nous basons la version ajustée, $MARI$, sur une distribution multinomiale au lieu d'une distribution hypergéométrique. Le modèle multinomiale est avantageux, car il ne force pas la taille des clusters, modélise correctement le caractère aléatoire, et est facilement étendu au cas dépendant. Nous montrons que l'ancien ARI est biaisé dans le modèle multinomiale et que la différence entre le ARI et le $MARI$ peut être importante pour les petits n, mais disparaît essentiellement pour les grands n, où n est le nombre d'individus. Enfin, nous fournissons un algorithme efficace pour calculer toutes ces quantités ($(A)RI$ et $M(A)RI$) en nous appuyant sur une représentation épurée du tableau de contingence de notre package `aricode`. La complexité spatiale et temporelle est linéaire en ce qui concerne le nombre d'échantillons et, surtout, ne dépend pas du nombre de clusters, car nous ne calculons pas explicitement le tableau de contingence.

Abstract. The Adjusted Rand Index (ARI) is arguably one of the most popular measures for cluster comparison. The adjustment of the ARI is based on a hypergeometric distribution assumption which is unsatisfying from a modeling perspective as (i) it is not appropriate when the two clusterings are dependent, (ii) it forces the size of the clusters, and (iii) it ignores randomness of the sampling. In this work, we present a new "modified" version of the Rand Index. First, we redefine the MRI by only counting the pairs consistent by similarity and ignoring the pairs consistent by difference, increasing the interpretability of the score. Second, we base the adjusted version, $MARI$, on a multinomial distribution instead of a hypergeometric distribution. The multinomial model is advantageous as it does not force the size of the clusters, properly models randomness, and is easily extended to the dependant case. We show that the ARI is biased under the multinomial model and that the difference between the ARI and $MARI$ can be large for small n but essentially vanish for large n, where n is the number of individuals. Finally, we provide an efficient algorithm to compute all these quantities ($(A)RI$ and $M(A)RI$) by relying on a sparse representation of the contingency table in our `aricode` package. The space and time complexity is linear in the number of samples and importantly does not depend on the number of clusters as we do not explicitly compute the contingency table.

3.1. Introduction

With the increasing amount of data available, development of clustering methods have become crucial in unsupervised learning to explore and find patterns in data sets. Despite the wealth of theoretical research on this subject, in practice selecting and validating a clustering is difficult. To answer these questions, one often resorts to a measure of clustering comparison: when the data is labeled, the quality of the clustering is evaluated by measuring the overlap with the original labeling; in the absence of labels, the reliability of the clustering can be assessed by evaluating its stability (see, e.g. Von Luxburg et al., 2010). This can be done by comparing several clusterings obtained by perturbing the initial data set (i.e. with resampling), or by running different clustering methods on the same data set. The idea of clustering stability is dug deeper in cluster ensembles (Strehl and Ghosh, 2002) and its variants, which involve measures of clustering comparison in the construction of the clustering itself.

Among the many measures proposed for pairwise clustering comparisons

(see Vinh et al., 2010, for an overview) one of the most popular is the Rand index (RI) (Rand, 1971) and its adjusted variant (Hubert and Arabie, 1985; Morey and Agresti, 1984). The RI is designed to estimate the probability of having a coherent pair, which is a pair for which its two observations are either in the same group in the two compared clusterings or in different groups. It is computed from the contingency table of the two classifications. However, the RI depends on the number of groups (Morey and Agresti, 1984) and is therefore difficult to interpret. To overcome this issue, the Adjusted Rand Index (in short ARI) is obtained by subtracting to the RI an estimator of its expected value obtained under the assumption of two independent clusterings.

To obtain such an estimator, a population distribution has to be assumed upon the two compared clusterings, or more specifically upon the marginals of the contingency table of the two clusterings. Considering either the clusters sizes fixed or not, the two natural hypotheses that arise are either the hypergeometric distribution or the multinomial distribution. In the literature, there is discordance as to which of these hypotheses to use.

The RI and ARI as defined by Brennan and Light (1974) and then adapted by Hubert and Arabie (1985) are based on the hypergeometric distribution hypothesis. In fact, considering fixed cluster sizes makes calculations easier and the expected value of the RI deterministic. However, this is a strong assumption that is violated in all cluster studies since no clustering algorithm fixes cluster sizes (see Wagner and Wagner, 2007, for a detailed discussion). Moreover, from a modeling perspective, it implicitly ignores any randomness of the sampling procedure and considers that the set of individuals that we observed is fixed. Hence under this model the $(A)RI$ are post-hoc quantities for which no inference to a parental population can be done, which limits the interpretation exclusively to the observed data points. Assuming the marginal to be fixed certainly simplifies the calculations under the hypothesis of independence between clustering. However, modeling dependency between clusterings under this assumption is not straightforward and rather unnatural compared to the multinomial model. Yet one certainly hopes to compare clusterings that are alike or dependant.

In comparison, the multinomial model does not assume the size of the clusters to be fixed, by considering a sample observed from an infinite population. Modeling dependent clusterings and adjusting accordingly is then greatly simplified. For all these reasons we argue that the multinomial model is more natural from a statistical perspective. Note that Morey and Agresti (1984) already

studied this model to propose an adjusted version of the RI. Nonetheless, as pointed out in Hubert and Arabie (1985); Steinley (2004); Steinley and Brusco (2018), Morey and Agresti made an error in their calculation of the expected value of the RI, assuming that the expected value of a squared variable is the square of the expected value, which is wrong in general. We are convinced that this error is the reason for the problem described in Steinley and Brusco (2018), advocating unfairly for the hypergeometric version of the $(A)RI$.

§

In this work, we essentially make a rigorous statistical analysis of the RI under the hypothesis of a multinomial distribution. In details, our contributions are the following:

1. Define new versions of the RI and the ARI, denoted by MRI and $MARI$ (for "modified" $(A)RI$), only counting consistent pairs by similarity. Indeed, we show that counting consistent pairs by dissimilarity is unnecessary and blurs the interpretation. In terms of our newly defined $MARI$, considering those pairs would simply result in a multiplication by 2.

2. Finalise the work of Morey and Agresti (1984) and derive an unbiased estimator of the expected value of the MRI under a multinomial distribution valid for data under $\mathcal{H}_1$ (dependent clusterings) and $\mathcal{H}_0$ (independent clusterings).

3. Provide an efficient algorithm to compute all these quantities ($(A)RI$ and $M(A)RI$) by relying on a sparse representation of the contingency table. The complexity is in $\mathcal{O}(n)$ time and space where n is the number of individuals. This is better than the usual $\mathcal{O}(n + KL)$ complexity, where K and L are the sizes of the two clusterings one which to compare, typically obtained when using the non-sparse contingency table. Our code is available in versions $\geq$ 1.0.0 of the `R` package `aricode` (Chiquet et al., 2020).

4. Investigate the difference with the hypergeometric Hubert and Arabie's ARI and show that it is biased under the multinomial distribution, even if the difference between the two estimators remains small. This is in contradiction with the results of Steinley and Brusco (2018) that used the faulty ARI of Morey and Agresti (1984).

3.2. Statistical Model

3.2.1. A new Rand Index - counting only pairs consistent by similarity

The Rand Index (RI) proposed by Rand (1971) counts all the consistent pairs in two given classifications. In details, let us consider two classifications C^1 and C^2 in respectively K and L classes of the same n individuals. The labels of individual i are given by $c_i^1 \in [1,\dots,K]$ and $c_i^2 \in [1,\dots,L]$. The consistent pairs are all pairs where observations i and j are in the same group (consistent by similarity), or in different groups (consistent by difference) in C^1 and C^2.

We introduce the two quantities c_{ij}^1 and c_{ij}^2 indicating whether i and j are in the same group for respectively classification C^1 and C^2 :

$$c_{ij}^1 = \begin{cases} 1 & \text{if } c_i^1 = c_j^1 = k, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad c_{ij}^2 = \begin{cases} 1 & \text{if } c_i^2 = c_j^2 = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

Note that c_{ij}^1 and c_{ij}^2 are the realisations of Bernoulli random variables denoted by C_{ij}^1 and C_{ij}^2 that will prove useful later in our statistical analysis, while studying the RI and other similar quantities as random variables.

Using these two quantities we see that a pair is consistent by similarity if $c_{ij}^1 c_{ij}^2 = 1$ and consistent by difference if $(1-c_{ij}^1)(1-c_{ij}^2) = 1$. Now considering all pairs, we get the following formula for the RI as defined by Rand:

$$\begin{aligned} RI(C^1,C^2) &= \frac{1}{\binom{n}{2}} \sum_{i<j} c_{ij}^1 c_{ij}^2 + \sum_{i<j} (1-c_{ij}^1)(1-c_{ij}^2) \\ &= 1 + \frac{1}{\binom{n}{2}} \left[2 \sum_{i<j} c_{ij}^1 c_{ij}^2 - \sum_{i<j} c_{ij}^1 - \sum_{i<j} c_{ij}^2 \right]. \end{aligned} \tag{3.1}$$

In Equation (3.1), we remark that only the product $\sum c_{ij}^1 c_{ij}^2$ depends on the joint distribution of C^1 and C^2: all other terms, coming exclusively from coherent pairs by difference, depend on the marginal distributions of C^1 and C^2. These terms will thus be cancelled out in any adjusted version of the RI, correcting for what would happen if C^1 and C^2 were drawn independently. Hence, we argue that considering the consistent pairs by difference unnecessarily complicates the reasoning and the probabilistic analysis of the RI. For simplicity we thus redefine the index and refer to it as the MRI (for "modified"

RI):

$$MRI(C^1,C^2) = \frac{1}{\binom{n}{2}} \sum_{i<j} c^1_{ij} c^2_{ij}. \tag{3.2}$$

Remark. *For the derivation of the expected value of MRI, RI and their adjusted version MARI and ARI, using the definition involving c^1_{ij} and c^2_{ij} (or more exactly C^1_{ij} and C^2_{ij} in a probabilistic perspective) considerably simplify the calculations compared to their classical combinatorial formulations. These combinatorial formulations are recalled in the next section as they are classically used to compute the RI and its variants.*

3.2.2. Computing the Rand Index from the $n_{k\ell}$ contingency table

The information from two observed classifications is usually summarized in a contingency table like Table 3.1, representing the number of observations $n_{k\ell}$ in group k in C^1 and in group ℓ in C^2.

$C^1 \backslash C^2$	c^2_1	$\cdots$	c^2_ℓ	$\cdots$	c^2_L	Sums
c^1_1	n_{11}	$\cdots$	$n_{1\ell}$	$\cdots$	n_{1L}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c^1_k	n_{k1}	$\cdots$	$n_{k\ell}$	$\cdots$	n_{2L}	$n_{k.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c^1_K	n_{K1}	$\cdots$	$n_{K\ell}$	$\cdots$	n_{KL}	$n_{K.}$
Sums	$n_{.1}$	$\cdots$	$n_{.\ell}$	$\cdots$	$n_{.L}$	$\sum_{k\ell} n_{k\ell} = n$

Table 3.1 – Contingency Table between clusterings C^1 and C^2; each entry $n_{k\ell}$ corresponds to the number of observations in group k in C^1 and group ℓ in C^2.

Using basics combinatorics we get the following relations between $n_{k\ell}, n_{k.}, n_{.\ell}$ and c^1_{ij},c^2_{ij}:

$$\sum_{i<j} c^1_{ij} = \sum_k \binom{n_{k.}}{2}, \quad \sum_{i<j} c^2_{ij} = \sum_\ell \binom{n_{.\ell}}{2} \text{ and } \sum_{i<j} c^1_{ij} c^2_{ij} = \binom{n_{k\ell}}{2}. \tag{3.3}$$

Expressions (3.1) and (3.2) of RI and MRI turn to

$$MRI(C^1,C^2) = \frac{1}{\binom{n}{2}} \sum_{k,\ell} \binom{n_{k\ell}}{2} = \frac{1}{2\binom{n}{2}} \sum_{k,\ell} n_{k\ell}^2 - n \tag{3.4}$$

$$RI(C^1,C^2) = 1 + \frac{2}{\binom{n}{2}} \sum_{k,l} \binom{n_{k\ell}}{2} - \frac{1}{\binom{n}{2}} \left[\sum_k \binom{n_{k.}}{2} + \sum_l \binom{n_{.l}}{2} \right]. \tag{3.5}$$

Using these formula, one can see that the minimum of the MRI is obtained when all $n_{k\ell}$ are equal, which has a simple and straightforward interpretation (as two perfectly independent and balanced clusterings). On the other hand the minimum of the RI is obtained for an extremely unbalanced table, *i.e.* when one of the two clustering consists of a single cluster and the other only of clusters containing single points. This makes the interpretation of the RI rather difficult (i.e. the lowest value is not obtained for two perfectly independent and balanced clusterings) and give more credibility to the definition of MRI that does not consider consistent pairs by difference.

3.2.3. Probabilistic model and properties of the Rand Index

So far, the $(M)RI$ have been computed from the *observed quantities* c_{ij}^1, c_{ij}^2, or equivalently from the observed contingency table $n_{k\ell}$. From now, we aim to study the statistical properties of the MRI and consider its status of random variable[1]:

$$MRI(C^1,C^2) = \frac{1}{\binom{n}{2}} \sum_{i<j} C_{ij}^1 C_{ij}^2, \tag{3.6}$$

where we recall that C_{ij}^1 and C_{ij}^2 are Bernoulli random variables indicating whether individual i and j are in the same groups in classification C^1 respectively C^2.

To derive the probability of success associated to C_{ij}^1 and C_{ij}^1, we need a probabilistic model for the classification of a given individual in C^1 and C^2, that is, a counterpart for generating the two observed clusterings c_i^1 and c_i^2 for the n data points. We denote by C_i^1 and C_i^2 the corresponding random variables. A natural model is the multinomial model, which give the joint

[1]By a slight abuse of notation, we use MRI for both its observed value and its definition as a random variable. We think that the context suffices for the reader to remove any ambiguity.

distribution of (C_i^1, C_i^2) as follows: for all $(k,\ell) \in \{1,\dots,K\} \times \{1,\dots,L\}$,

$$\mathbb{P}(C_i^1 = k, C_i^2 = \ell) = \pi_{k\ell}, \quad \text{s.c.} \sum_{k,\ell}^{K,L} \pi_{k\ell} = 1.$$

The marginal probabilities of a given group is defined for k in C^1 by $\sum_\ell^L \pi_{k\ell} = \pi_{k.}$ and for ℓ in C^2 by $\sum_k^K \pi_{k\ell} = \pi_{.\ell}$. See Table 3.2 for a global picture. Compared to the hypergeometric model, the multinomial model easily deals with dependent classifications and does not force the size of the clusters.

$C^1 \backslash C^2$	$c_i^2 = 1$	$\cdots$	$c_i^2 = \ell$	$\cdots$	$c_i^2 = L$	Sums
$c_i^1 = 1$	π_{11}	$\cdots$	$\pi_{1\ell}$	$\cdots$	π_{1L}	$\pi_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$c_i^1 = k$	π_{k1}	$\cdots$	$\pi_{k\ell}$	$\cdots$	π_{2L}	$\pi_{k.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$c_i^1 = K$	π_{K1}	$\cdots$	$\pi_{K\ell}$	$\cdots$	π_{KL}	$\pi_{K.}$
Sums	$\pi_{.1}$	$\cdots$	$\pi_{.\ell}$	$\cdots$	$\pi_{.L}$	$\sum_{k\ell} \pi_{k\ell} = 1$

Table 3.2 – Multinomial model: probabilistic distributions $\pi_{k\ell} = \mathbb{P}(C_i^1 = k, C_i^2 = \ell)$ and marginal distributions $\pi_{k.} = \mathbb{P}(C_i^1 = k)$ and $\pi_{.\ell} = \mathbb{P}(C_i^2 = \ell)$

Based on this multinomial model for C_i^1 and C_i^2, it is then relatively straightforward to derive the joint distribution and marginals of C_{ij}^1 and C_{ij}^2. In particular we have:

$$\mathbb{P}(C_{ij}^1 = 1) = \sum_k \pi_{k.}^2, \quad \mathbb{P}(C_{ij}^2 = 1) = \sum_\ell \pi_{.\ell}^2 \text{ and } \quad \mathbb{P}(C_{ij}^1 C_{ij}^2 = 1) = \sum_{k,\ell} \pi_{k\ell}^2. \tag{3.7}$$

However, in order to derive the expectation, variance and unbiased adjustment of the MRI under the multinomial model, one not only needs to characterize events on the classification C^1 and C^2 on (unordered) pairs of individual $\{i,j\}$, but also on *pairs of pairs* of individual $\{i,j\}$ and $\{i',j'\}$, with terms like the expectation of $C_{ij}^1 \times C_{i'j'}^2$. The following section derives a couple of technical – yet simple – lemmas, on events implying such random variables so that the final calculation of the moments of MRI under the multinomial model are straightforward.

Remark. *To our knowledge most derivations of the expectation and variance of the RI found in the literature are based on the combinatorial formulation*

given in Equation (3.5)*: these derivations rely on general results on the moments of either the multinomial or the generalized hypergeometric distribution and involve tedious calculations. In contrast, our proofs, found in the next sections, are short, self-contained and easily accessible to any reader with some basic knowledge in probability and statistics. For this reason we argue that our proofs are interesting in their own rights.*

Subsets of Pairs of Pairs - preparing the derivations of the moments of the *MRI*

Consider $\{i,j\}$ and $\{i',j'\}$ the $\mathcal{P}\times\mathcal{P}$ set of unordered pairs of $\{1,\dots,n\}^2$ such that $i<j$ and $i'<j'$. This set is composed by pairs of pairs, and can equivalently be seen as the set of all quadruplets of $\{1,\dots,n\}^4$ such that $i<j$ and $i'<j'$. We partition this set into the three following subsets:

1. the unordered pairs $\mathcal{P}$,
2. the ordered-triplets $\mathcal{T}$,
3. the ordered quadruplets $\mathcal{Q}$.

These three subsets $\mathcal{P},\mathcal{T}$ and $\mathcal{Q}$ makes a partition of $\mathcal{P}\times\mathcal{P}$ and in particular,

$$|\mathcal{P}|^2=|\mathcal{P}|+|\mathcal{T}|+|\mathcal{Q}|.$$

We now study respectively $\mathcal{P},\mathcal{T}$ and $\mathcal{Q}$ in the three following lemmas: we derive their cardinality and compute some expectations involving these subsets and the C^1_{ij}, C^2_{ij} variables under the multinomial model. These three lemmas will be the building blocks for the characterization of the MRI.

Lemma 3.2.1 (Subset of unordered pairs $\mathcal{P}$)**.** *With a slight abuse of notation, we consider $\mathcal{P}$ as a subset of $\mathcal{P}\times\mathcal{P}$:*

$$\mathcal{P}=\{\{i,j,i',j'\}:|\{i,j\}\cup\{i',j'\}|=2.\}$$

The cardinality of $\mathcal{P}$ is $|\mathcal{P}|=\binom{n}{2}$ and

$$\mathbb{E}\left(\sum_{i,j\in\mathcal{P}}C^1_{ij}C^2_{ij}\right)=\binom{n}{2}\sum_{k\ell}\pi^2_{k\ell}. \tag{3.8}$$

Proof. For any $i, j \in \mathcal{P}$, we have from (3.7) that $\mathbb{E}(C^1_{ij}C^2_{ij}) = \sum_{k\ell} \pi^2_{k\ell}$. We just need to sum over all possible pairs to get the desired result. □

Lemma 3.2.2 (Subset of ordered triplets $\mathcal{T}$)**.** *Consider the subset $\mathcal{T}$ of $\mathcal{P} \times \mathcal{P}$*

$$\mathcal{T} = \{\{i, j, i', j'\} : |\{i, j\} \cup \{i', j'\}| = 3\}.$$

The cardinality of $\mathcal{T}$ is $|\mathcal{T}| = n(n-1)(n-2)$ and

$$\mathbb{E}\left(\sum_{\mathcal{T}} C^1_{ij}C^2_{ij'}\right) = n(n-1)(n-2)\sum_{k\ell} \pi_{k\ell}\pi_{k.}\pi_{.\ell}. \tag{3.9}$$

$$\mathbb{E}\left(\sum_{\mathcal{T}} C^1_{ij}C^2_{ij}C^1_{ij'}C^2_{ij'}\right) = n(n-1)(n-2)\sum_{k\ell} \pi^3_{k\ell}. \tag{3.10}$$

Proof. For the cardinality of $\mathcal{T}$, one can map to the set of arrangements of $\{1, \ldots, n\}^3$.

For (3.9), remark that $C^1_{ij}C^2_{ij'}$ is a Bernoulli variable equal to 1 only when i and j are in the same cluster k in C^1 and i and j' are in the same cluster ℓ in C^2 . Hence, j can be in any cluster ℓ' in C^2 and j' can be in any cluster k' in C^1. From here one easily get its expectation,

$$\mathbb{E}(C^1_{ij}C^2_{ij'}) = \sum_{k\ell k'\ell'} \pi_{k\ell}\pi_{k'\ell}\pi_{k\ell'} = \sum_{k\ell k'\ell'} \pi_{k\ell} \sum_{k'} \pi_{k'\ell} \sum_{\ell'} \pi_{k\ell'} = \sum_{k\ell} \pi_{k\ell}\pi_{.\ell}\pi_{k.}$$

and we get the desired result by summing over all triplets.

For (3.10), remark that $C^1_{ij}C^2_{ij}C^1_{ij'}C^2_{ij'}$ is a Bernoulli variable equal to 1 if and only if i, j and j' are in the same clusters for both classifications. Summing over all $\mathcal{T}$ we get (3.10). □

Lemma 3.2.3 (Subset of ordered quadruplets $\mathcal{Q}$)**.** *Consider the following subset $\mathcal{Q}$ of $\mathcal{P} \times \mathcal{P}$:*

$$\mathcal{Q} = \{\{i, j, i', j'\} : |\{i, j\} \cup \{i', j'\}|\} = 4\}.$$

The cardinality is $|\mathcal{Q}| = 6\binom{n}{4}$ and

$$\mathbb{E}\left(\sum_{\mathcal{Q}} C^1_{ij}C^2_{i',j'}\right) = 6\binom{n}{4}\sum_{k,\ell} \pi^2_{k.}\pi^2_{.\ell}. \tag{3.11}$$

Proof. There are $\binom{n}{4}$ ways to pick 4 distinct elements of $\{1, ..., n\}^4$. We can then arrange those in $\binom{4}{2}$ to get an element of $\mathcal{Q}$. Hence, all together there

are $6\binom{n}{4}$ quadruplets. We get $\mathbb{E}(C^1_{ij}C^2_{ij'})$ using the fact that i, j, i', j' are all different and that their classes are drawn independently. We then sum over $\mathcal{Q}$. □

Expectation and Variance of the Rand Index

With Lemmas 3.2.1, 3.2.2 and 3.2.3, we are now equipped to easily derive the moments of the MRI. We use $\mathbb{E}$ for stating the expectation understood under the multinomial model in general. With the additional assumption of independence between the classification, what we refer to as the *null hypothesis*, we use $\mathbb{E}_{\mathcal{H}_0}$. This terms is classically used for adjusting the Rand index.

Proposition 3.2.1 (Expectations of the MRI)**.** *Let θ denote the expectation of the MRI and θ_0 the expectation under $\mathcal{H}_0$. Then,*

$$\theta = \mathbb{E}(MRI) = \sum_{k\ell} \pi^2_{k\ell}, \qquad \theta_0 = \mathbb{E}_{\mathcal{H}_0}(MRI) = \sum_{k\ell} \pi^2_{k.}\pi^2_{.\ell}$$

Proof. By Definition 3.6 and Lemma 3.2.1 we obtain θ. For θ_0, it suffices to replace π_{kl} by $\pi_{k.}\pi_{.l}$ in the previous formula. □

Similarly, we derive the expectation of the "usual" RI.

Proposition 3.2.2. *Let θ^{RI} denotes the expectation of the RI and θ^{RI}_0 the expectation under $\mathcal{H}_0$. Then,*

$$\begin{aligned}
\theta^{RI} &= \mathbb{E}(RI) = 1 + 2\sum_{k\ell} \pi^2_{k\ell} - \sum_k \pi^2_{k.} - \sum_\ell \pi^2_{.\ell} \\
\theta^{RI}_0 &= \mathbb{E}_{\mathcal{H}_0}(RI) = 1 + 2\sum_{k\ell} \pi^2_{k.}\pi^2_{.\ell} - \sum_k \pi^2_{k.} - \sum_\ell \pi^2_{.\ell}
\end{aligned}$$

Proof. Compared to the MRI, the only additional terms are $1 + \sum_{i,j} C^1_{ij} + \sum_{i,j} C^2_{ij}$. Using (3.7) and summing over all pairs $\mathcal{P}$ we get the desired results. □

We now continue with the variance of the MRI.

Proposition 3.2.3. *Let $\sigma^2 = \mathbb{V}(MRI)$ be the variance of the MRI. Then,*

$$\sigma^2 = \frac{1}{\binom{n}{2}}\left(\sum_{k,\ell}\pi^2_{k\ell} - \left[\sum_{k,\ell}\pi^2_{k\ell}\right]^2\right) + \frac{n(n-1)(n-2)}{\binom{n}{2}^2}\left(\sum_{k,\ell}\pi^3_{k\ell} - \left[\sum_{k,\ell}\pi^2_{k\ell}\right]^2\right)$$

Proof. To obtain the variance of the MRI, first rewrite the variance in terms of covariance:

$$\begin{aligned}\sigma^2 &= \frac{1}{\binom{n}{2}^2}\mathbb{V}\Big(\sum_{\mathcal{P}\times\mathcal{P}} C^1_{ij}C^2_{ij}\Big)\\ &= \frac{1}{\binom{n}{2}^2}\mathrm{Cov}\Big(\sum_{\mathcal{P}\times\mathcal{P}} C^1_{ij}C^2_{ij}, \sum_{\mathcal{P}\times\mathcal{P}} C^1_{ij}C^2_{ij}\Big)\\ &= \frac{1}{\binom{n}{2}^2}\sum_{\mathcal{P}\times\mathcal{P}}\mathrm{Cov}\Big(C^1_{ij}C^2_{ij}, C^1_{i'j'}C^2_{i'j'}\Big)\end{aligned}$$

We then split this final sum using our partition of $\mathcal{P}\times\mathcal{P}$. Also noticing that for all $\{i,j\},\{i',j'\}\in\mathcal{Q}$ we have $\mathrm{Cov}(C^1_{ij}C^2_{ij}, C^1_{i'j'}C^2_{i'j'}) = 0$, we get,

$$\begin{aligned}\sigma^2 &= \frac{1}{\binom{n}{2}^2}\left[\sum_{\mathcal{P}}\mathrm{Cov}\Big(C^1_{ij}C^2_{ij}, C^1_{ij}C^2_{ij}\Big) + \sum_{\mathcal{T}}\mathrm{Cov}\Big(C^1_{ij}C^2_{ij}, C^1_{ij'}C^2_{ij'}\Big)\right]\\ &= \frac{1}{\binom{n}{2}}\mathbb{V}\Big(C^1_{ij}C^2_{i,j}\Big) + \frac{n(n-1)(n-2)}{\binom{n}{2}^2}\mathrm{Cov}\Big(C^1_{ij}C^2_{ij}, C^1_{ij'}C^2_{ij'}\Big)\\ &= \frac{1}{\binom{n}{2}}\left(\sum_{k,\ell}\pi^2_{k\ell} - \left[\sum_{k,\ell}\pi^2_{k\ell}\right]^2\right) + \frac{n(n-1)(n-2)}{\binom{n}{2}^2}\left(\sum_{k,\ell}\pi^3_{k\ell} - \left[\sum_{k,\ell}\pi^2_{k\ell}\right]^2\right)\end{aligned}$$

We get the second line by enumerating the elements of $\mathcal{P}$ and $\mathcal{T}$. We get the third line using the definition of the covariance (for any two variable X and Y: $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$) and Lemmas 3.2.1 and 3.2.2.

□

Remark. *Importantly, for a fixed $\pi_{k\ell}$, σ^2 goes towards 0 when n grows to infinity: the larger n, the better the estimation of θ.*

The Rand Index depends on the number of groups

In the multinomial model with uniform clusters (equal cluster size), Morey and Agresti (1984) showed that θ_0^{RI} depends on the number of groups in C^1 and C^2. This is also true for MRI and easier to prove since it does not include the marginal terms of coherence by difference. We also prove the following lemma showing that if one splits a cluster of C^1 or C^2 into two, the MRI always decreases. Note that this latter lemma does not assume independence between classifications.

Lemma 3.2.4. *Consider two classifications C^1 and C^2 in $K+1$ respectively L clusters. Let $C^{1\prime}$ be the classification obtained by fusing two clusters of C^1. Then,*

$$MRI(C^1, C^2) \leq MRI(C^{1\prime}, C^2).$$

Also, for any distribution on C^1 and C^2 we have

$$\theta(C^1, C^2) \leq \theta(C^{1\prime}, C^2))$$

Proof. Assuming without loss of generality that clusters 1 and 2 were merged, we get

$$\begin{aligned} MRI(C^1, C^2) - MRI(C^{1\prime}, C^2) &= \frac{1}{2\binom{n}{2}}\Big(\sum_\ell n_{1\ell}^2 + n_{2\ell}^2 - (n_{1\ell} + n_{2\ell})^2\Big) \\ &= -\frac{1}{\binom{n}{2}}\sum_\ell n_{1\ell} n_{2\ell} \leq 0. \end{aligned}$$

Since the expectation is linear, we can consider any particular model on C^1 and C^2 to get the final result. □

3.2.4. The Adjusted version of the Rand Index

Since the $(M)RI$ depends on the number of groups, it needs to be adjusted for chance. A way to do so, is to subtract its expectation under the null hypothesis $\mathcal{H}_0$ (as motivated in Brennan and Light, 1974; Hubert and Arabie, 1985; Morey and Agresti, 1984). Ideally one would like to get $\theta - \theta_0$ with their true values. Under our multinomial model this quantity is

$$\theta - \theta_0 = \sum_{k\ell} \pi_{k\ell}^2 - \sum_{k\ell} \pi_{k.}^2 \pi_{.\ell}^2$$

which is equal to zero under $\mathcal{H}_0$ (independence of the classifications), that is, when $\pi_{k.}\pi_{.\ell} = \pi_{k\ell}$ for all k,ℓ. In practice, one can only estimate the quantities $\theta - \theta_0$ from observed classifications. Our goal is therefore to get an unbiased estimator of $\theta - \theta_0$.

The MRI being by definition an unbiased estimator of θ, we only need an unbiased estimator of θ_0, that is $\sum_{k\ell} \pi_{k.}^2 \pi_{.\ell}^2$. However, under the alternative $\mathcal{H}_1$ (i.e. when the compared classifications are not independent, the most natural case), deriving an unbiased estimator of θ_0 is trickier and depends on the model

assumption. Morey and Agresti (1984) proposed a plug-in estimator for the multinomial model, but as pointed out by Hubert and Arabie (1985); Steinley (2004); Steinley and Brusco (2018), they made errors in their calculations. In the next section we continue their work by proposing an unbiased estimator for θ_0. We also show that the hypergeometric estimator of Hubert and Arabie (1985) for θ_0, used as correction in the "traditional" ARI, is biased under our multinomial $\mathcal{H}_1$.

A new Adjusted Rand Index. We now define our own adjusted version of the MRI that we denote $MARI$:

$$MARI = \widehat{\theta} - \widehat{\theta}_0. \tag{3.12}$$

with

$$\widehat{\theta} = \sum_{\mathcal{P}} C_{ij}^1 C_{ij}^2 \Big/ \binom{n}{2} \qquad \widehat{\theta}_0 = \sum_{\mathcal{Q}} C_{ij}^1 C_{i'j'}^2 \Big/ 6\binom{n}{4}.$$

and its observed value,

$$MARI^{obs} = \sum_{\mathcal{P}} c_{ij}^1 c_{ij}^2 \Big/ \binom{n}{2} - \sum_{\mathcal{Q}} c_{ij}^1 c_{i'j'}^2 \Big/ 6\binom{n}{4}.$$

where we recall that c_{ij}^1 and c_{ij}^2 are the observed counterparts of C_{ij}^1, C_{ij}^2 and $\mathcal{P}, \mathcal{Q}$ are defined in Section 3.2.3.

Lemma 3.2.5. *Under the multinomial model, the $MARI$ is unbiased, that is,*

$$\mathbb{E}(MARI) = \theta - \theta_0.$$

Proof. The proof is straightforward using Lemma 3.2.3. □

Computing the $MARI$ from a contingency table. In practice, the comparison of two classifications is given as a contingency table as Table 3.1, and we thus need a formulation of the $MARI$ defined in (3.12) as a function of $n_{k\ell}$.

We already gave in (3.4) an expression of $\widehat{\theta}$ as a function of $n_{k\ell}$. As we will see, $\widehat{\theta}_0$ can as well be computed from the $n_{k\ell}$ contingency Table 3.1 even if summing over all elements of $\mathcal{Q}$ rather than $\mathcal{P}$ is a bit less straightforward. To get $\sum_{\mathcal{Q}} c_{ij}^1 c_{i'j'}^2$, we will use the term $\sum_{k\ell} n_{k.}^2 n_{.\ell}^2$ from which we will, as a direct

result of Definition (3.3), derive the $(\sum_{\mathcal{P}} c^1_{ij})(\sum_{\mathcal{P}} c^2_{i'j'})$ terms. These latter can be decomposed as follows:

$$(\sum_{\mathcal{P}} c^1_{ij})(\sum_{\mathcal{P}} c^2_{i'j'}) = \sum_{\mathcal{P}} c^1_{ij}c^2_{ij} + \sum_{\mathcal{T}} c^1_{ij}c^2_{ij'} + \sum_{\mathcal{Q}} c^1_{ij}c^2_{i'j'}. \tag{3.13}$$

It is then sufficient to subtract the terms of $\mathcal{P}$ and $\mathcal{T}$ from the left side of Equation (3.13) to get $\sum_{\mathcal{Q}} c^1_{ij}c^2_{i'j'}$. All terms summing over $\mathcal{P}$ are easy to recover (see Definition 3.3). However, the terms involving elements of $\mathcal{T}$ are more tedious to obtain and are derived in Lemma 3.2.6. The terms of $\mathcal{Q}$ derived in Lemma 3.2.7.

Lemma 3.2.6. *We have the following expression of $\sum_{\mathcal{T}} c^1_{ij}c^2_{ij'}$ in terms of $n_{k\ell}$:*

$$\sum_{\mathcal{T}} c^1_{ij}c^2_{i'j} = 2n + \sum_{k,\ell} n_{k.}n_{k\ell}n_{.\ell} - \sum_{k,\ell} n^2_{k\ell} - \sum_{k} n^2_{k.} - \sum_{\ell} n^2_{.\ell}$$

Proof. We need to consider all i in $\{1, ..., n\}$. Assuming for now that i is in classes (k,ℓ), that is $c^1_i = k$ and $c^2_i = \ell$, let us consider all j, j' such that $c^1_{ij}c^2_{ij'} = 1$. The term $c^1_{ij}c^2_{ij'}$ is equal to one if $c^1_j = k$ and $c^2_{j'} = \ell$. We then get different scenarios according to whether $c^1_{j'} = k$ or not and whether $c^2_j = \ell$. Those scenarios are enumerated in Table 3.3.

	j in ℓ	j not in ℓ
j' in k	$(n_{k\ell} - 1)(n_{k\ell} - 2)$	$(n_{k\ell} - 1)(n_{.\ell} - n_{k\ell})$
j' not in k	$(n_{k.} - n_{k\ell})(n_{k\ell} - 1)$	$(n_{k.} - n_{k\ell})(n_{.\ell} - n_{k\ell})$

Table 3.3 – Four scenarios to be considered for j and j' in the calculation of the terms in$\sum_{\mathcal{T}} c^1_{ij}c^2_{ij'}$ when i is in class (k, ℓ).

Summing all terms of Table 3.3 we get $n_{k.}n_{.\ell} + 2 - n_{k\ell} - n_{k.} - n_{.\ell}$. To account for all i belonging to class (k, ℓ) we then multiply by $n_{k\ell}$. Eventually we sum over all k, ℓ to recover

$$\begin{aligned}\sum_{\mathcal{T}} c^1_{ij}c^2_{ij'} &= \sum_{k,\ell} n_{k\ell}(2 + n_{k.}n_{.\ell} - n_{k\ell} - n_{k.} - n_{.\ell}) \\ &= 2n + \sum_{k,\ell} n_{k.}n_{k\ell}n_{.\ell} - \sum_{k,\ell} n^2_{k\ell} - \sum_{k} n^2_{k.} - \sum_{\ell} n^2_{.\ell}\end{aligned}$$

□

Lemma 3.2.7. *We have the following expression of $\sum_{\mathcal{Q}} c^1_{ij}c^2_{i'j}$ in terms of $n_{k\ell}$:*

$$\sum_{\mathcal{Q}} c^1_{ij} c^2_{i'j'} =$$
$$\Bigg[\sum_{k\ell} n^2_{k.} n^2_{.\ell} - \Bigg(4 \sum_{k\ell} \binom{n_{k\ell}}{2} + 4(2n + \sum_{k,\ell} n_{k.} n_{k\ell} n_{.\ell} - \sum_{k,\ell} n^2_{k\ell} - \sum_{k} n^2_{k.} - \sum_{\ell} n^2_{.\ell}) + 2n\Big(\sum_{k} \binom{n_{k.}}{2} + \sum_{\ell} \binom{n_{k.}}{2} \Big) + n^2 \Bigg) \Bigg] \Big/ 4$$

Proof. From Equation (3.3) we can derive $\sum_{k\ell} n^2_{k.} n^2_{.\ell}$ as a function of $\sum_{\mathcal{P}\times\mathcal{P}} c^1_{ij} c^2_{i'j'}$ and n, since, $\sum_k n^2_{k.} = n + 2\sum_{\mathcal{P}} c^1_{ij}$ and $\sum_\ell n^2_{.\ell} = n + 2\sum_{\mathcal{P}} c^2_{ij}$ with,

$$\begin{aligned} \sum_{k\ell} n^2_{k.} n^2_{.\ell} &= (2 \sum_{i<j} c^1_{ij} + n)(2 \sum_{i'<j'} c^2_{i'j'} + n) \\ &= 4 \sum_{\mathcal{P}\times\mathcal{P}} c^1_{ij} c^2_{i'j'} + 2n \Big(\sum_{\mathcal{P}} c^1_{ij} + \sum_{\mathcal{P}} c^2_{i'j'} \Big) + n^2 \end{aligned} \tag{3.14}$$

Using equation (3.13), we decompose $\sum_{\mathcal{P}\times\mathcal{P}} c^1_{ij} c^2_{i'j'}$ into terms of $\mathcal{P}$, $\mathcal{T}$ and $\mathcal{Q}$ and get,

$$\sum_{\mathcal{Q}} c^1_{ij} c^2_{i'j'} = \Big[\sum_{k\ell} n^2_{k.} n^2_{.\ell} - \Big(4 \sum_{\mathcal{P}} c^1_{ij} c^2_{ij} + 4 \sum_{\mathcal{T}} c^1_{ij} c^2_{ij'} + 2n \Big(\sum_{\mathcal{P}} c^1_{ij} + \sum_{\mathcal{P}} c^2_{ij} \Big) + n^2 \Big) \Big] \Big/ 4$$
$$= \Bigg[\sum_{k\ell} n^2_{k.} n^2_{.\ell} - \Bigg(4 \sum_{k\ell} \binom{n_{k\ell}}{2} + 4(2n + \sum_{k,\ell} n_{k.} n_{k\ell} n_{.\ell} - \sum_{k,\ell} n^2_{k\ell} - \sum_{k} n^2_{k.} - \sum_{\ell} n^2_{.\ell}) + 2n\Big(\sum_{k} \binom{n_{k.}}{2} + \sum_{\ell} \binom{n_{k.}}{2} \Big) + n^2 \Bigg) \Bigg] \Big/ 4$$

□

3.3. Implementation - package `aricode`

We implemented code for fast computation of the MRI and its adjusted version the $MARI$, as well as a number of other clustering comparison measures in the `R/C++` package `aricode`, which is available on CRAN.

Computing these measures is straightforward by means of the whole $K \times L$

contingency table. However, the time and space complexity is in $\mathcal{O}(n + KL)$, which is somewhat inefficient when K and L are large. Our implementation in `aricode` is in $\mathcal{O}(n)$: the key idea is that, given n observations, at most n elements of the $n_{k\ell}$ contingency matrix can be non zero. To recover these non zero elements one can proceed in two simple steps: first, all observations are sorted in lexicographical order in terms of their first and second cluster index. This can be done in $\mathcal{O}(n)$ using `bucket sort` (Cormen et al., 2001) or `radix sort` (as implemented in `R` (R Core Team, 2019b)). Note that once the observations are sorted, all i that are in clusters k and ℓ are one after the other in the data table. Thus, in a second step `aricode` counts all non zero n_{kl} in a single path over the data table. Internally this is done using `Rcpp` (Eddelbuettel et al., 2011).

In Figure 3.1 we compare our implementation of the standard ARI with the implementation of `mclust` (Scrucca et al., 2016a) (that uses the whole contingency table). As can be noted, the cost of the latter can be prohibitive for large vectors.

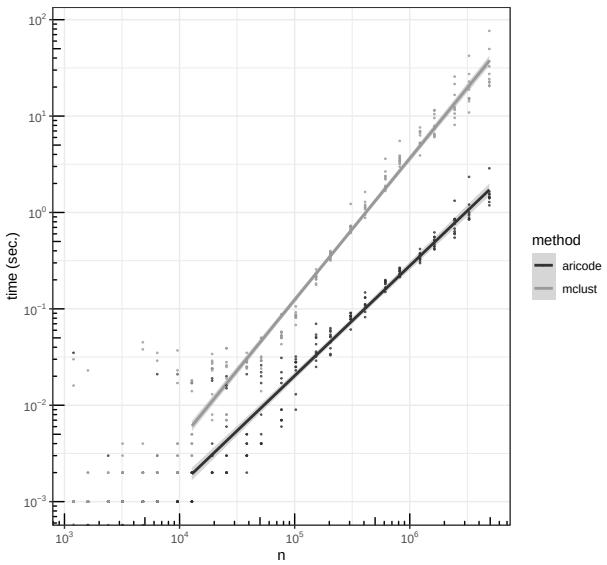


Figure 3.1 – Timings comparing the cost of computing the ARI with `aricode` or with the commonly used function `adjustedRandIndex` of the `mclust` package.

3.4. Hubert and Arabie's ARI

In this section we study the expectation of the 'standard' RI of Brennan and Light (1974) (by contrast with our MRI); the expression of which results from the hypergeometric model. This expression was used by Hubert and Arabie (1985) for adjusting the RI and producing the usual ARI. We study this expected value when the expectation corresponds to the multinomial distribution. We show that this estimator is biased in general under the alternative hypothesis, that is, when the two compared clusterings are not independent.

3.4.1. Expectation of Hubert and Arabie's ARI

Consider the observed value of the ARI proposed by Brennan and Light (1974); Hubert and Arabie (1985): in order to analyse this quantity in our multinomial setup, we first give its definition in terms of c_{ij}^1 and c_{ij}^2, that is

$$\begin{aligned} ARI^{\text{obs}} &= \frac{2}{\binom{n}{2}} \sum_{kl}^{KL} \binom{n_{kl}}{2} - \frac{2}{\binom{n}{2}^2} \sum_{kl}^{KL} \binom{n_{k.}}{2} \binom{n_{.l}}{2} \\ &= \frac{2}{\binom{n}{2}} \sum_{\mathcal{P}} c_{ij}^1 c_{ij}^2 - \frac{2}{\binom{n}{2}^2} \sum_{\mathcal{P}} c_{ij}^1 \sum_{\mathcal{P}} c_{ij}^2, \end{aligned}$$

where we recall that c_{ij}^1,c_{ij}^2 are realisations of the Bernoulli variables C_{ij}^1, C_{ij}^2. In a probabilistic perspective, we consider the ARI as a random variable:

$$ARI = \underbrace{\frac{2}{\binom{n}{2}} \sum_{\mathcal{P}} C_{ij}^1 C_{ij}^2}_{\widehat{\theta}^{RI}} - \underbrace{\frac{2}{\binom{n}{2}^2} \sum_{\mathcal{P}} C_{ij}^1 \sum_{\mathcal{P}} C_{ij}^2}_{\widehat{\theta}_0^{RI}}, \tag{3.15}$$

where, as for the MRI, we ignored the marginal terms in our definitions of $\widehat{\theta}^{RI}$ and $\widehat{\theta}_0^{RI}$ that cancel in the ARI. We now claim the following proposition.

Proposition 3.4.1. *Under the multinomial model we have*

$$\mathbb{E}(ARI) = \mathbb{E}(\hat{\theta}^{RI}) - \mathbb{E}(\hat{\theta}_0^{RI}),$$

with

$$\mathbb{E}(\hat{\theta}^{RI}) = 2\sum_{k\ell}^{KL} \pi_{k\ell}^2 \text{ and}$$

$$\mathbb{E}(\hat{\theta}_0^{RI}) = \frac{2}{\binom{n}{2}^2}\left[\binom{n}{2}\sum_{k\ell}^{KL}\pi_{k\ell}^2 + n(n-1)(n-2)\sum_{k\ell}^{KL}\pi_{k\ell}\pi_{k.}\pi_{.\ell} + 6\binom{n}{4}\sum_{k\ell}^{KL}\pi_{k.}^2\pi_{.\ell}^2\right]$$

Assuming we are under the null this simplifies so that $\mathbb{E}_{\mathcal{H}_0}(ARI) = 0$.

Proof. Using Lemma 3.2.1, we have

$$\mathbb{E}(\sum_{\mathcal{P}} C_{ij}^1 C_{ij}^2) = \binom{n}{2}\sum_{k\ell}^{KL}\pi_{k\ell}^2.$$

Using Definition (3.13) and Lemmas 3.2.1, 3.2.2, 3.2.3 we obtain

$$\mathbb{E}(\sum_{\mathcal{P}} C_{ij}^1 \sum_{\mathcal{P}} C_{ij}^2) = \binom{n}{2}\sum_{k\ell}^{KL}\pi_{k\ell}^2 + n(n-1)(n-2)\sum_{k\ell}^{KL}\pi_{k\ell}\pi_{k.}\pi_{.\ell} + 6\binom{n}{4}\sum_{k\ell}^{KL}\pi_{k.}^2\pi_{.\ell}^2.$$

Under the null we have $\pi_{k\ell}^2 = \pi_{k.}^2\pi_{.\ell}^2$ and we get

$$\mathbb{E}_{\mathcal{H}_0}(\sum_{\mathcal{P}} C_{ij}^1 C_{ij}^2) = \binom{n}{2}\sum_{k\ell}^{KL}\pi_{k.}^2\pi_{.\ell}^2$$

$$\begin{aligned}\mathbb{E}_{\mathcal{H}_0}(\sum_{\mathcal{P}} C_{ij}^1 \sum_{\mathcal{P}} C_{ij}^2) &= \sum_{k\ell}^{KL}\pi_{k.}^2\pi_{.\ell}^2\left[\binom{n}{2} + n(n-1)(n-2) + 6\binom{n}{4}\right] \\ &= \binom{n}{2}^2\sum_{k\ell}^{KL}\pi_{k.}^2\pi_{.\ell}^2.\end{aligned}$$

The expectations $\mathbb{E}(\hat{\theta}^{RI})$ and $\mathbb{E}(\hat{\theta}_0^{RI})$ are obtained by scaling respectively with $2/\binom{n}{2}$ and $2/\binom{n}{2}^2$; $\mathbb{E}(ARI)$ is their difference.

□

From these results we conclude that Hubert and Arabie's ARI is biased under the multinomial model in general, since the term used for the adjustment is biased as $\mathbb{E}(\hat{\theta}_0^{RI}) \neq \theta_0^{RI}$. Note, however, that this estimator is not biased under the null $\mathcal{H}_0$.

3.4.2. Study of the bias Hubert and Arabie 's *ARI*

The quantity that we study in this section is

$$\begin{aligned}\text{bias}_n(\theta_0^{RI}) &= \theta_0^{RI} - \mathbb{E}(\widehat{\theta}_0^{RI}) \\ &= \sum_{k,\ell}^{K,L} \pi_{k.}^2 \pi_{.\ell}^2 - \left[\binom{n}{2} \sum_{kl}^{K,L} \pi_{k\ell}^2 + 6\binom{n}{3} \sum_{k\ell}^{K,L} \pi_{k\ell}\pi_{.\ell}\pi_{k.} + 6\binom{n}{4} \sum_{k\ell}^{K,L} \pi_{k.}^2 \pi_{.\ell}^2\right] \Big/ \binom{n}{2}^2\end{aligned}$$

Bias disappear when n goes to infinity. The bias can be rewritten as

$$\text{bias}_n(\theta_0^{RI}) = \frac{4n-6}{n(n-1)} \sum_{k,\ell}^{K,L} \pi_{k.}^2 \pi_{.\ell}^2 - \frac{2}{n(n-1)} \sum_{kl}^{K,L} \pi_{k\ell}^2 - \frac{4(n-2)}{n(n-1)} \sum_{k\ell}^{K,L} \pi_{k\ell}\pi_{.\ell}\pi_{k.}$$

From this expression we get

Lemma 3.4.1.

$$|bias_n(\theta_0^{RI})| \leq \frac{8}{n}$$

$$|bias_n(\theta_0^{RI})| = O(1/n), \quad \textit{and} \quad \lim_{n\to+\infty} bias_n(\theta_0^{RI}) = 0.$$

Proof. As seen in Equation (3.4.2), the bias consist of three terms. The absolute value of the sum of these three terms is bounded by the sum of their absolute values. Then, using that $\sum_{k,\ell} \pi_{k\ell} = 1$ and all $\pi_{k\ell} \geq 0$, we bound $\sum_{k,\ell} \pi_{k.}^2 \pi_{.\ell}^2$, $\sum_{kl} \pi_{k\ell}^2$ and $\sum_{k\ell} \pi_{k\ell}\pi_{.\ell}\pi_{k.}$ by 1 and we get $|\text{bias}_n(\theta_0^{RI})| \leq \frac{4(2n-3)}{n(n-1)}$. We have, $(2n-3) < 2(n-1)$ and the result follows.

□

Empirical bias. In the case of independence the bias is zero. In the case of dependence, using Lemma 3.4.1 we get that the bias is smaller than 0.04 for n larger than 200. Following the work of Steinley and Brusco (2018), we study the importance of the difference empirically for small value of n in the next paragraph. In summary for n larger than 64 we observe a small bias, typically smaller than 10^{-2}. For smaller values of n the bias can be larger.

Simulation setting. We study the evolution of the bias by comparing two classifications with equal number of groups ($K = L$), with values varying in $K \in \{2, 4, 8, 16, 32, 64, 128\}$ and a growing number of individuals. For drawing the two compared classifications under the multinomial model, see Table 3.2.

We consider three scenarios described below where we tune the level of difficulty by controlling the balance between group sizes with the parameters ϵ.

Scenario 1. In the first scenario we investigate a π_{kl} distribution with a disproportionate diagonal. All other entries being null.

$$\pi_{k\ell} = \begin{pmatrix} 1-\epsilon & 0 & \cdots & 0 \\ 0 & \frac{\epsilon}{K-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\epsilon}{K-1} \end{pmatrix}$$

Scenario 2. In the second scenario we investigate a π_{kl} distribution with a proportional diagonal and extra diagonal dependency. All other entries being null.

$$\pi_{k\ell} = \begin{pmatrix} (1-\epsilon)/K & \epsilon/K & \cdots & 0 \\ 0 & (1-\epsilon)/K & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon/K & 0 & \cdots & (1-\epsilon)/K \end{pmatrix}$$

Scenario 3. In the third scenario we investigate a π_{kl} distribution with one line and one column being disproportional and all other entries being null.

$$\pi_{k\ell} = \begin{pmatrix} 1-\epsilon & \frac{\epsilon}{K+L-2} & \cdots & \frac{\epsilon}{K+L-2} \\ \frac{\epsilon}{K+L-2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\epsilon}{K+L-2} & 0 & \cdots & 0 \end{pmatrix}$$

Results. The results are shown in Figure 3.2 where the bias is shown in its absolute value with $\log_2 / \log_{10}$ scales. For the different scenarios, the parameter of imbalanceness ϵ,is fixed to 0.3 and 0.8.

In the different scenarios, the bias remains moderates for most values of K and n. When the number of individuals is small however, the difference turns to be more important and using the $(A)RI$ lead to misguiding conclusions.

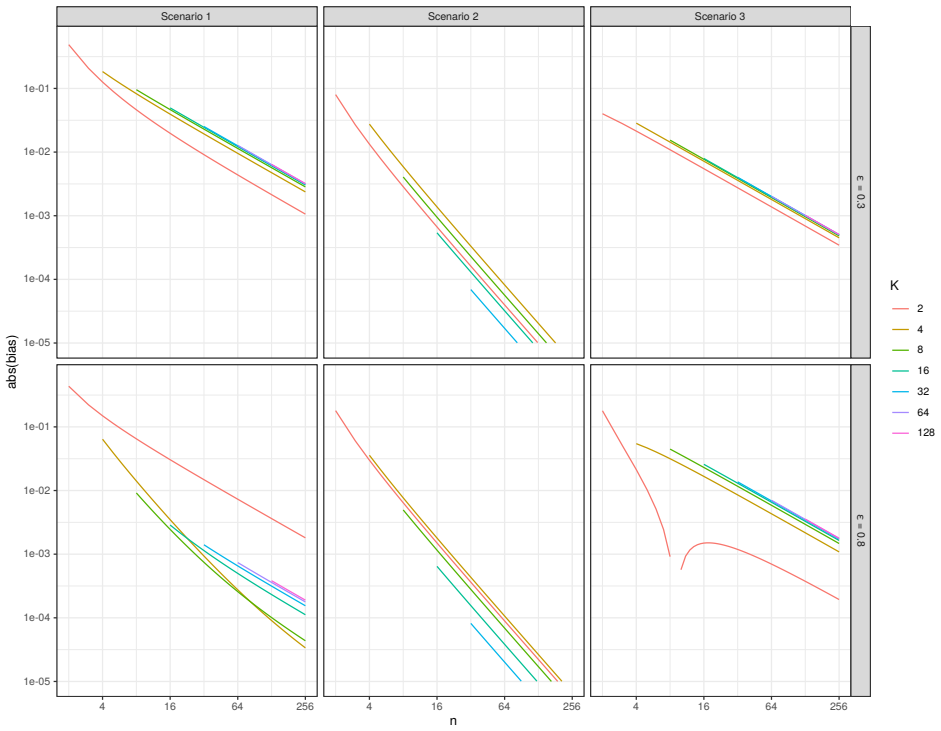


Figure 3.2 – Hubert and Arabie's ARI bias for different scenarios of $\pi_{k\ell}$-distribution

3.5. Conclusion

As a conclusion, we argue that one should always prefer our $M(A)RI$ to the $(A)RI$. There are four main reasons for this.

- The adjustment of the RI is based on a hypergeometric distribution which is unsatisfying from a modeling perspective. In particular, it forces the size of the clusters to be the same and it ignores randomness of the sampling (see the introduction). A multinomial model of the $MARI$ does not force the size of the clusters and properly model randomness. Furthermore, the model easily extends to the dependant case.

- The difference between the ARI and $MARI$ can be large for small n but essentially vanish for large n (see Section 3.4.2).

- The $M(A)RI$ can be computed just as fast as the $(A)RI$ in only $O(n)$ rather than $O(n + KL)$ using our `aricode` package.

- The $M(A)RI$ does not take into account pairs coherent by difference which – as argued in Section 3.2.1 – unnecessarily complexify the analysis and interpretation of the $(A)RI$.

Chapter 4

Cluster Stability for class discovery

Résumé. La stabilité des clusters pour la découverte des classes est devenue particulièrement populaire en génomique et en oncologie. Son principe est de sélectionner le nombre de groupes donnant lieu à la classification la plus stable en effectuant, par exemple, du sous-échantillonnage. Malgré sa popularité, on sait encore peu de choses sur la façon et sous quelles conditions dont cette méthode fonctionne en pratique. Afin d'étudier cette question, j'ai mis en place un package R `clustRstab` qui permet de mesurer la stabilité de clusters de manière flexible et unifiée. Par rapport aux autres packages R, il est avantageux, car il permet (1) de combiner différents paramètres algorithmiques afin de calculer la stabilité, (2) d'ajuster les critères de stabilité pour le hasard (3) de générer un grand nombre de répétitions (*nism*) dans un laps de temps raisonnable et (4) de générer la moyenne et l'écart-type comme sorties. Ce package R m'a permis de réaliser une simulation et une étude d'application. Dans l'étude de simulation, j'ai testé, (1) si nous avions accès à un grand nombre de jeu de données, le regroupement le plus stable corresponderait au nombre correct de groupes et (2) si nous sommes capables d'estimer cette stabilité lorsque, comme dans la pratique, nous n'avons accès qu'à un seul jeu de données. Dans l'étude d'application, j'ai appliqué cette méthode au jeu de données de NCI60, composé de lignées cellulaires de 9 types de cancer différents, et j'ai examiné si nous sommes en mesure de récupérer ces 9 types de cancer en utilisant la stabilité de clusters. Les résultats de ces deux études montrent que (1) la stabilité des clusters ne permet pas toujours d'estimer correctement le nombre de groupes, et ce, même dans des configurations de données simples, (2) nous ne sommes pas toujours en mesure d'estimer la stabilité, (3) le clustering le plus stable n'est pas toujours le plus intéressant ou celui qui révèle le plus d'informations sur la structure des données et enfin (4) l'estimation de la stabilité des clusters dépend du paramétrage. Pour toutes ces raisons, la stabilité des clusters, même si elle est une caractéristique intéressante pour un

clustering, doit être utilisée avec prudence. Par exemple, en la testant avec différents réglages de paramètres algorithmiques, ce qui est possible avec le paquet `clustRstab`, ainsi qu'en la combinant avec d'autres critères de validation de clusterings.

Abstract. Cluster stability for class discovery has become particularly popular in genomics and oncology. Despite its popularity, little is still known about how or under which conditions this method works in practice. In order to investigate this question, I implemented an R package `clustRstab` that allows to measure cluster stability in a flexible manner for the task of selecting the number of groups in unsupervised clustering. Compared to other R packages, it is advantageous since it allows to (1) combine different algorithmic parameters in order to compute the stability, (2) adjust the stability criteria for chance, (3) generate a large number of repetitions (*nism*) in a reasonable amount of time and (4) generates both the mean and the standard deviation as outputs. This R package allowed me to conduct a simulation and an application study. In the simulation study I tested whether, (1) if we had access to a large number of datasets, the most stable clustering corresponds to the correct number of groups and (2) whether we are able to estimate this stability when we, as in practice, only have access to a single dataset. In the application study, I applied this method to the NCI60 cancer dataset consisting of cell lines from nine different types of cancer and investigated whether we are able to recover these nine cancer types by using cluster stability. The results of these two studies show that (1) cluster stability does not always correctly estimate the number of groups, and this, even in simple data settings, (2) we are not always able to estimate the stability, (3) the most stable clustering is not always the most interesting or the one that reveals the most information of the data structure and finally (4) the estimation of cluster stability is parameter dependent. For all these reasons, cluster stability, even though it is an interesting characteristic for a clustering, should be used with caution. For example, by testing it with different algorithmic parameter settings, which is possible with the `clustRstab` package, as well as combining it with other clustering validation criteria.

4.1. Introduction

Unsupervised clustering is an important part of data exploration and can reveal important structure or patterns present in the dataset. The particularity with unsupervised clustering is that there is no ground truth to which the results can be compared. Therefore, the number of groups, K, needs to be estimated. This is a difficult task and an open question in statistics. Also, without any external information available, the choice of K can easily become somewhat subjective. In order to make this choice more objective, several clustering validation criteria have been developed.

One of these methods that has become particularly popular in genomics and oncology is cluster stability. The basic principle of this method is that a clustering should be stable, *i.e.* reproducible on similar datasets and robust to changes such as subsampling Ben-Hur et al. (2001); Ben-Hur and Guyon (2003); Von Luxburg et al. (2010). To measure the stability of a clustering, ideally, one would like to have access to a large number of datasets upon the same n observations and p variables that one could cluster and then compare. However, since the data acquisition procedures are very long, tedious and expensive, especially in biology, in practice we will only have access to one dataset. The stability is therefore estimated from this initial dataset. The most common manner to do so is to perturb the initial dataset, for example by sampling observations or variables, in order to obtain a large number of perturbed datasets. These perturbed datasets are then clustered into a given number of groups, and the obtained clusterings are compared. The more similar these compared clusterings are, the more the initial clustering is considered as stable. This procedure is repeated for several numbers of groups and K can then be selected as the one giving the most stable clustering.

Cluster stability and the idea of subsampling is particularly adapted to genomics since genes are often highly correlated with each other (Ben-Hur and Guyon, 2003). Also, the method is supported by some promising theoretical works, developed, among others, by Ben-David et al. (2006, 2007); Bubeck et al. (2009); Shamir and Tishby (2008a,b, 2009). These works indicate that, at least in some idealized settings, a clustering is stable when the number of groups K of a clustering corresponds to the true number $K^\star$ of groups in the data, and is unstable when $K \neq K^\star$. Yet, these results have all been derived under the assumption of an infinite number of observations, that is, $n \to \infty$, and has not yet been extended to the finite case.

Hence, despite the popularity of cluster stability, little is still known about how or under which conditions this method works in practice. Also, the wish for stability can be questionable, and its pertinence for selecting the number of groups might depend on the context. On one hand, being stable is a good property for a clustering and one would not like to have a clustering that is unstable. On the other hand, the most stable clustering for a dataset occurs for $K = 1$, *i.e.* when all observations are in the same cluster. Also, in a simulation study conducted by Şenbabaoğlu et al. (2014) the authors investigated the validity of *Consensus Clustering* (Monti et al., 2003), a clustering method based on stability. They found that this method could (1) cluster randomly generated unimodal data into stable clusters and that (2) it poorly identified the correct number of groups in data with known structure. Hence, a stable clustering will not necessarily be interesting and the correct number of groups is not always detected by the "most stable" clustering.

Moreover, the estimation of the stability of a clustering depends on a large number of parameters such as the data perturbation method, the clustering algorithm, the cluster comparison strategy *etc.* One could imagine that different parameter settings could yield to different stability estimations. An example of this is found in a comparison study conducted by Rozmus (2017). The author showed that different cluster stability methods led to different answers in the question of the "correct" number of groups, indicating parameter dependence of this method.

In order to get a better understanding of when and under which conditions this method works, we are, in the present work, going to test this method in different parameter settings.

First, a simulation study, we are going to simulate the "true" and the estimated stability of a clustering. The aim of this study is to find out whether, if we had access to the "true" stability of a clustering, (1) the most stable clustering would correspond to the true number of groups and if (2) we are able to estimate this "true" stability correctly. This simulation study is motivated by the theoretical results derived by Ben-David et al. (2006, 2007) and cluster stability is going to be presented in a statistical setting based on the work of Von Luxburg et al. (2010).

Second, in an application study, this method is going to be illustrated upon

the NCI60 cancer dataset (Ross et al., 2000) and the results are going to be compared to other cluster stability methods as well as to other clustering validation criteria. The aim of this section is to (1) test how different parameters impact the estimation of cluster stability, (2) find some of the necessary parameter settings for correctly interpreting the stability and (3) investigate how cluster stability articulate with other clustering validation methods.

To conduct these two studies we implemented an R package `clustRstab` allowing to estimate the stability of a clustering in a flexible manner. This package is going to be presented and compared to some other methods for estimating cluster stability before presenting these two studies.

4.2. clustRstab: an R package for flexible estimation of cluster stability for class discovery

In this section we are going to present the `clustRstab` package. To do so, we will first present its general pipeline with its different algorithmic steps. The different possible parameter settings will then be presented in detail. Finally, we will compare the package to other R packages estimating cluster stability.

4.2.1. The global structure of the `clustRstab` R-package

The `clustRstab` package allows to measure the stability of a clustering in a flexible manner with the aim to select the number of groups in unsupervised clustering. It consists of a major function, `clustRstab`, that takes a dataset as input, perturbs and cluster it in different number of groups (K) and gives as output a stability measure (mean, standard deviation, minimum and maximum) for each K. Its estimation of cluster stability, is based on the generic cluster stability algorithm as presented in Algorithm 4 and its general steps can be described as follows:

1. Generate a large number of perturbed datasets from the initial dataset by, for example, subsampling variables or observations.

2. Cluster each perturbed dataset using a given cluster algorithm.

3. Compare the obtained clusterings using:

(a) A given comparison strategy, for example comparing all or some of the obtained clusterings

(b) A given cluster comparison score *e.g* the $MARI$ of Sundqvist et al. (2020a)

The stability of the clustering is then estimated as the arithmetic mean of these comparisons.

Several variants of cluster stability have been implemented, for example, the R packages `clv` (Nieweglowski, 2020), `clusterStab` (MacDonald et al., 2018), `ClusterStability` (Lord et al., 2016), `ConsensusClusterPlus` (Wilkerson et al., 2010) `fcp` (Hennig, 2020). Even though these packages are often presented with a promise of novelty, in reality they often boils down to a specific setting of these three steps. Hence, the advantage of the `clustRstab` package is that all these different parameter setting can be tested with the same function.

The `R`-code for the `clustRstab` function, with its most important arguments and their default values, are printed below with

```
clustRstab(data,
           perturbedDataFun = subSample,
           nsim = 500,
           clAlgo = clAlgoKmeans,
           kVec = 2:15,
           clCompScore = MARI,
           typeOfComp = "all",
           plot = TRUE)
```

where `data` is the initial dataset, `perturbedData` is the data perturbation function, `nsim` is the number of perturbed datasets that should be generated, `clAlgo` is the clustering algorithm to be used, `kVec` is the vector of the number of groups K to be tested, `clCompScore` is the cluster comparison score, `typeOfComp` the type of comparison that should be conducted and finally `plot`, a Boolean indicating whether a plot of the results should be printed. Hence, this function allows to easily test different parameter settings in order to evaluate their impact on a given dataset. This avoids to use different R packages and functions since their parameter settings can be implemented by this function. The general pipeline of the `clustRstab` function is presented in Figure 4.1.

As can be seen in this pipeline, the `clustRstab` function takes as input a dataset $\mathbb{X} \in \mathbb{R}^{n \times p}$, where n is the number of observations and p the number of variables. Then, $nsim$ perturbed datasets, $D_1, \ldots, D_{nism}$, are genereated using the `dataPerturbFun` and consisting of n' observations and p' variables. If the `dataPerturbFun = subSample,` then n' and p' depend on the proportions of subsampled items otherwise, if the data is perturbed by adding noise or randomly projected in a smaller dimension, then, $n' = n$ and $p' = p$. Each perturbed dataset is then clustered into `K=kVec` different number groups with `kVec` $= [K_{min}, \ldots, K_{max}]$ using the cluster algorithm `clAlgo`. This results in a cluster label matrix where the cluster labels (group belongings) for each observation are given for the different `kVec` clusterings. For computational ease, the labels for each K are then regrouped into $n' \times nsim$ matrices. This allows to easily compare the clusterings for each K using the `clCompScore`. For each value of K, $nComb$ comparisons will be conducted and as will be seen later, $nComp$ depends on the type of comparison (`typeOfComp`). The mean-value, the standard deviation, the minimum and the maximum values of the `clCompScore` are then extracted for each K and given as an output. An example of a resulting `clustRstab` plot is found in Figure 4.2.

For computational efficiency, the `clustRstab` function is parallellized using the `mc.lapply` function from the R-package `parallel` (R Core Team, 2018). The user can fix the number of cores that should be used to parallellize the calculations, `mc.cores`, as an argument of the `clustRstab` function, by default it is fixed to `mc.cores=2`.

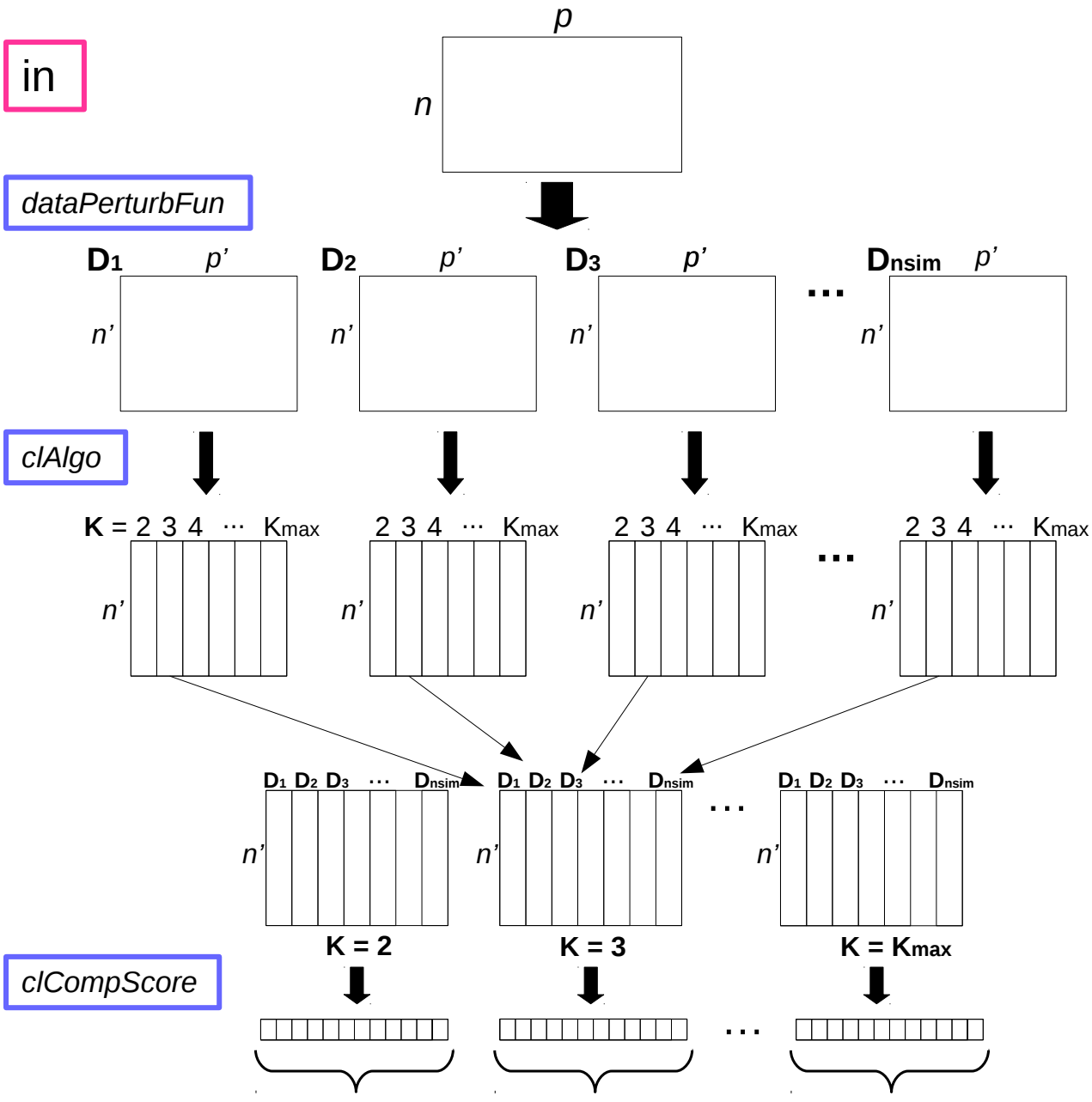


Figure 4.1 – Pipeline for the `clustRstab` function

4.2.2. Details of the `clustRstab` R-package

I will now present the steps of this algorithm and their different parameter settings.

perturbedDataFun: generate *nsim* perturbed versions of the dataset

The most crucial step when estimating the stability of a clustering is to perturb the initial dataset. This can be done in different manners whereof the most popular is subsampling of observations and/or variables. Two other perturbation strategies are adding random noise (Möller and Radke, 2006) and randomly projecting the data in a low dimensional space if the data is high dimensional (Smolkin and Ghosh, 2003). These three perturbing strategies are implemented in the `clustRstab` package. For each of these functions, the user can tune the level of perturbation directly in the `clustRstab` function with the arguments described below.

- Subsampling observations and/or variables:

  ```
  subSample(data,
      nProp = 0.7,
      pProp = 0.9, ...)
  ```

  In the `clustRstab` function the default proportions of subsambled observations is set to `nProp = 0.7` and the proportion of subsampled variables `pProp = 0.9`. These values are arbitrary and the user is encouraged to test different values in order to see what impact this might have on the results.

- Adding Gaussian noise:

  ```
  noiseGaussian(data,
          noiseGaussianMean = 0,
          noiseGaussianSD = 1, ...)
  ```

  We implemented a function making it possible for the user to add Gaussian noise to each datapoint in the dataset. The parameters that need to be fixed are the mean-value and the standard deviation of Gaussian distribution of the added noise. In the `clustRstab` package, their default values are, `noiseGaussianMean = 0` and `noiseGaussianSD = 1`. Again these are arbitrary values.

- Projecting in a random smaller space:

```
randProjData(data,
    randProjDim = 100 ,
    randProjMethod = "Haar", ...)
```

We implemented a function making it possible to randomly project the data into a lower number of dimensions. To do so, we use the `RPGenerate` function from the `RPEnsemble` package (Cannings and Samworth, 2017). This function generates a random p by `randProjDim` matrices according to Haar measure, Gaussian or axis-aligned projections, where p is the number of dimensions in the initial dataset (number of variables) and `randProjDim` is the number of dimensions in which the dataset should be projected. This later value is set to `randProjDim = 100` by default in the `clustRstab` package and the default projection method is the Haar measure `randProjMethod = "Haar"`. The Haar measure assigns an "invariant volume" to subsets of locally compact topological groups, consequently defining an integral for functions on those groups. However, the user can select Gaussian or axis-aligned projections, see Cannings and Samworth (2015) for further explications.

Von Luxburg et al. (2010) pointed out that there is a fine balance between perturbing the dataset sufficiently enough in order to obtain different clusterings, but not perturbing it too much which would destroy the signal of interest present in the dataset. Doing so might not be evident and depend probably on the dataset in question.

clAlgo: Cluster the perturbed datasets

The next step for measuring the stability of a clustering is to cluster all the obtained perturbed versions of the datasets in `kVec = Kmin,..., Kmax` groups.This can be done by using any clustering algorithm. We implemented some of the most popular in the `clustRstab` package, that is, the k-means, hierarchical ascendant clustering and the GMM.

- The k-means algorithm

```
clAlgo = clAlgoKmeans
```

The k-means algorithm is implemented with the `kmeans` function from the `stats` package (R Core Team, 2019a) and in order to avoid that the algorithm gets "stuck" in local minimums (suboptimal solutions) we conduct 10 initializations for each "run".

- Hierarchical ascendant clustering

  clAlgo = clAlgoHCWard
  clAlgo = clAlgoHCComplete

  In the HAC algorithm, each observation is considered as a distinct class. The algorithm will then regroup the two closest classes and repeat this until all observations constitute one class. The HAC algorithm is implemented with the `hclust` function from the R-package `stats` (R Core Team, 2019a) with an *euclidean* distance. There are different methods to compute the distance between clusters. For the moment we implemented the Ward and Complete methods, see Section 1.5 for more information.

- The Gaussian Mixture model

  clAlgo = clAlgoGmmEII
  clAlgo = clAlgoGmmEEI

  The Gausian Mixture Models (GMM) is conducting clustering in a probabilistic setting and aims to fit K Gaussian distributions upon the n observations. This is done using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). We implemented the GMM in the `clustRstab` package by the `Mclust` function from the `mclust` R-package (Scrucca et al., 2016a) for different covariance models such as diagonal or spherical distribution for equal volume and shape (EEI vs EII). These are the covariance models with the least number of parameters to be adjusted, hence, with the highest probability to converge when K is large.

Also, the user have access to these different functions and can therefore easily define their own clustering algorithm.

typeOfComp: Type of comparison

Once the clusterings are obtained the next step is to measure their similarity or distance. Let us define $\mathcal{S}_K$ the stability score of a clustering of K groups.

This score is be defined in its statistical setting in Section 4.2.4 and see Definiton (4.5). The estimated stability index $\widehat{\mathcal{S}}_K$, defined in (4.8), is then computed for each K as the mean value of the cluster comparison score and will hence depend on the type of comparison conducted. Different strategies can be employed when comparing the clusterings and we implemented three of them in the `clustRstab` package.

- Pairwise comparisons between clusterings of perturbed datasets

  The argument `typeOfComp = "all"` allows to compute pairwise comparisons between all the $nsim$ obtained clusterings from the perturbed datasets. Computationally this is quite costume since there are $ncomb = \binom{nsim}{2}$ pairwise comparisons to compute, the cluster stability index $\widehat{\mathcal{S}}_K$ is then computed as:

$$\widehat{\mathcal{S}}_K = \frac{1}{\binom{nsim}{2}} \sum_{d<d'}^{\binom{nsim}{2}} \widehat{Sc}(\widehat{C}(\mathbb{X}'_d),\widehat{C}(\mathbb{X}'_{d'})), \tag{4.1}$$

  where $\widehat{Sc}()$ is the cluster comparison score, $\widehat{C}()$ is the cluster algorithm, $\mathbb{X}'_d$ and $\mathbb{X}'_d$ are two different perturbed datasets with $d < d'$ and $d = 1, \ldots, nsim$ of the same initial dataset $\mathbb{X} \in \mathbb{R}^{n\times p}$.

- A randomly selected subsample of the pairwise comparisons between clusterings of perturbed datasets The argument `typeOfComp = "random"` allows to decrease the number of comparisons by only comparing a random set selected of clusterings. The cluster stability index $\widehat{\mathcal{S}}_K$ is then computed as:

$$\widehat{\mathcal{S}}_K = \frac{1}{|sComp|} \sum_{d<d'\in sComp}^{|sComp|} \widehat{Sc}(\widehat{C}(\mathbb{X}'_d),\widehat{C}(\mathbb{X}'_{d'})), \tag{4.2}$$

  where $sComp$ is a sampled proportion of all the possible comparisons between the different $d < d'$ perturbed datasets.

- Compare each obtained clustering from the perturbed data with the initial dataset `typeOfComp = "toInitial"` allows to compare the clusterings obtained from the perturbed datasets with the clustering of the

original data. There are then only $ncomb = nsim$ comparisons to compute, largely decreasing the computationally efforts of the algorithm. This sampling method is proposed for example in Levine and Domany (2001) and the cluster stability index $\widehat{\mathcal{S}}_K$ is then computed as:

$$\widehat{\mathcal{S}}_K = \frac{1}{nsim} \sum_{d=1}^{nsim} \widehat{Sc}(\widehat{C}(\mathbb{X}),\widehat{C}(\mathbb{X}'_d)), \tag{4.3}$$

where $\mathbb{X}$ is the initial dataset.

clCompScore: Cluster comparison score

To compare the clusterings, one need a clustering comparison score. A multitude of scores exist for this purpose and most of them are somewhat counting agreeing or non agreeing pairs of observations from the contingency table of two clusterings (this will be seen later on). The advantage with the `clustRstab` function is that the user can put any given score as `clCompScore` argument to the function. It only needs that the score takes two clusterings as an input and gives a similarity or distance as output. We recommend the user to use the $MARI$ or the NID scores implemented in the `aricode` package (Chiquet et al., 2020) for their statistical properties and their efficient computational implementation. Indeed, the space and time complexity of the computation of these scores is linear in the number of samples and importantly does not depend on the number of clusters as it is not explicitly computed from the contingency table. However, it should be noticed that cluster comparison scores depend on the number of groups of the two clustering. Indeed, the larger the number of group is, the more information (or the more similar) will the two clusterings share (be) Morey and Agresti (1984); Vinh et al. (2010); Von Luxburg and Ben-David (2005). Some few scores such as the ARI or the $MARI$ are intrinsically corrected for chance, whereas others, such as the NID or the Jaccard coefficient are not.

In the Section 1.5.4 I conducted a simulation experiment investigating how the ARI and NID depend on the number of groups. The result of this experiment stresses out two things, (1) the cluster comparison scores need to be corrected for chance and (2) as already stated, the variance (or standard deviation) needs to be taken into account when interpreting the stability of a clustering. It is rarely the case that cluster comparison scores, as for the NID, are intrinsically corrected for chance. We therefore propose a permutation

strategy to estimate the scores for the numbers of groups in the `clustRstab` package. To do so, the user needs to set, `baseLineCorrection=TRUE` as an argument in the `clustRstab` function. A "baseline" score for each comparison between two clusterings will then be computed by permuting all the cluster labels of one of the clusterings and compare it to the other. This permutation comparison is repeated $\kappa = 10$ times, and the obtained mean-value is then used to scale the initial cluster comparison score. This process will obviously make the computational effort longer, but it is necessary if one wants to use a non corrected score and compare the stability for different numbers of K. If `typeOfComp = "toInitial"` and `baseLineCorrection=TRUE`, the permuted stability score $\widehat{\mathcal{S}}_{K_{perm.}}$ is then be computed as follows

$$\widehat{\mathcal{S}}_{K_{perm.}} = \frac{1}{nsim} \sum_{d=1}^{nsim} \frac{\widehat{Sc}(\widehat{C}(\mathbb{X}),\widehat{C}(\mathbb{X}'_d))}{\frac{1}{\kappa}\sum_1^\kappa \widehat{Sc}(\widehat{C}(\mathbb{X}),perm(\widehat{C}(\mathbb{X}'_d)))}, \tag{4.4}$$

where $perm()$ is the function of permuting the labels of a clustering.

Select the number of groups

The number of groups can now be selected as the one optimizing the stability index $\widehat{\mathcal{S}}_K$. For this, the `clustRstab` function gives as output the mean stability value but also the standard deviation, the minimum and the maximum for each K. If `plot = TRUE`, a `clustRstab` plot will be printed. In Figure 4.2 an example of such a plot is printed for the NCI60 dataset that will be presented in Section 4.2.5. The stability of the NCI60 dataset is maximized for $K = 3$.

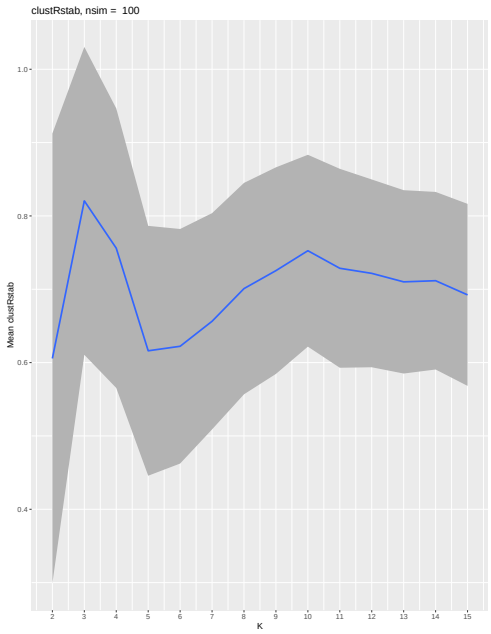


Figure 4.2 – Example of `clustRplot`. The blue curve corresponds to the mean and the gray shaded area to the SD of the $MARI$ score (Sundqvist et al., 2020a) for the NCI60 dataset with `perturbedDataFun = subSample`, `clAlgo = clAlgoKmeans`, `typeOf Comp = "all"`. The $MARI$ score is a similarity score and should therefore be maximized.

Computational time

The time to run the `clustRstab` function depends on some of the choices for the cluster stability algorithm. Particularly, it depends on the number of groups K tested, the number of simulations $nsim$ and the type of comparisons conducted. The calculations will be long if `typeOfComp = "all"` since it will then compute $\binom{nsim}{2}$ comparisons for each K. Also, if one adds the `baseLineCorrection = TRUE` this will also make the calculations longer. However, we still advice the user to fix the number of simulations (the number of generated perturbed datasets) $nsim > 100$ in order for the result to be as robust as possible.

4.2.3. Comparison with other R-packages for cluster stability

In this section we will present different R-packages that all allow to estimate the stability of a clustering. It exists a large variety of R-packages allowing

to estimate the stability of a clustering. As already mentioned, they all follow the same general pipeline as presenter for the `clustRstab` package, but with the particularity of fixing some of the parameters. Yet, it is important to see how these parameters can be set and the different arguments. The following R-packages are going to be presented: `clusterStab`, `cvl`, `ClusterStability`and `Consensus Clustering`.

clv. The `clv` package (Nieweglowski, 2020) is based on the concept of stability proposed by Ben-Hur and Guyon (2003). That is, if a clustering is well representing the underlying structure of a dataset, the clustering should be robust to some small changes in the data. In this package, two functions allow to measure the stability of a clustering; `cls.stab.sim.ind` and `cls.stab.sim.opt`. We are only going to focus on the first one.
The `cls.stab.sim.ind` function measures the similarity between the clustering of the original dataset and the clustering obtained from the perturbed datasets.

Here, the data is perturbed by subsampling the observations (default proportion is set to 0.75) and clustered into different numbers of groups $K = K_{min}, \ldots, K_{max}$. These clusterings are then compared to the clustering of the original dataset using a similarity score. The user can choose between a large number of implemented similarity measures and cluterings algorithms and can as well use its own defined clustering algorithm with the `cls.stab.sim.ind.usr` function.

An interesting detail is that the user will be notified if there is a clustering with groups of singletons (single observations) which might help to avoid considering uninteresting clusterings. Moreover, this package contains a number of interesting functions characterizing a clustering. Also, with the this `cls.stab.sim.ind` function one can test several clustering algorithms and similarity scores within one run. The output is a list containing the results for each clustering method and similarity score, *e.g.* `results$kmeans$rand`, with the results for each comparison. The number of repetitions, $nsim = 99$, is the maximum value that this function can take. This might not be enough in certain settings. Also, it should be noted that the function changes automatically the number of simulations to 10 if it is set to larger than 99 without noticing the user.

clusterStab. This package is also based on the method of Ben-Hur and Guyon (2003) and allows to estimate the stability of clustering solutions using microarray data. The `benhur` function uses a similar method as described from the `clv` package (subsampling observations and comparing the obtained clusterings to the one of the original dataset) but only allow to use the HAC algorithm and the Jaccard index. The clustering can then be validated by computing the same clustering procedure with the `ClusterComp` function, but instead of subsampling the observations, the variables are subsampled. The authors argue that, if the clustering is stable, it is a "good" clustering that did not occur by chance. However, validating a clustering, that was defined by its stability, by assessing its stability do not seem to be of much of interest. Especially that one would hope that the stability of a clustering does not depend on the fact the one has subsampled the observations or variables. Also, basing the selection of the number of groups on a distribution of an index and not its moments (mean-value and variance) increase the difficulty of the task for the user, making this choice more prone to subjectivity.

ClusterStability. This package (Lord et al., 2016) implements the methods for computing the stability of a clustering and its observations proposed by Lord et al. (2017). These authors use the fact that the k-means and the k-medoids (similar to the k-means algorithm but defines datapoints as clustercenters) algorithm usually gets stuck in local optimums. The stability of a clustering is assessed by conducting a large number of clusterings upon the whole initial dataset varying the initialization of the cluster centers. Hence, the dataset itself will not be perturbed. Skipping this "perturbing" step makes the computational implementation efficient and they fix the number of runs by default to 1000. Measuring the stability of the observations of the dataset also allow to remove unstable items and hence, obtain an even more stable clustering. The assessment of the global and the individual stability of the clustering is done by the `ClusterStability` function which computes four different similarity indexes. A drawback of this function is that it only takes as input a single number of groups K and therefore demands some extra work if the user wants to compare the stability for different number of groups. Also, it only gives the similarity index but not its variance making the evaluation of the stability incomplete.

ConsensusClusterPlus. A particularly popular method for genomics and oncology research, is the *Consensus Clustering* method proposed by (Monti et al., 2003). Rather than computing an index score for each pair of clustering and average them, *Consensus Clustering* computes a matrix for each pair of clustering and averages those in a "consensus matrix". The cluster stability score is then derived from this matrix. The consensus matrix indicates whether two observations are classified together or not in a given clustering of the perturbed data, with, 1 if two observations are classified together, and 0 otherwise. This is done for each clustering obtained and the consensus matrix is computed as the sum of all "individual" matrices normalized by a matrix indication the number of times two observations have appeared together in a subsampled dataset. If the classification is perfectly stable, the consensus matrix would only consist of 0 and 1. Given the value of the consensus matrix, the cumulative distribution function is drawn and the area under the curve is computed. The number of groups is then selected as the value provoking the largest change in this area compared to the other number of groups. However, there is not a clear-cut criterion and the choice is prone for subjective interpretations.

Different R-packages computing cluster stability, conclusion. In the present section, we presented different R-packages allowing to measure the stability of a clustering. These methods led to different answers to the question of the most stable clustering and hence, the estimation of the "correct" number of groups for the data. This is not surprising since we have seen that the estimation of the stability of a clustering is parameter dependent, and that these methods use slightly different parameters settings. It should be noticed that some of these packages lack the possibility to correct the stability criteria for the number of groups as well not giving any dispersion index making the interpretation incomplete. Particularly, the results of the `clusterStab` and the `ConsensusClusterPlus` are given as distribution plots, making the decision of the "most" stable clustering even more prone to subjectivity.

Also, some of these packages mix the procedures of (1) proposing a clustering and (2) validating it. Indeed, in the `ClusterStability` and the `ConsensusClusterPlus` packages, some of the "unstable" observations are removed in order to obtain a more "stable" clustering. In this case, one is no longer measuring the stability of a clustering but basing the clustering on its stability. Obviously, if one removes the unstable observations of a clustering,

the clustering will become more stable and in this case also more probable to be "validated". Using an external cluster validation criterion, such as the Gap-statistic (Tibshirani et al., 2001) or an information criterion, then seem to be necessary for validation.

We therefore argue that our package `clustRstab` has its interest of its own since it allows to (1) adjust the stability criteria for chance, (2) generates a large number of repetitions ($nism$) in a reasonable amount of time, (3) gives both the mean and the standard deviation as outputs and (4) allows to combine different parameters in order to compute the stability.

4.2.4. Cluster stability and the data structure - A simulation study

In this simulation study we will investigate how cluster stability behaves in different data settings. The simulation study is motivated by the theoretical work of Ben-David et al. (2006, 2007) showing that the stability of the k-means algorithm depends on the number of global optimums, that is, whether there are a single or several optimums in the k-means objective function. This number is related to how the groups are distributed in a given data structure. In the "ideal case", where the algorithm always finds the global optimum, *i.e.* the optimal solution and we have access to an infinity of observations n, they show that, if there is a unique optimum, the clustering is stable. This will for example appear when the groups are distributed homogeneously in equidistant from each other and the number of groups in the clustering corresponds to the correct number of groups in the dataset. By simulations, we will test whether this result is observed in a more realistic case (n being finite) and with the actual k-means algorithm (that hence can get stuck in local optima). This will allow to investigate whether:

1. In the case that we had access to a large number of datasets and n being finite, would the most stable clustering correspond to the true number of groups?

2. If we only had access to one dataset, would we be able to estimate this stability correctly?

Theoretical motivation

Before conducting this simulation study we are going to present the theoretical results of Ben-David et al. (2006, 2007) as well as the k-means algorithm. In order for the reader to easily understand the presented results, we will first introduce some notations that will remain the same throughout this report. The notations are based on the work of Von Luxburg et al. (2010).

Definitions and the k-means algorithm

Notations. Let us consider a dataset $\mathbb{X}$, consisting of n observations, also referred to as datapoints in this section, and p variables, *e.g.* genes. The datapoints, $x_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$ are assumed to be drawn independently from some unknown underlying distribution P on some space $\mathcal{X}$. A *clustering* C is a function that takes the datapoints of $\mathbb{X}$ as inputs and assigns labels to them such that $C : \mathbb{X} \to \{1, \ldots, K\}$ which results in a partition of K clusters. Since in unsupervised clustering, the ground truth partition of the datapoints is unknown, a clustering algorithm, $\mathcal{A}$, will be used to obtain an estimation $\widehat{C}$ of C. The aim is that $\widehat{C}$ should be an, as good, representation of the underlying dataspace $\mathcal{X}$, as possible. Figure 4.3a. illustrates a clustering structure (dashed lined) that represents correctly the groups of the underlying dataspace $\mathcal{X}$ (the blue circles), whereas Figure 4.3c. illustrates a cluster structure that is not correctly representing $\mathcal{X}$.

The *stability index*, $\mathcal{S}_K$ of a clustering in K groups is then defined as the expected similarity, or distance, measured by a given *cluster comparison score* $Sc()$, between two clusterings upon the same observations with

$$\mathcal{S}_K := \mathbb{E}(Sc(\widehat{C}_K(\mathbb{X}_1), \widehat{C}_K(\mathbb{X}_2)) \tag{4.5}$$

where $\mathbb{X}_1$, $\mathbb{X}_2$ are two datasets of the same n observations and p variables drawn from the same underlying distribution P of the space $\mathcal{X}$ and K the number of clusters.

Remark: For the stake of simplicity, the presented theoretical results have been derived ignoring any effects linked to the sampling of datapoints. That is, it is assumed that we have access to an infinity of datapoints (that we do not have to sample) and that we can therefore work directly on the dataspace $\mathcal{X}$. Von Luxburg et al. (2010) referrers this to the limit case with $n \to \infty$. Also it is assumed that we have have access to the true minimal distance between

clusterings. That is, the distance obtained when any score or sampling effects are ignored.

The k-means algorithm. One of the most popular clustering algorithms is the k-means algorithm. This algorithm attempts to optimize the clustering objective function by minimizing the euclidean L^2-norm between each observation x_i, for $i = 1, \ldots, n$, and its closest barycenter (cluster mean), noted μ_k for $k = 1, \ldots, K$. Hence, the stake for the algorithm will be to "place" the barycenters in a manner that minimizes this distance.

For a given number of observations n and groups K the objective function $Q_K^{(n)}$ is defined as,

$$Q_K^{(n)}(\mu_1, \ldots, \mu_K) = \frac{1}{n} \sum_{i=1}^{n} \min_{k=1,\ldots K} \|x_i - \mu_k\|^2 \tag{4.6}$$

where $\mu_1, \ldots, \mu_K$ denote the barycenters for the K clusters. When considering directly the dataspace $\mathcal{X}$ and hence $n \to \infty$, we have

$$Q_K^{(\infty)}(\mu_1, \ldots, \mu_K) = \int_{\mathcal{X}} \min_{k=1,\ldots K} \|x - \mu_k\|^2 dP(x), \tag{4.7}$$

where P in the underlying probability distribution. Whereas the objective function $Q_K^{(n)}$, as defined in (4.6) depends on the n sampled datapoints, $Q_K^{(\infty)}$ can directly integrate on the probability distribution P of the underlying space $\mathcal{X}$, *i.e.* ignoring data sampling procedures. In this theoretical setting, Ben-David et al. (2006, 2007) assumed to have access to an infinity of n observations, hence considering directly the space $\mathcal{X}$ and working with the objective function $Q_K^{(\infty)}$ and not $Q_K^{(n)}$.

To minimize the distance between each observation and its closest barycenter, the k-means algorithm uses an iterative refinement technique described in Algorithm 3.

Algorithm 3 k-means

1: Initialization of K barycenter (cluster means) $\mu^0 = \{\mu_1^0, \ldots, \mu_K^0\}$
2: **for** each step t **do**
3: Assign each datapoint to its closest cluster center:
$\forall i = 1, \ldots, n : C^t(x_i) := \underset{k=1,\ldots,K}{argmin} \|x_i - \mu_k^t\|$
4: Given the clusters, refine the cluster means
$\forall k = 1, \ldots K : \mu_k^{t+1} := \frac{1}{n_k} \sum_{\{i|C^t(X_i)=k\}} x_i,$
where n_k denotes the number of datapoints in cluster k
5: **end for**
6: Stop when change in mean is smaller than a given ϵ: $\sum_{k=1}^{K} |\mu_k^t - \mu_k^{t-1}| < \epsilon$

The major drawback of the k-means algorithm is that the objective function $Q_K^{(n)}$ or $Q_K^{(\infty)}$ is not convex and hence the algorithm "gets stuck" in local optimums, *i.e.* giving a suboptimal result of $\widehat{C}$. As a consequence, the k-means algorithm is sensitive to the initialization of the μ_k barycenters.

Two scenarios will therefore be considered.

- In the *idealized scenario*, the 'oracle' k-means algorithm is considered, that is, the algorithm that will always find the global optimum, the "best" clustering solution (considering the minimization of $Q_K^{(\infty)}$) and never get stuck in local optimums.

- In the *realistic scenario*, the actual k-means algorithm is considered. This algorithm can hence get stuck in local minimums and the resulting clustering will therefore depend on the initialization of the cluster barycenters.

The stability index $\mathcal{S}_K$, will, in the following section, be based on a distance function and as consequence, a clustering will be stable when $\mathcal{S}_K = 0$ and unstable when $\mathcal{S}_K > 0$ (if the $\mathcal{S}_K$ would have been based on a similarity probability measure, it would have been stable when $\mathcal{S}_K = 1$ and unstable when $\mathcal{S}_K < 1$).

Ben-David et al. (2006, 2007) investigated the behavior of the idealized k-means algorithm when the objective function Q_K^{∞} has one or several global optimums and they show that,

1. If Q_K^∞ has a unique global minimum, then the idealized k-means algorithm is perfectly stable when $n \to \infty$, that is

$$\lim_{n\to\infty} \mathcal{S}_K = 0.$$

2. If Q_K^∞ has several global minima, then the idealized k-means algorithm is unstable, that is,

$$\lim_{n\to\infty} \mathcal{S}_K > 0.$$

These results do not depend on whether $K = K^\star$, with $K^\star$ being the correct number of groups. The question is then to know for which K there is a unique solution. Figure 4.3 illustrates different scenarios, or group settings, for which there are a unique solution, or not, depending on K. The different settings in this figure illustrate what we will refer to as symmetrical and unsymmetrical group settings. A symmetric group setting describes a data structure where the barycenters of the groups are equidistant from each other and an unsymmetrical group setting when this is not the case.

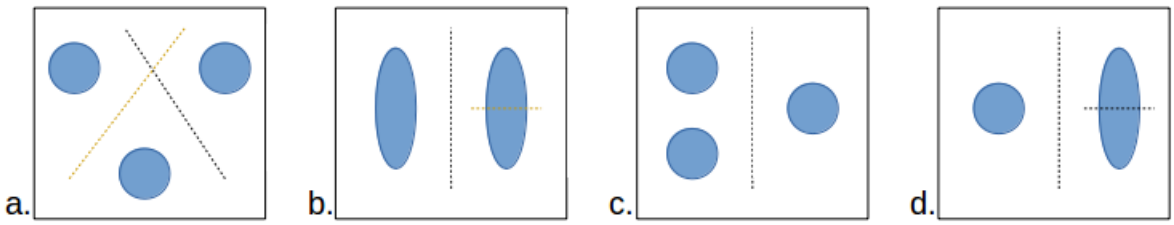


Figure 4.3 – Different data structures for clustering when $K \neq K^\star$ for symmetrical data structures in a. and b. and unsymmetrical data structures in c. and d.. The blue geometrical forms illustrate the data structures in the dataspace $\mathcal{X}$ and their number correspond to the true number of groups $K^\star$ and the dashed lines the cluster boundaries with the resulting K number of groups. The Figure is adapted from Ben-David et al. (2006).

These results can be interpreted as follows:

1. *When the group setting of the data structure is unsymmetrical, there will be a unique global optimum for all K. Hence, the cluster structure will be stable even when the number of groups of the clustering corresponds to the wrong number of groups, that is $K \neq K^\star$.*

This is illustrated in Figure 4.3c. and 4.3d., consisting of unsymmetrical group settings. Indeed, In Figure 4.3c., the data structure consists of $K^\star = 3$ groups but is clustered into $K = 2$ groups. Since the two groups to the left are much closer to each other than to the one to the right, the "best" solution is to separate the left groups from the right. Hence, there will be a unique optimum for $Q_K^{(\infty)}$. The clustering will therefore be stable even though $K < K^\star$. In Figure 4.3d.. there is only $K^\star = 2$ groups but when clustered into $K = 3$ groups the "best" solution is to split the larger group (and not the smaller) into two. Again there will be a unique optimum for $Q_K^{(\infty)}$ and the clustering will be stable even though $K < K^\star$.

2. *When the data structure is symmetrical, a clustering will be stable only when the number of groups corresponds to the true number of groups, that is $K = K^\star$ and unstable when $K < K^\star$ or $K > K^\star$.*

   This is illustrated in Figure 4.3a. and 4.3b., consisting of symmetrical group setting with $K^\star = 3$ respectively $K^\star = 2$ groups. In Figure 4.3a., the three groups are clustered into $K = 3$ by the two dashed cluster separators, there is a unique solution to do so, and the clustering will be stable. However, if we want to cluster this data into $K = 2$ groups, hence $K < K^\star$, there are three equal solutions, or global optima, for $Q_K^{(\infty)}$. Each of these three solutions regroups two of the groups and separates them from the third, and the clustering will therefore be unstable. Equally, if we want to cluster the data in Figure 4.3b., into $K = 3$ groups, hence $K > K^\star$, one of the groups will be split into two. This is illustrated by the yellow dashed line. Here, there are two solutions that are equivalent; the yellow separator can either split the group to the left or the one to the right with the same probability. As a result, there are two global minimums for $Q_K^{(\infty)}$ and the clustering is unstable.

As a conclusion, these results indicate that a clustering will be stable, or not, depending on whether $K = K^\star$ but also depending on the symmetrical structure of the group setting. Whereas the second result is encouraging, the clustering is only stable when it consists of the true number of groups ($K = K^\star$), the first is a bit concerning. Indeed, it states that a clustering will be stable even though $K \neq K^\star$ as long as it has a unique global optimum for the objective function. However, as pointed out by Von Luxburg et al. (2010),

this latter result is probably an artifact of the idealized k-means algorithm and hence, not extendable to the realistic case. Indeed, when considering the realistic k-means algorithm, the algorithm will get stuck in local optimums and hence not "reach" the global optimum. Instability will therefore be induced by the initialisation procedure of the cluster means even when there is an unique global optimum. Yet, this should be tested in a more realistic setting.

Also, these results (together with other non-presented results from, for example, Bubeck et al. (2009); Shamir and Tishby (2008a,b, 2009)) converge towards the following observations:

- When $K > K^\star$, the clustering is unstable
- When $K = K^\star$, the clustering is stable
- When $K < K^\star$, the clustering can be stable or unstable depending on subtle differences of the distribution

This observation is promising for the cluster stability as a method for model selection in unsupervised clustering and one can hope that these results are transferable to other clustering algorithms. However, it should be noticed that for the case of simplicity, in these proofs, the stability of a clustering is treated as a binary state, where the clustering is either stable $S_K = 0$ or unstable $S_K > 0$. Yet, in practice the stability is a continuum and it is the responsibility of the user to judge whether a clustering is "stable enough". Also, one would like to know what happens when we are no longer in this theoretical "ideal" setting, that is, when we no longer have access neither to an infinity of observations, nor to the true distance between clusterings.

Experimental setting

In this study, we explored whether (1) if we had access to a large number of datasets, with a finite number of n, but we no longer have access to the true distance between clusterings, would we still able to detect the correct number of groups, and (2) if we only have access to one dataset, are able to correctly estimate the stability of a clustering.

To be consistent with the notations presented in the previous section, the algorithm for estimating cluster stability will be presented in a statistical context. Yet, it remains the same algorithm as described for the `clustRstab` function.

Algorithm and notations. Let's consider $\mathbb{X} \in \mathbb{R}^{n \times p}$, the dataset from which we want to estimate the stability of a clustering. To estimate the stability we will start by perturbing the initial dataset in order to obtain a large number of D datasets. This is done by using a given perturbation function f, for example subsampling datapoints or variables, such that $f : \mathbb{X} \to \mathbb{X}'_1, \ldots, \mathbb{X}'_D$, where $\mathbb{X}'_d$ are the resulting perturbed datasets for $d = 1, \ldots, D$. Then each perturbed dataset is clustered, using a given clustering algorithm $\mathcal{A}$, in order to obtain a clustering $\widehat{C}_K(\mathbb{X}'_d)$ of the n observations x_i in $k = 1, \ldots, K$ groups. This is done for different numbers of groups with $K = 2, \ldots, K_{max}$. For each K these obtained clusterings are then compared using a cluster comparison score $\widehat{Sc}()$ (based on distance or similarity). The estimation of the stability of a clustering is defined as the arithmetic mean of these scores with

$$\widehat{\mathcal{S}}_K := \frac{1}{nb.comp} \sum_{d<d'}^{D} \widehat{Sc}(\widehat{C}_K(\mathbb{X}'_d), \widehat{C}_K(\mathbb{X}'_{d'})), \tag{4.8}$$

where *nb.comp* is the number of comparisons done between the different clusterings. As already seen, there are different manners to compare the clusterings as well as different cluster comparison scores. In this simulation study we are going to compute pairwise comparisons between all the clusterings obtained from the perturbed dataset using a distance based score. The number of comparisons are then set to, $nb.comp = \binom{D}{2}$ and the stability index corresponds to a distance that we want to minimize in order to select the number of groups that corresponds to the most stable clustering with

$$\widehat{K} := \underset{K}{Argmin}\ \widehat{\mathcal{S}}_K.$$

. These general steps for computing cluster stability are presented in Algorithm 4.

Algorithm 4 Clustering Stability

1: **Generate D perturbed versions of the dataset $\mathbb{X}$ using perturbation function f such that**
$f : \mathbb{X} \to \{\mathbb{X}'_1, \ldots, \mathbb{X}'_D\}$
2: **for** $K = 2, \ldots, K_{max}$ **do**
3: **for** $d = 1, \ldots, D$ **do**
4: **Cluster each perturbed dataset** using a given clustering algorithm $\mathcal{A}$ such that
$\widehat{C}_K : \mathbb{X}'_d \to \{1, \ldots, K\}$
5: **end for**
6: **Compare the D obtained clustering** using a cluster comparison score $\widehat{Sc}$:
$\widehat{Sc} : \widehat{C}_K(\mathbb{X}'_1), \widehat{C}_K(\mathbb{X}'_2) \to$ *distance or similarity*
7: **Compute stability index** as the arithmetic mean of these comparisons with
$\widehat{\mathcal{S}}_K = \widehat{\mathbb{E}}(\widehat{Sc}(\widehat{C}_K(\mathbb{X}'_1), \widehat{C}_K(\mathbb{X}'_2)))$
8: **end for**
9: Choose the parameter K that gives the best stability, *for a distance*:

$$\widehat{K} := \underset{K}{Argmin}\ \widehat{\mathcal{S}}_K$$

The question is hence whether $\widehat{\mathcal{S}}_K$ is a good estimator of the "true" cluster stability $\mathcal{S}_K$. To test this, we generated (1) the "true" cluster stability by generating a large number of independent datasets from the same underlying probability distribution and (2) an estimation of this true cluster stability by generating a single dataset from the same probability distribution that was then subsampled. To test whether we could observe similar results as Ben-David et al. (2006, 2007) for the "idealized theoretical setting", this was done in two for two different data structures consisting of seven homogeneous groups. In the first data structure, the groups were all in equidistant from each other. To recall the work of Ben-David et al. (2006, 2007) we refer this data structure as symmetrical. In the second data structure, two of the groups were brought closer to each other than to the others. We refer this data structure as unsymmetrical. In order to relate our simulation study to the presented theoretical results, we are going to use the k-means algorithm and a distance based cluster comparison score.

Data generation. Data was generated for $K^\star = 7$ groups with different group means, μ_k for $k = 1, \ldots, 7$, and the same variance σ^2. Each group had $n_k = 50$ observations and each observation x_{ki}, with $i = 1, \ldots, 50$, and k indicating the group membership, had $p = 20$ repeated measures of the same variable generated from a Gaussian distribution with $X_{p'} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $p' = 1, \ldots, p$, and $\mu = [\mu_1, \ldots, \mu_7]$. The dataset $\mathbb{X} \in \mathbb{R}^{n \times 20}$ is the result of the concatenation of the different $X_{p'}$ variables where $n = 350$.

- **"Symmetrical" data structure** In the "symmetrical" data structure the different group means μ_k^s are in equidistant from each other with $\mu^s = [-6, -4, -2, 0, 2, 4, 6]$.

- **"Unsymmetrical" data structure** In the "unsymmetrical" data structure we change the mean value of 6^{th} group from $\mu_6^u = 4$ to $\mu_6^u = 5$. In this manner the 6^{th} and 7^{th} group are closer to each other than the other groups. Hence, the different group means no longer in equidistant from each other and $\mu^u = [-6, -4, -2, 0, 2, 5, 6]$.

In both data structures, the variance for each group was fixed to $\sigma^2 = 1$ in order for the groups to be slightly overlapping.

These two data structures are represented in Figure 4.4, where $V1$ and $V2$ are two variables generated from the symmetrical setting, *i.e.* $V1, V2 \sim \mathcal{N}(\mu^s, \sigma^2)$, and $V'1$ and $V'2$ are generated from the unsymmetrical setting, *i.e.* $V'1, V'2 \sim \mathcal{N}(\mu^u, \sigma^2)$. The dotted lines indicate the group mean-values for $V1$ and $V'1$, *i.e.* μ^s and μ^u. The density of $V1$ and $V'1$ is also plotted. As can be seen, when the mean-value for the 6^{th} group changes value from $\mu^u = 4$ to $\mu^u = 5$, the 6^{th} and 7^{th} the group becomes almost completely overlapping.

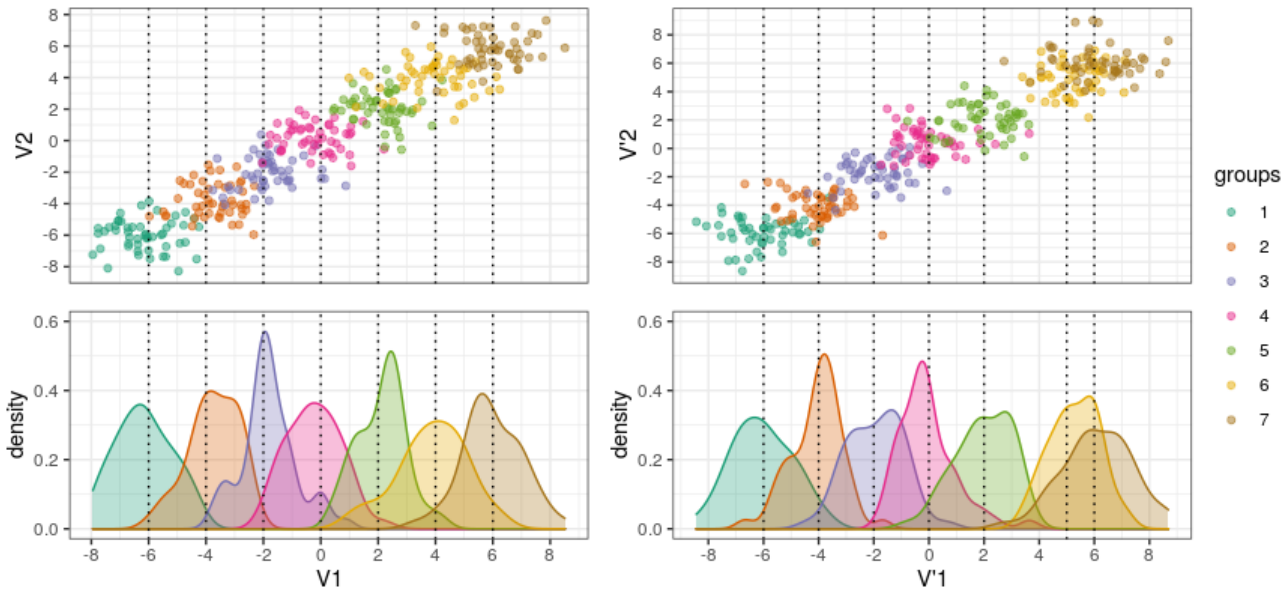


Figure 4.4 – Top: scatter plot for two variables (V1 and V2) generated according to the '*symmetrical*' data setting (left) and two variables (V'1 and V'2) generated according to the '*unsymmetrical*' data setting (right). Bottom: density plot V1 and V'1 The dashed lines indicate group mean values for V1 and V'1.

The datasets were generated based on a customized function based on the `rnorm` from the R-package `stats` (R Core Team, 2019a).

Computing cluster stability.

- **Clustering the datasets.** To cluster these datasets, we used the k-means algorithm from the R-package `stats` (R Core Team, 2019a), optimizing the $Q_K^{(n)}$ objective function. In order to avoid the algorithm to get stuck in local (suboptimal) optimum, and hence avoiding that the stability depends on the algorithm, we generated κ random initialization of the centroids μ, with $\kappa = 10$. The k-means algorithm then chooses the initialization that minimizes better the objective function; ending up with a clustering that should not depend on the initialization of μ.

- **Compare the obtained clustering.** In order to compare the obtained clusterings, we are going to compute pairwise comparisons using the Normalized Information Distance (NID, Vinh et al. (2010)). The NID is a distance function and a metric *i.e.* satisfying the positive definiteness,

symmetry and triangle inequality. It is based on the entropy of the clusterings and computed from the contingency table of the two clusterings. The NID is normalized within the range [0,1], equaling zero when the two clustering are identical, and 1 when they are independent. We use the NID score implemented in the R-package `aricode` (Chiquet et al., 2020).

- **Computing the cluster stability index.**

  - **Simulating the "true" cluster stability S_K.** In order to compute the "true" stability index, we generated a large number D of datasets $\mathbb{X}^1, \ldots, \mathbb{X}^D$ from the Gaussian probability distribution, $\mathcal{N}(\mu, \sigma^2)$, described above. The stability index was defied as

$$\mathcal{S}'_K := \frac{1}{\binom{D}{2}} \sum_{d<d'}^{D} NID(\widehat{C}_K(\mathbb{X}^d), \widehat{C}_K(\mathbb{X}^{d'}))$$

    where the $'$ in $\mathcal{S}'$ indicate that we are working on and estimation of the true minimal distances between two clusterings, by the NID, but not the true minimal distance itself.

  - **Estimating the simulated "true" cluster stability $\widehat{\mathcal{S}}_K$.** In order to estimate the "true" cluster stability, as would be done in practice, we only generated a single dataset $\mathbb{X}$ from the Gaussian distribution described above. A given proportion λ of the $p = 20$ variables of $\mathbb{X}$ was then randomly subsampled (without replacement). This was repeated D times, so that $subsampling_\lambda : \mathbb{X} \to \mathbb{X}'^{\lambda}_1, \ldots, \mathbb{X}'^{\lambda}_D$. This was done for $\lambda = 0.25, 0.35, \ldots, 0.95$ and the estimated stability index was defined as

$$\widehat{\mathcal{S}}'_{K_\lambda} := \frac{1}{\binom{D}{2}} \sum_{d<d'}^{D} NID(\widehat{C}_K(\mathbb{X}'^{\lambda}_d), \widehat{C}_K(\mathbb{X}'^{\lambda}_{d'})))$$

    For both these settings, the number of simulations (generated datasets) was fixed to $D = 500$. The estimated cluster stability was computed with the function `clustRstab` from the `clustRstab` R-package Sundqvist et al. (2020b).

Questions and hypotheses concerning the behavior of $\mathcal{S}_K$ and $\widehat{\mathcal{S}}_{K_\lambda}$.

Q1 - Is $\mathcal{S}'_K$ minimized at $K^\star$?

The first question that we want to investigate is whether the "true" cluster stability, $\mathcal{S}'_K$, will detect the correct number of groups, that is if the clustering will be the most stable for $K = K^\star = 7$ groups with $K^\star = \underset{K}{Argmin}\ \mathcal{S}'_K$.

Hypotheses: If the theoretical results of Ben-David et al. (2006, 2007) are transferable to our more realistic scenario (considering empirical values and not the limit case $n \rightarrow$), our hypothesis is that, $\mathcal{S}'_K$ should detect the correct number of groups in the symmetrical data setting, *i.e.* when the cluster centroids are all equidistant from each other with $\mu^s = [-6, -4, -2,0,2,4,6]$. Indeed, in this setting, there should be only one global optimum for the objective function $Q^{(n)_K}$ when $K = 7$ but several when $K \neq 7$. However, for the unsymmetrical data setting, *i.e.* $\mu^s = [-6, -4, -2,0,2,5,6]$, this should not be the case. Indeed, in this latter case, and due to the asymmetry between the cluster centroids (compared to each other), there will be unique global optimums even when $K \neq K^\star$. Particularly, this should be the case when $K = 6$ since two of the groups are now almost "merged" as well as for $K = 2$ since this new data structure is separating two of the groups from the five others. As noted by Von Luxburg et al. (2010), small changes in cluster boundaries, inducing "jittering" to the k-means algorithm by sampling variation (see Shamir and Tishby (2008a,b, 2009)) should not have a big impact on these results. Also, since we do several initializations for the k-means algorithm, we hope that the instability will not be induced from the k-means algorithm, but from the actual data structure.

Q2 - Is $\widehat{\mathcal{S}}'_{K_\lambda}$ a good estimator of $\mathcal{S}'_K$?

The second question we want to investigate is whether $\widehat{\mathcal{S}}'_{K_\lambda}$ is a good estimator of $\mathcal{S}'_K$. For this, there are two questions two answer; (1) is $\widehat{\mathcal{S}}'_{K_\lambda}$ a globally good estimator of $\mathcal{S}'_K$, that is, will they behave in a similar manner for different values of K, and (2) will they have the same minimum. Also, we want to investigate whether $\widehat{\mathcal{S}}'_{K_\lambda}$ behaves differently for different values of λ, *i.e.* the proportion of subsampled variables. If the latter is the case, we want to investigate if there is a given value for λ for

which the estimator behaves in a more similar manner to $\mathcal{S}'_K$ than the other values of λ.

Hypotheses. For the questions concerning the estimated cluster stability $\widehat{\mathcal{S}}'_{K_\lambda}$, it is much more delicate to make any hypothesis. Indeed, the theoretical results presented earlier in this paper do not cover this estimator. One can only hope that $\widehat{\mathcal{S}}'_{K_\lambda}$ is a good estimator of $\mathcal{S}'_K$ and hence a possible method for model selection in unsupervised clustering.

Simulation results. The results of this simulation study are presented in Figure 4.5. We will first consider the results for the "symmetrical" data structure. Here the "true" stability $\mathcal{S}'_K$ is minimized for $K = 7$, hence "detecting" the correct number of groups. The estimated stability $\widehat{\mathcal{S}}'_{K_\lambda}$ is also minimized for $K = 7$ for all values of λ except $\lambda = 0.95$ for which it is minimized at $K = 2$. It should also be noticed that the curves of $\widehat{\mathcal{S}}'_{K_\lambda}$ differ from each others and from the curve of $\mathcal{S}'_K$ when $K < 7$.

We now consider the results for the "unsymmetrical" data structure. The "true" stability $\mathcal{S}'_K$ has a "peak" at $K = 6$ but is minimized for $K = 2$. However, the minimum of $\widehat{\mathcal{S}}'_{K_\lambda}$ depends on the value of λ and is either $K = 2$ or $K = 6$. The standard deviations of theses indices are are large for $K \leq K^\star$ but small for $K > K^\star$ in both data settings (except for when $K = 2$ in the "unsymmetrical" setting).

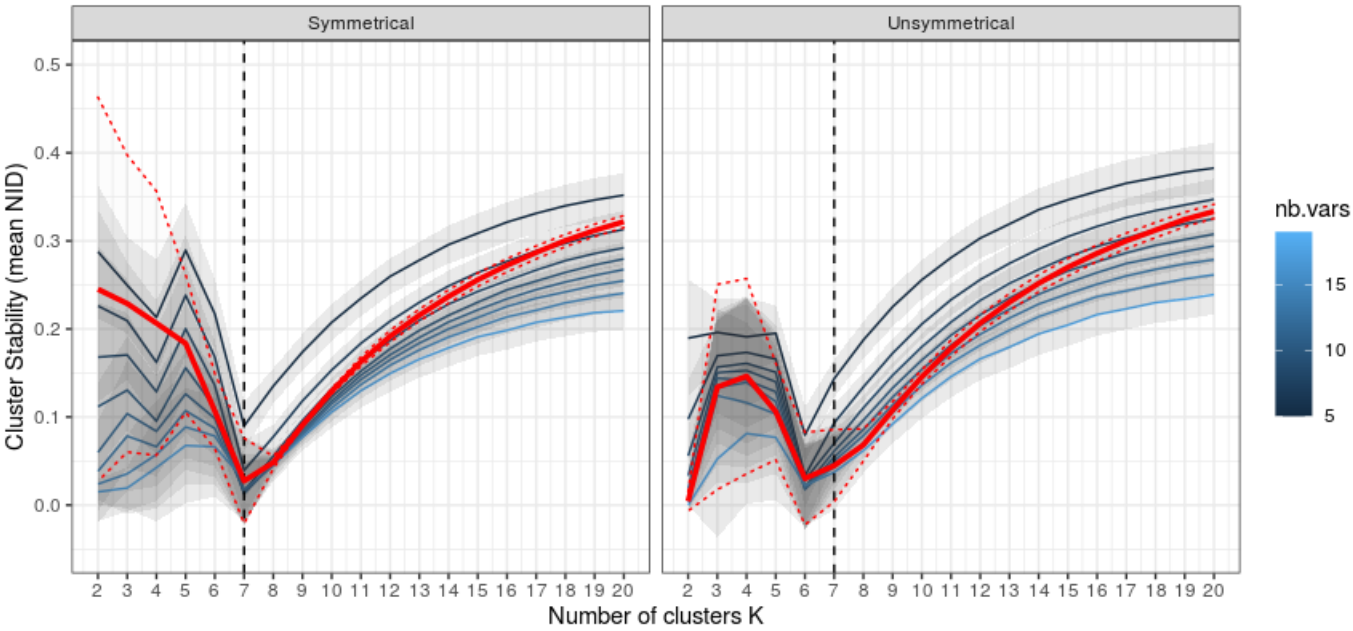


Figure 4.5 – Simulation results for cluster stability. In red: *mean* (solid line) and *standard deviation* (dashed lines) for $\mathcal{S}'_K$. In blue: *mean-values* for $\widehat{\mathcal{S}}'_{K_\lambda}$. Different nuances of blue correspond to different values of λ (here converted to the number of subsampled variables) with the lighter the blue, the larger λ. The gray areas correspond to the *standard deviations* of the different $\widehat{\mathcal{S}}'_{K_\lambda}$. The black dashed line indicates the true number of groups with $K^\star = 7$. To the left: results correspond to the "symmetrical" data setting with $\mu^s = [-6,-4,-2,0,2,4,6]$, *i.e.* all the cluster centroids are in equidistance from each others. To the right: results correspond to the "unsymmetrical" data setting with $\mu^u = [-6,-4,-2,0,2,5,6]$.

Discussion of simulation results. As a first observation, it should be noticed that, for none of the two simulations studies, the clustering was stable for $K > K^\star$. This is in line with the theoretical results presented earlier in this report and a promising indication for the use of cluster stability as a method for model selection in unsupervised clustering.

We will now consider the simulation results for the "symmetrical" data setting. As predicted by the theoretical work of Ben-David et al. (2006, 2007), in this setting the most stable clustering occurs for $K = K^\star$. As argued by these authors, it might be that there is a unique minimum for the objective function $Q_K^{(n)}$ when the clustering is in $K = K^\star$ groups, but that there are several global minimums when $K \neq K^\star$, hence inducing instability. In this setting, the correct "detection" of $K^\star$ is observed for both the "true" cluster

stability $\mathcal{S}'_K$ and its estimation $\widehat{\mathcal{S}}'_{K_\lambda}$ and, is again, a promising indication for the use of cluster stability as a method for model selection in unsupervised clustering. Nonetheless, it should be noticed that, this kind of symmetry in the distribution of the data is rarely (or never) observed in real data settings. Also, as soon as one of the group centroids (mean-value) was slightly shifted, both the "true" cluster stability and its estimation behaved in a more unpredictable manner.

Considering now the simulation results for the "unsymmetrical" data setting, *i.e.* the setting where the mean-value of the 6^{th} group have been slightly shifted towards the mean-value of the 7^{th} group and hence almost merging the datapoints of these two groups. In this setting, the minimum of the "true" cluster stability $\mathcal{S}'_K$ did not occur nor for $K = K^\star$ groups, neither for $K = 6$ (as could have been expected) but for $K = 2$ groups. This can be understood since the "merging" of the datapoints from the 6^{th} and the 7^{th} group separates them from the rest of the datapoints. As a consequence, and according to the work of Ben-David et al. (2006, 2007), this might create a situation where there is a unique global minimum for the objective function $Q_K^{(n)}$ when $K = 2$, resulting in a stable cluster structure, but several global (and local) minima for the other values of K. This indicates that the most stable clustering is not always the most interesting. Indeed, in this setting, a clustering of $K = 6$ groups would tell us more about the data distribution than a clustering of $K = 2$ groups. Yet, it is the latter one that is the most stable. Even more concerning, in this data setting, the minimum of the estimated stability $\widehat{\mathcal{S}}'_{K_\lambda}$ is shifting between $K = 2$ and $K = 6$ groups, depending on the value of the parameter λ, *i.e.* the proportion of subsampled variables. This illustrates the importance of testing different values of such parameters when perturbing the initial dataset. Yet, as seen eariler, in most of the implementations of cluster stability methods, this parameter is fixed by default. This simulation study shows that such default choice can have important consequences since, like in this "unsymmetrical" setting, different values of λ would result in two completely different clusterings.

Conclusion All together, the presented theoretical results, as well as our simulation study, show that in some simple data settings, especially when the groups of the data are "symmetrically" distributed, cluster stability works well as a method for model selection in unsupervised clustering. However, when the datapoints are distributed in a more complex manner, which is usually

the case, this is no longer true. Indeed, as predicted by the theoretical work of Ben-David et al. (2006, 2007), we observed that the most stable clustering does not necessarily correspond, neither to the correct number of groups, nor to the most interesting clustering.

4.2.5. The NCI60 cancer study

In order to investigate how cluster stability can be used in a more realistic setting, we are going to use the NCI60 cancer dataset (Ross et al., 2000). The advantage with a dataset like the NCI60 is that the ground truth partition of the observations with their respective labels is known. Hence, we are able to compare the obtained clustering results to these labels. The NCI60 dataset consists of 60 cell lines of 9 different cancer types for which 6830 genes (RNA-microarray) have been measured. The different cancer types are: Central Neuronal System (CNS) ($n = 5$), Breast ($n = 7$), Non-Small-Lung (NSCLC) ($n = 9$), Ovarian ($n = 6$), Leukaemia ($n = 6$), Colon ($n = 5$), Melanoma ($n = 8$), Prostate ($n = 2$), Renal ($n = 9$), Unknown ($n = 1$). The dataset is implemented in the `clustRstab` package which is accessible by the command `data(NCI60)`, resulting in a list with two objects: `NCI60$expr`, a 6830×64 data frame containing the gene expression for the different cell lines and `NCI60$type` a vector indicating the cancer type of each cell line.

Clustering the NCI60 dataset

As often done in oncology studies, we conducted the clustering analysis on the most varying genes with the hypothesis that these latter "carry" the biological signal of interest. We therefore selected the genes with the *standard deviation:* $SD > 0.8$ (a commonly used threshold) resulting in $p = 1\,689$ genes. We also removed the cell line with an unknown cancer type as well as some technical replicates and ended up with a dataset of $n = 59$ observations and $p = 1\,689$ genes. The resulting clusterings of this dataset into $K = 9$ groups are shown in Figure 4.6. The dendrogram in this figure corresponds to the hierarchical ascendant clustering algorithm with the Ward distance (HAC-Ward), the gray-scaled bars correspond to the k-means algorithm and the Gaussian Mixture Model (GMM, with EII, spherical distribution for equal volume and shape, covariance model).

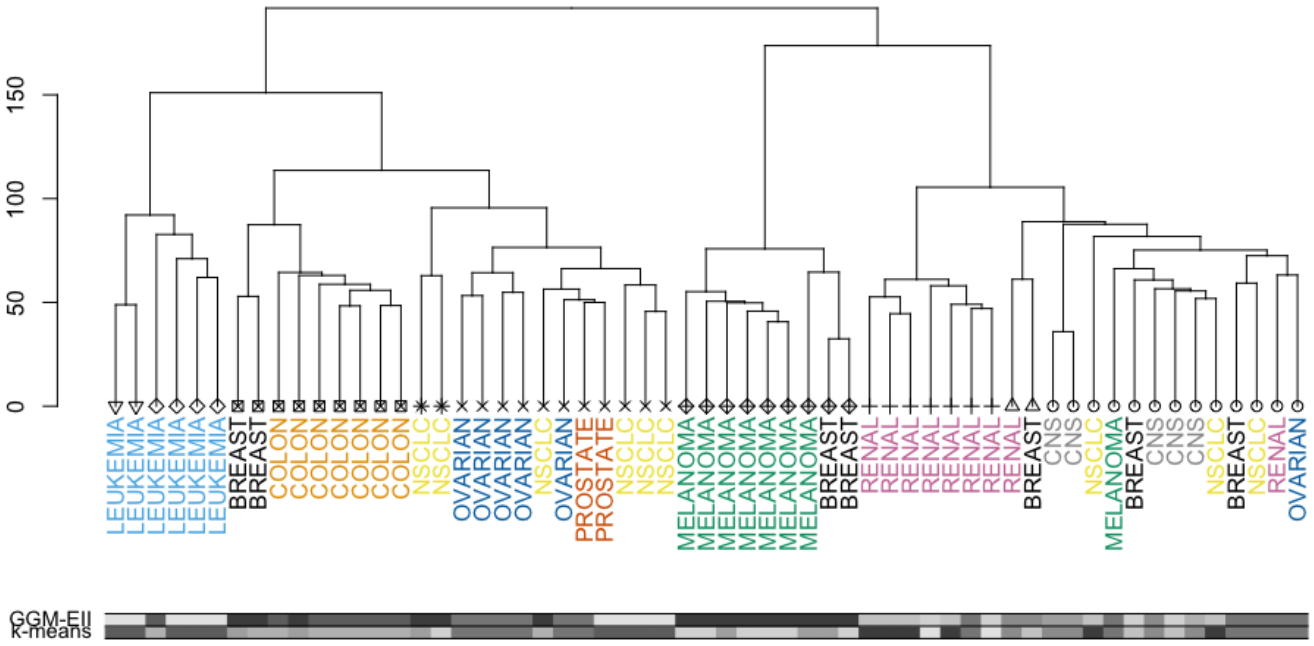


Figure 4.6 – Clustering of the NCI60 dataset: the color code indicates cancer type. Symbol on dendrogram leaf indicates group belonging for HAC-Ward clustering when $K = 9$. Colored bares: GGM-EII (upper line) and k-means clustering for $K = 9$. Different scales of gray indicate different group belonging.

As can be seen in Figure 4.6, the initial 9 cancer types are quite well recovered by the different clusterings algorithms. Indeed, it is only the breast and the NSCLC tumors that are dispersed among the other cancer types. This is inline with the results of Ross et al. (2000).

Estimating the number of groups in the NCI60 dataset

In a regular clustering setting, we would have had to estimate the number of groups before analyzing the clustering results. To do so, we are going to estimate the stability of the clustering using the `clustRstab` function. In the simulation study, we saw that this method depends on the proportion of sub-sampled variables. It is therefore reasonable to think that, other parameters, such as the method for perturbing the data (subsampling, adding noise to the data *etc.*), the type of clustering algorithms, as well as the cluster similarity score *etc.*, could have a similar impact on the stability estimation. In order to avoid the influence of such "algorithmic" parameters, we are going to estimate the cluster stability with different parameter settings. We are also going to

compare the stability results with other cluster validation criteria.

Using cluster stability

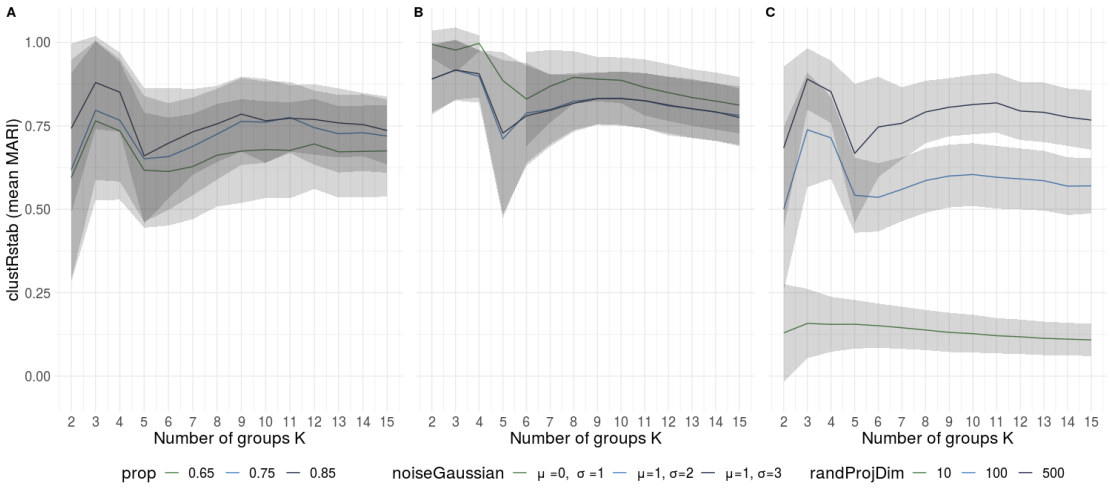


Figure 4.7 – `clustRstab` for the NCI60 dataset for different strategies and levels of data perturbation. A. `perturbedDataFun = subSample` and the color code indicates the level of subsampled items with `nProp = pProp`. B. `perturbedDataFun = noiseGaussian`, the color code indicates different values of the mean (μ) and standard deviation (σ) of the added noise. C. `perturbedDataFun = randProjData`, the color code indicates the number of dimension in which the data are projected. For A., B. and C.: `nsim = 500`, `clAlgo = clAlgoKmeans`, `typeOfComp = "all"`, `clCompScore = MARI`

Data perturbation functions. In Figure 4.7 we measured the cluster stability of the NCI60 dataset using the three perturbation functions defined for the `clustRstab` package (subsampling, adding noise and random projection). For each of these perturbation functions, we varied the level of perturbation. The cluster comparison score $MARI \in [0:1]$ (Sundqvist et al., 2020a) was used. This score estimates the probability for two observations to be clustered together in two clusterings. Thus, the most stable clustering corresponds to the one with the highest $MARI$ score. As can be seen in this figure, the most stable clustering occurs for $K = 3$ for most of the stability estimations. This is the case in Figure 4.7A. where the data have been subsampled for different levels of `nProp=pProp` and in Figure 4.7C. where the data have been randomly projected in different number of dimensions. However, when Gaussian noise is added to each datapoint, as done in Figure 4.7C., the most stable clustering

occurs for $K = 3$ or $K = 4$ depending on the values of `noiseGaussianMean` (μ) and `noiseGaussianSD` (σ). An example of when the dataset is too much perturbed is found in Figure 4.7B. Indeed, when the dataset is projected in only 10 dimensions, the structure disappears for all K. Other technical details of these stability estimations are given in the caption of the figure.

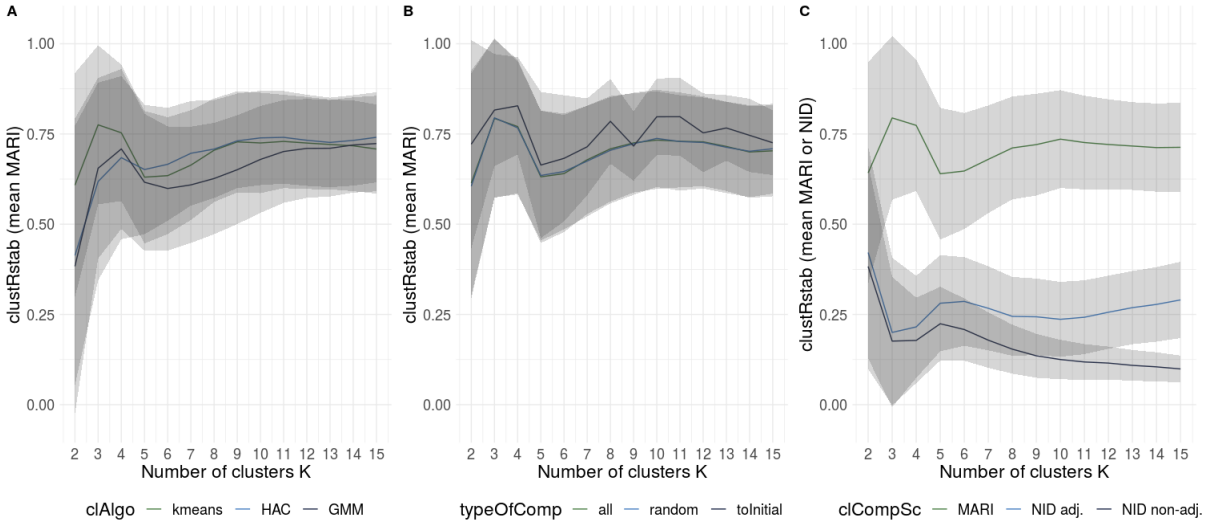


Figure 4.8 – `clustRstab` A. `perturbedDataFun = subSample` and the color code indicates the level of subsampled items with `nProp = pProp`. B. `perturbedDataFun = noiseGaussian`, the color code indicates different values of the mean (μ) and standard deviation (σ) of the added noise. C. `perturbedDataFun = randProjData`, the color code indicates the number of dimension in which the data are projected. For A., B. and C.: `nsim = 500`, `clAlgo = clAlgoKmeans`, `typeOfComp = "all"`, `clCompScore = MARI`

Clustering algorithms. As can be seen in Figure 4.6, different cluster algorithms result in different partitions of the observations. In order to investigate whether the type of clustering algorithm also has an impact on the estimated clustering stability, we measured the stability of the NCI60 dataset differing only the clustering algorithm. The results for this can be found in Figure 4.8A. As can be seen in this figure, whereas the k-means algorithm generates the most stable clustering for $K = 3$, the HAC and the GMM algorithms generate the most stable clustering for $K = 4$.

Cluster comparison strategy In order to test whether the comparison strategy had any effect on the NCI60 dataset, we computed these different

similarity index while keeping all the other parameters constant. The results are found in Figure 4.8B. As can be seen in this figure, the most stable clustering $K = 3$ or $K = 4$ depends on the type of comparison. The random sampling of pairwise comparisons (`typeOfComp = "random"`) and the complete pairwise comparisons (`typeOfComp = "all"`) are behaving very similarly. This is not surprising since the number of simulations is large, $nsim = 500$. However, this indicates that one can probably use the `typeOfComp = "random"` argument to gain computational effort, at least when $nsim$ is large.

Type of cluster comparison score. We also investigated the effect of the $MARI$ and the NID (adjusted for chance or not) on the stability of the obtained clusterings. The results are to be found in Figure 4.8C. For recall, the $MARI$ is a similarity based score that one hence wants to maximize, whereas the NID is a distance-based score that one wants to minimize. First, it can be noticed that the $MARI$ and the adjusted NID have quite similar stability curves. That is, they are both optimized at $K = 3$ and then decrease for $K = 4$,5 and then increase to "flatten out". The non-adjusted NID score, however, has a local minimum (at $K = 3 = 4$), then increase a bit before decreasing again ($K > 5$). That is, with the uncorrected NID value, the clusterings become the more and more stable with the increase of K. Again, this shows the importance for adjusting cluster comparison scores for chance.

Using other validation criteria

In Figure 4.9 we compare the results from the `clustRstab` (D) function upon the NCI60 dataset with the following clustering criteria: the (A) Elbow method (computing the intragroup variance of a clustering), (B) the Silhouette-average (measuring how close each observation in a cluster is to observations in the neighboring clusters), (C) the Gap statistic (comparing the change in within-cluster dispersion with the one expected under an appropriate reference null distribution) and (E) the Bayesian Information Criterion (BIC - computing the log-likelihood penalized by the number of groups in a GMM probabilistic setting). For the Elbow method, since the variance will decrease with the number of groups, one search for a change in the curve. In this case it is quite difficult to find such chnage, but it might appear for $K = 4$ or $K = 5$. For the Silhouette average, one wants the observations to be as "far as possible" from the neighboring cluster, hence this criterion should be maximized. For

the NCI60 dataset this is the case for $K = 13$, but there is also a peak at $K = 4$. For the Gap-statistic, one wants that the "gab" (difference) between the obtained clustering, and what would have been expected by chance, to be as large as possible without increasing for larger K. This occurs for $K = 6$ in the NCI60 dataset. Finally, the BIC criteria should be maximized which is the case for the three tested covariance GMM at $K = 4$.

Hence, these different clustering criteria yield in different results concerning "the best number of groups" for the NCI60 clustering with K estimated from $K = 4$ to $K = 6$ groups. It should be noticed that these criteria indicate a larger K than the K found with the clustering stability estimation where the "best number of groups" varied from $K = 2$ to $K = 5$ groups but mostly indicated $K = 3$. In sight of the theoretical results of Ben-David et al. (2006, 2007) showing that a clustering might be stable for a number of groups inferior, but not superior, to the correct number of groups depending on the data structure, this makes sense.

Globally, these results stress out that model selection in unsupervised clustering is a difficult task and that combining different criteria and considering different number of groups might be useful in order to understand the underlying data structure.

Ground truth comparison

To better understand the obtained clusterings in the NCI60 dataset, we compared these clusterings with the NCI60 ground truth (the 9 cancer types). To do so, the MARI was computed for the groud truth and each of the obtaned clusterings. The results of these comparisons are found in Figure 4.10 and in Table 4.1. In Figure 4.10, we see that the most similar clustering to the 9 cancer types partition occurs for $K = 6$ to $K = 9$ groups. In the contingency tables presented in Table 4.1, we see that (1) the clustering of NCI60 in $K = 3$ groups separates the melanomia cancer type from the other cancer types, (2) when $K = 4$, both melanomia and leukemia are separated from the other cancer types, (3) when $K = 6$, melanomia, leukemia and colon are separated from the other cancer types and (4) when $K = 9$, three groups with only two observations appear. As a consequence, the most stable clustering ($K = 3$ or $K = 4$), does neither correspond to the most similar clustering to the groud truth ($K = 6$ to $K = 9$), nor to the one that recover the most cancer types. This, again, indicates that a stable clustering is not necessarily interesting.

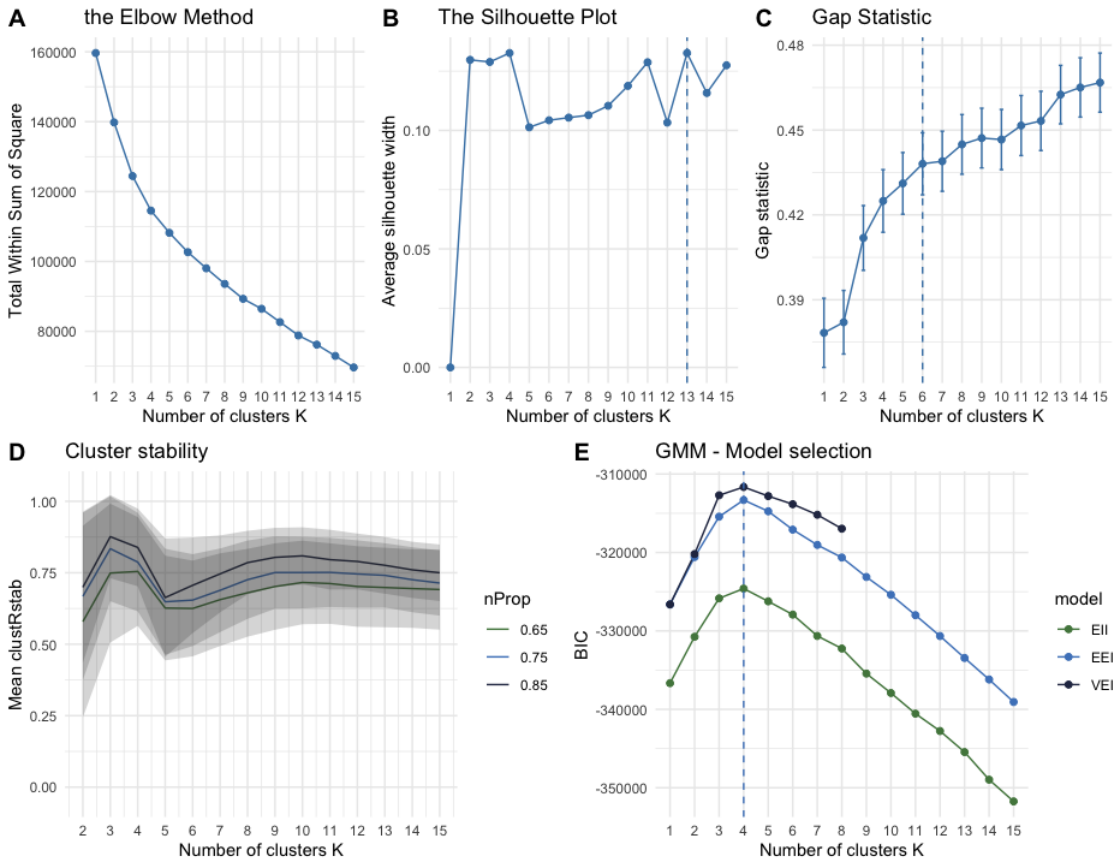


Figure 4.9 – Different methods for selecting the number of groups in the NCI60 dataset. For all methods except for the GMM, the k-means algorithm was used with 30 initializing points. The GMM was otained using the *mclust* R package (Scrucca et al., 2016a) and the covariance models are: "EII", spherical, equal volume, "VEI", diagonal, varying volume, equal shape and "VVI" diagonal, varying volume and shape. Plots A, B, C and E were obtained using the *factoextra* R package, proposed by Kassambara and Mundt (2017). Plot D was constructed by the `clustRstab` package. It uses the $MARI$: the higher the $MARI$, the more stable the classification. $nProp$ corresponds to the proportion of subsampled observations.

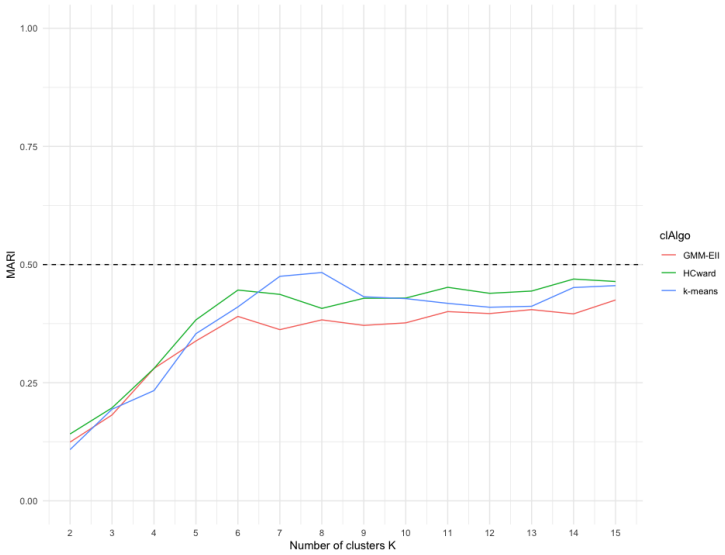


Figure 4.10 – Comparison (MARI) between the clusterings of the NCI60 dataset into different number of groups and for different clustering algorithms with the ground truth *i.e.* the 9 cancer types.

	$K=3$			$K=4$				$K=6$					
Cancer Type	1	2	3	1	2	3	4	1	2	3	4	5	6
BREAST	3	2	2	3	2	0	2	3	0	0	0	2	2
CNS	5	0	0	5	0	0	0	5	0	0	0	0	0
COLON	0	7	0	0	7	0	0	0	0	0	0	7	0
LEUKEMIA	0	6	0	0	0	6	0	0	0	0	6	0	0
MELANOMA	1	0	7	1	0	0	7	1	0	0	0	0	7
NSCLC	3	6	0	3	6	0	0	3	0	6	0	0	0
OVARIAN	1	5	0	1	5	0	0	1	0	5	0	0	0
PROSTATE	0	2	0	0	2	0	0	0	0	2	0	0	0
RENAL	9	0	0	9	0	0	0	2	7	0	0	0	0
	$MARI = 0.19$			$MARI = 0.28$				$MARI = 0.44$					

	$K=9$								
Cancer Type	1	2	3	4	5	6	7	8	9
BREAST	2	1	0	0	0	0	2	0	2
CNS	5	0	0	0	0	0	0	0	0
COLON	0	0	0	0	0	0	7	0	0
LEUKEMIA	0	0	0	0	4	2	0	0	0
MELANOMA	1	0	0	0	0	0	0	0	7
NSCLC	3	0	0	4	0	0	0	2	0
OVARIAN	1	0	0	5	0	0	0	0	0
PROSTATE	0	0	0	2	0	0	0	0	0
RENAL	1	1	7	0	0	0	0	0	0
	$MARI = 0.42$								

Table 4.1 – The contingency tables for the NCI cancer types (in lines) and the HAC-Ward clustering (columns) for $K = 3,4,6,9$. The bottom row indicates the MARI value between the two classifications. In order to have a better visibility of the tables, the cells containing 0 are indicated in gray.

Conclusions NCI60 application study

In this application study we investigated how different parameter settings can affect the measured stability of a clustering. Do do so, we used our R-package `clustRstab` and the NCI60 dataset (Ross et al., 2000) consisting of 59 cell lines of 9 different cancer types. We saw that the most stable clustering for this dataset occurred when $K = 3$ or $K = 4$ and hence not $K = 9$. The cluster stability measures found similar results compared to other clustering validation criteria but tended to estimate the number of groups K a bit smaller. More importantly, we saw that the most stable clustering of the NCI60 dataset was not the most similar to the partition of the 9 cancer types. Hence, in this case, the most stable clustering was not the most interesting or the one that reviled the most about the underlying data structure.

4.3. Conclusion

In this chapter we presented the `clustRstab` R package that allows to measure cluster stability in a flexible manner for the task of selecting the number of groups in unsupervised clustering. Compared to other R packages, it is advantageous since it allows to:

- combine different algorithmic parameters in order to compute the stability,
- adjust the stability criteria for chance,
- generate a large number of repetitions ($nism$) in a reasonable amount of time
- generate both the mean and the standard deviation as outputs.

We therefore argue that our package `clustRstab` has its interest of its own.

Using the `clustRstab` package, we conducted a simulation study and an application study. In these two studies we showed that:

1. the most stable clustering does not always, even in simple settings, correspond to the correct number of groups
2. we are not always able to correctly estimate this stability

3. the most stable clustering is not always the most interesting or the one that reveals the most information of the data structure

4. the estimation of cluster stability is parameter dependent.

For all these reasons, cluster stability, even though it is an interesting characteristic for a clustering, it is not a "magical" measure and should therefore be used with caution. For example, by testing it with different algorithmic parameter settings, which is possible with the `clustRstab` package, as well as combining it with other clustering validation criteria.

Chapter 5

TNBC classification of the TCGA dataset

Résumé. L'identification des sous-types de cancer du sein triple négatif (TNBC) est une priorité en oncologie, car elle pourrait permettre d'identifier de nouveaux traitements ciblés pour les patients atteints de TNBC. À cette fin, Lehmann et al. (2011) a proposé une classification transcriptomique de six sous-types de TNBC. Plusieurs groupes de recherche ont depuis essayé de reproduire ces groupes sur d'autres bases de données avec un succès mitigé. Dans cette étude, j'ai classifié une large cohorte d'échantillons de tumeurs TNBC obtenus de la base de données TCGA sur la base de l'expression transcriptomique RNAseq. Pour ce faire, j'ai utilisé deux stratégies : (1) j'ai prédit les sous-types Lehmann et al. (2011) en utilisant l'outil TNBCtype de Chen et al. (2012) et (2) j'ai utilisé des méthodes de classification non supervisées et des critères de validation que j'ai développé dans cette thèse. Mon objectif était de déterminer si ces deux stratégies, en utilisant les mêmes données (échantillons et gènes), permettraient d'obtenir des classifications similaires. Mon hypothèse était que, si c'était le cas, le signal observé correspondrait à un signal biologique "robuste", présent dans les tumeurs TNBC. Les résultats montrent que c'est le cas en particulier pour deux des sous-types de Lehmann et al. (2011), l'mésenchymateux et l'immunomodulateur qui ont tous les deux été détectés par l'outil TNBCtype et par la classification la plus stable ($K = 2$). Un troisième sous-type, le basal-like 2, a été identifié dans une classification ayant $K = 6$ groupes et a montré des taux de survie plus faibles que les autres groupes. Ceci est très intéressant, car ceci pourrait lier ce groupe à un résultat clinique. Ainsi, dans cette étude, le TNBCtype s'est révélé d'être un outil robuste pour le sous-typage TNBC. Cependant, cette méthode a également ses limites et, par exemple, une grande partie des échantillons a été considérée comme non spécifiée. Afin de comprendre pleinement l'hétérogénéité des échantillons de tumeurs TNBC, il serait donc intéressant de les comparer avec les résultats d'autres classifications.

Abstract. Identifying Triple Negative Breast Cancer (TNBC) subtypes is a priority in oncology since it could allow to identify new targeting treatments for TNBC patients. For this aim Lehmann et al. (2011) proposed a transcriptomic classification of six TNBC subtypes. Several research groups have since then tried to replicate these groups on other datasets with mixed success. In this study I classified a large cohort of TNBC tumor samples obtained from the TCGA database, based on RNAseq transcriptomic expression. In order to do so, I used two strategies: (1) I predicted the Lehmann et al. (2011) subtypes using the TNBCtype tool (Chen et al., 2012) and (2) I used unsupervised clustering methods and clustering validation criteria that I had developed in this thesis. My goal was to investigate if these two strategies, when using the same data (samples and genes) would result in similar classifications. My hypothesis was that, if that was the case, the observed signal correspond to a biological "robust" signal, present in the TNBC tumors. The results show that this is the case especially for two of the Lehmann et al. (2011) subtypes, the mesenchymal and the immunomodulatory which were both detected by the TNBCtype tool and by the most stable clustering ($K = 2$). Also, a third subtype, basal-like 2, was identified the in the clustering of $K = 6$ groups and showed lower survival rates than the other groups. This is very interesting since it could eventually link this group to a clinical outcome. Hence, in this study the TNBCtype showed to be a robust tool for TNBC subtyping. However, this method also has its limits and, for example, a large proportion of the samples, were considered as unspecified. In order to fully understand the heterogeneity of the TNBC tumor samples, it would therefore be interesting to compare with other classification results.

5.1. Introduction

Triple Negative Breast Cancer (TNBC) is an aggressive form of cancer for which no molecular specific treatment currently exists. The TNBC tumors are heterogeneous and, in contrary to other breast cancer types, do not express hormone (oestrogen or progesterone) or HER2 receptors whereof the label "triple negative". As a consequence, finding homogeneous subtypes of TNBC has become a priority in oncology. Indeed, this could lead to the development of molecular based treatments, specific for each group. For this reason, several TNBC classifications have been proposed (Bonsang-Kitzis et al., 2016; Burstein et al., 2014; Lehmann et al., 2011, 2016; Masuda et al., 2017), de-

scribed in Section 1.3.2. In this thesis I also classified the TNBC tumors of the RATHER consortium, using transcriptomics Section 1.5 and using proteomic 2. By classifying this latter dataset, as well as taking into account the results of the other classification studies, I realized how difficult the task of subtyping TNBC tumors is. Indeed, the obtained classifications depend, not only on biological signals, but also on many different experimental parameters, such as the patient cohort used, the normalization of data, the gene (or protein) selection as well as the clustering methods.

Comparing different classification results is therefore important and could hopefully allow to distinguish the biological results from these methodological constraints. Indeed, it is possible to think that, if a subgroup, associated to a specific gene signature, is found in several classifications, it is revealing a robust biological signal present among the TNBC tumors. Yet, comparing the results of different classification results is difficult to implement, since, for example, data cohorts differ, normalization pipelines of data are not always published, variables differ in between different studies *etc.* Correctly comparing different classification results would imply to have access to the data and/or the scripts, for data normalization and statistical analysis, of the different TNBC classification studies. I did not manage to obtain this during my thesis. However, the research group of Lehmann et al. (2011) has made such comparison possible by proposing (1) a transcriptomic signature (by abuse of notation I will refer to this transcriptomic signature as to a gene signature) that can be used for clustering TNBC samples, as well (and more importantly) as a TNBC subtype tool (TNBCtype) that allows to predict the subtype of a TNBC sample based on their initial classification (Chen et al., 2012). This group was the first to propose a TNBC classification and they identified 6 subgroups associated to the following different gene signatures: basal-like 1 (BL1); basal-like 2 (BL2); immunomodulatory (IM); mesenchymal (M); mesenchymal stem–like (MSL); luminal androgen receptor (LAR); and finally a group of samples that were unstable (UNS). A limitation of this tool is that, if there are differences between the compared classification and the predicted subtypes, it is difficult to investigate from where these differences origin. Indeed, they might have been induced from methodological choices made by either the group of Lehmann et al. (2011) or the one conducting the classification study. However, this tool remains very practical and has so far been used for classification comparison in several TNBC classification studies (Bonsang-Kitzis et al., 2016; Burstein et al., 2014; Masuda et al., 2017). The results of these comparisons have been

varying. For example, Bonsang-Kitzis et al. (2016) found descriptive (but no statistical) similarities between their identified 6 subgroups and the ones of Lehmann et al. (2011) and Burstein et al. (2014) found similarities between their identified 4 subgroups and the ones of Lehmann et al. (2011) but did not manage to replicate the classification of Lehmann et al. (2011) when they used the gene signature proposed by these latter. Also, when I clustered the RATHER consortium dataset in Chapter 1.5, I found a subgroup of IM, a subgroup of M and a subgroup of LAR samples in the dataset whereas the samples from the other groups were mixed.

In order to get a better understanding of how the TNBCtype tool can be used for TNBC subtype prediction, I am going to cluster a large TNBC dataset, using the clustering methods and strategies that I have developed in my thesis and compare the clustering results to the predicted subtypes of the TNBCtype tool. My hypothesis is that, if similar groups are found by these two classification strategies, these groups correspond to a robust biological signal present among the TNBC tumors. However, and as already mentioned, if the classification results are different, they will be difficult to interpret since the differences might as well correspond to different biological as to different methodological factors.

5.2. Data

5.2.1. The TCGA dataset

The data used in this study were generated by The Cancer Genome Atlas (TCGA) Research Network: `https://www.cancer.gov/tcga`. I downloaded the transcriptomic data from the cbioportal (TCGA 2018) portal. It contained RNAseq data for $n = 1084$ breast cancer (BC) patients. In order to get the hormonal status of these patients, I downloaded the clinical data from the GDC portal of the National Cancer Institute. I also obtained survival data for the patients, with an average of 10 years of surveillance, from the study of Liu et al. (2018).

5.2.2. Inclusion criteria and breast cancer types

Inclusion criteria

Only female patients with a defined status (positive or negative), of estrogen receptors (ER), progesterone receptors (ER) and epidermal growth factor receptor (HER2), measured by immunohistochemistry (IHC), were included in this study. For those patients who had "Equivocal" IHC-HER2 status, the HER2 fish status was used. This resulted in an inclusion of $n = 854$ BC patients.

Breast cancer type definition.

I manually classified the tumor samples based on their ER/PR and HER2 statuses and the definitions of Perou et al. (2000); Sorlie et al. (2003) into 4 BC types with: two hormonal positive BC types (luminal A and luminal B), a HER2 positive type (HER2(+)) and the TNBC. The ER/PR/HER2 IHC statuses of each BC type are shown in Table 5.1. It should be noticed that there are two BC types that are HER2 positive, the luminal B and the HER2(+). This latter is actually ER(-)HER2(+), but I will refer to it as HER2(+) in order to avoid making notations too heavy.

Breast Cancer Subtypes	Definition	Patients Counts	Patients Percentage
Luminal A	ER/PR+, HER2-	528	62%
Luminal B	ER/PR+, HER2+	130	15%
TNBC	ER/PR-, HER2-	152	18%
HER2 (+)	ER/PR-, HER2+	43	5%

Table 5.1 – Cancer type: Definition and number of patients

As can be seen in Table 5.1, the TNBC patients represent 18% of the included patients. This corresponds to what would be expected since the TNBC represents around 15% of all BC (Severson et al., 2015). A supplementary TNBC tumor sample was excluded since it was considered as possible ER+ by the TNBCtype tool.

5.3. Methods

Analysis pipeline. Before classifying the TNBC dataset, I will explore the dataset including all the BC tumor samples. This will allow me to do two things:

1. Investigate how similar or different the TNBC tumors are compared to the other BC tumors.

2. Exclude the TNBC (manually) IHC classified tumors that have gene expression patterns similar to the other BC types.

To do so, I will cluster the BC tumors into $K = 4$ groups (the same number as the BC types) using the k-means and hierarchical ascendant clustering with Ward distance (noted HAC-WARD) algorithms, using the R package `stats` (R Core Team, 2016), as well as the Gaussian Mixture Model with the equal sphere and size covariance model (noted GMM-EII) algorithm from the R package `mclust` (Scrucca et al., 2016a). These methods are described in Section 1.5. The obtained clusterings will then be compared to the 4 IHC BC types. Survival analysis for the BC types will also be conducted by estimating the Kaplan-Meier estimator using the `survival` R package (Therneau, 2015). The survival plots will be obtained with the `survminer` R package (Kassambara et al., 2020).

For the TNBC clustering and subtyping, I will only include the TNBC tumors that are clustered together in this BC clustering analysis.

5.3.1. TNBC classification strategies

I will use two different **classification strategies** to subtype the TNBC tumor samples. First I will use the TNBCtype tool (Chen et al., 2012) in order to **predict** the TNBC subtypes. Second I will use unsupervised clustering methods and different clustering validation criteria to **cluster** the TNBC tumor samples.

TNBC subtype prediction: The TNBCtype tool.

The TNBCtype tool (Chen et al., 2012) predicts the subtypes of TNBC samples based on the Lehmann et al. (2011) classification. To construct this tool, they conducted differential analysis for their 6 identified TNBC subgroups and defined a gene signature based on the most differentially expressed genes for each subgroup. Based on these genes they defined a centroid (the arithmetic mean) for each subgroup. To predict the subtype of a candidate TNBC tumor (sample or cell line) Spearman correlation is conducted between this candidate and each of the six subtypes centroids. The candidate is then assigned to the

TNBC subtype (BL1, BL2 IM, M, MSL, or LAR) with the highest correlation coefficient. Those candidates that have low correlation coefficients (correlation coefficient < 0.1 or $p.value > 0.05$), or are similar between subtypes (difference of two largest correlation coefficients < 0.05) are considered unclassified (UNS).

TNBC clustering

In order to cluster the TNBC samples, I will use the same clustering algorithms as described for the clustering analysis of the BC types. In order to select the number of groups, I will use various cluster validation criteria described in Section 1.5.2. These cluster validation criteria will be compared with the stability of the k-means clusterings computed by the `clustRstab` package (Sundqvist et al., 2020b) that I implemented in my thesis.

I will then compare the obtained clusterings with the subtypes predicted by the TNBCtype tool using the Modified version of the Adjusted Rand Index ($MARI$) (Sundqvist et al., 2020a), that I proposed in this thesis and that is described in Chapter 3. Survival analysis will also be conducted for the different obtained classifications using the same methods as described for the BC types.

5.3.2. Data prepossessing and gene selection

RNA preprocessing.

The RNAseq data contained $p = 20\ 531$ genes (median) counts for each BC patient. I normalized the data by the `voom` function of the R-package `limma` (Ritchie et al., 2015). As described in the following tutorial, this function: (1) transforms the counts to log2 counts per million reads (CPM: based on calculated normalization factors), (2) fits a linear model to the log2 CPM for each gene (here based on the BC type), and the residuals are calculated. Genes with low variation ($CPM < 5$) were excluded.

BC clustering gene selection.

For classifying the BC samples, I selected the most varying genes. For this, I tested several thresholds for the standard deviation (SD) ($threshold = \{1, 10, 16\}$) and I selected $threshold = 1$ since it gave the best clustering of the TNBC samples. As a consequence, $p = 7\ 636$ genes where included ($SD > 1$).

TNBC clustering gene selection.

In order to not influence the normalization procedure for the RNAseq data of the TNBC tumors, I normalized the raw RNAseq data (as described above) in the TNBC clustering analysis, for only the included tumor samples. This clustering analysis was based on the $p = 2\ 110$ genes that were in common between the present dataset and the $p = 2\ 188$ Lehmann et al. (2011) genes. The same data (sample, genes) were given as input to the TNBCtype tool.

5.4. Results

The results for the BC samples will be presented before presenting the results for the TNBC classifications.

Breast cancer type analysis

BC clustering

The results of the BC samples ($n = 853$) clusterings are shown in Figure 5.1. This figure shows the results of the HAC-Ward clustering (by the dendrogram) and k-means and GGM-EEI clustering by the color bars. As it can be seen in this figure, a large proportion of the TNBC tumors are clustered together by all the three clustering algorithms (indicated by violet, regrouped to the left in the dendrogram, and the homogenous gray scales in the color bars just below). Also, as it can be seen in the dendrogram, the first branch of samples to separate from the others is the one of the TNBC samples. This is an indication on how different these TNBC tumors are in transcriptomic expression patterns compared to the other BC types and stresses out the importance of analyzing them apart. Also, a group of HER(+) samples are recovered (indicated in red) whereas the two luminal types are mixed (indicated in blue and greed).

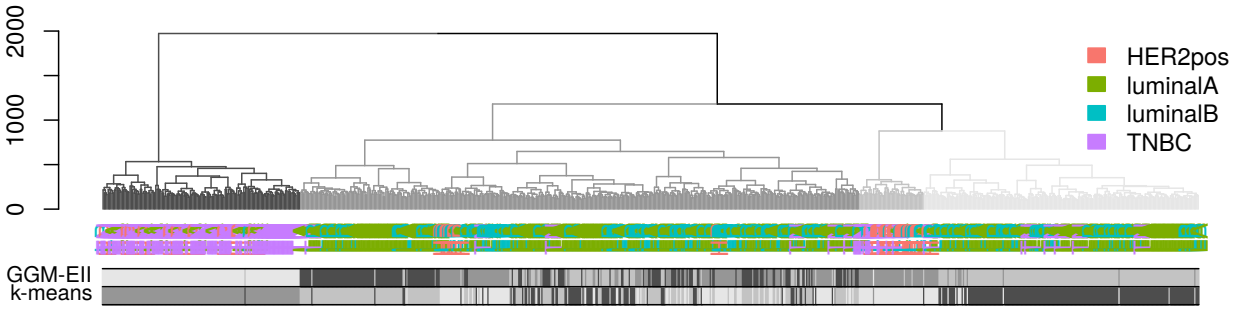


Figure 5.1 – Clustering of the TCGA-BRCA dataset. Color code of dendrogram leaves corresponds to the TCGA breast cancer types. Color code of the dendrogram branches corresponds to HAC-Ward clustering of $K = 4$. Color code of colorbars correspond to clusterings for k-means (bottom) and GMM-EII (top) ($K = 4$).

When the GMM-EEI clustering was compared to the BC IHC classification, it showed that only 126 of the 152 included TNBC tumors were clustered together (the results were similar for the two other clustering algorithms). To avoid including any HER2(+) or luminal like tumors in the following TNBC analyses, the remaining 26 TNBC samples were excluded. I then projected these groups, before and after the exclusion of these TNBC samples into the first two axes of the Principal Component Analysis (PCA) space of RNAseq data. The results are shown in Figure 5.2. As it can be seen in this figure, the samples of the luminal BC types are overlapping but they are distinguished from the HER2(+) separating them from the TNBC samples. These first two axes that represents around 18% of the total variance are therefore, probably, corresponding to a ER/PR/HER2 signature. My hypothesis is that HER2 expression is corresponding to the second axis and ER/PR expression is corresponding to the first axis (based on the distribution of the BC groups). In order to test this, correlations or regression models should be computed between these axes and the ER/PR/HER2 transcripts, which I did not have time to do during my thesis. Figure 5.2A. shows the PCA results before the TNBC sample exclusion and Figure 5.2B. shows the results after the TNBC sample exclusion. Once these samples are excluded, the separation between the TNBC samples and the others BC types becomes clearer. Indeed, the exclusion of these 26 TNBC samples removed those who were overlapping with the HER(+) and luminal types.

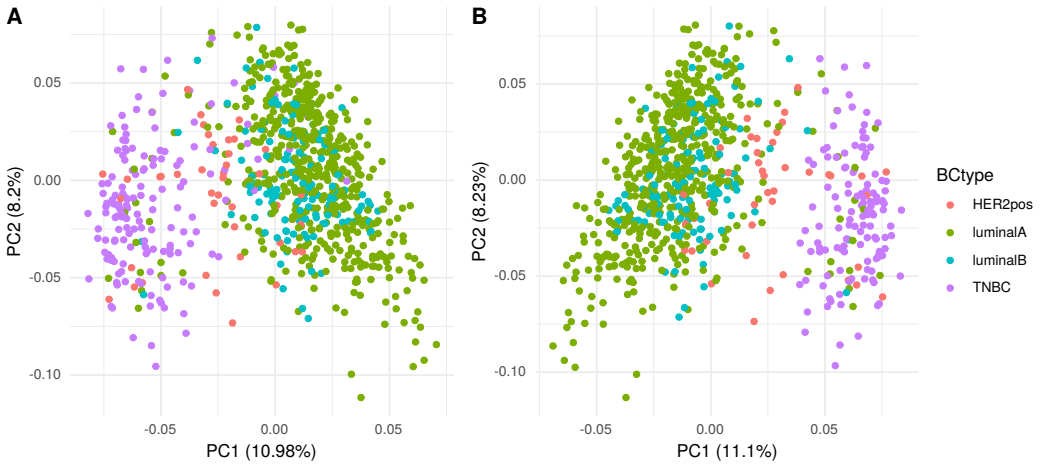


Figure 5.2 – First two dimensions of the PCA space for the TCGA-BRCA RNAseq dataset. The color code indicates the breast cancer type. A. the PCA space include all TNBC samples. B. the PCA space include only TNBC samples clustered together by the GMM-EII. The PCA was implemented with the `prcomp` function of the R-package `stats` (R Core Team, 2016).

Survival study.

The results of the survival analysis for the first 5 years of follow-up are shown in Figure 5.3. There is a significant difference in survival rates between the different groups ($\chi^2 = 12.7$, $df = 3$, $p.value = 0.005$). Pairwise comparisons should be conducted to investigate the origin of this difference. However, it can be noted that the survival curves for the TNBC and the HER(+) groups are lower than the curves for the two luminal groups.

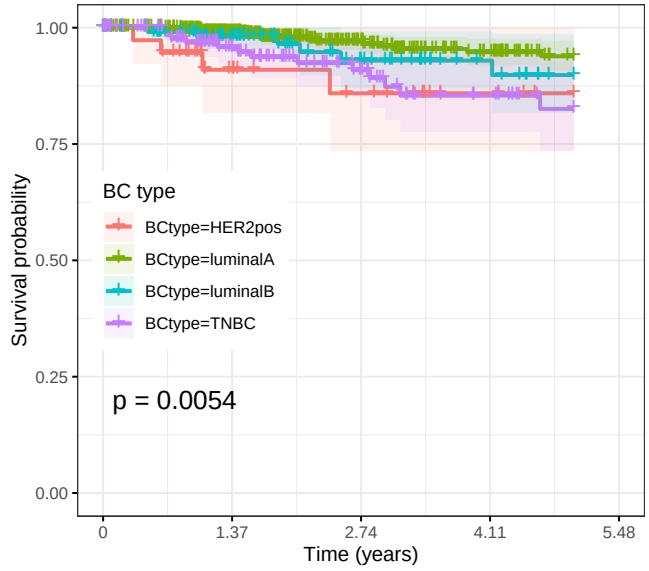


Figure 5.3 – Survival plot for TCGA BC patients, the color code indicates the BC type

Conclusion of BC analysis. In this first section, I analyzed and clustered all the BC tumors based on the most varying transcripts. By doing so I identified a HER(+) group, a TNBC group and a group of luminal (A and B) samples. These groups differ in survival rates. More importantly, I identified a group of $n = 126$ TNBC tumor samples that were very different from the other sample groups in transcriptomic expression (seen by the clustering results and the PCA projection). I will now continue the analysis using only these $n = 126$ TNBC tumor samples.

5.4.1. TNBC classification

In order to classify the included $n = 126$ TNBC tumors I used two strategies, first I predicted their subtypes based on the classification of Lehmann et al. (2011) using the TNBCtype tool, and then I clustered them using unsupervised clustering methods. I then compared the results of these two classifications.

TNBCtype prediction.

The results of the TNBCtype tool predictions are shown in Table 5.2.

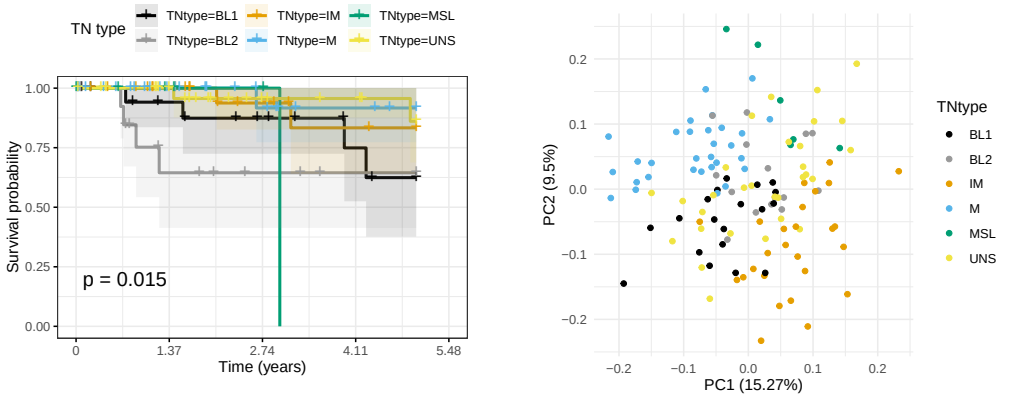
BL1	BL2	IM	M	MSL	UNS
19	15	26	30	6	30

Table 5.2 – TNBCtype tool prediction of the TCGA TNBC tumor samples ($n = 126$) into the subtypes of Lehmann et al. (2011). Subtypes correspond to:basal-like 1 (BL1); basal-like 2 (BL2); immunomodulatory (IM); mesenchymal (M); mesenchymal stem–like (MSL). No sample was predicted as luminal androgen receptor (LAR).

The two largest predicted groups were the unspecified (UNS) and the mesenchymal (M) with $n = 30$ samples each. No sample was predicted as the luminal androgen receptor (LAR) subtype. The results of the survival analysis for the 5 first years of follow up is shown in Figure 5.4a. The two basal like subtypes, and especially BL2, had lower survival curves then the other groups. Also, there was a significant difference in survival rates among the subgroups ($\chi^2 = 14.1$, $df = 5$, $p.value = 0.01$). Pairwise comparisons should be conducted to investigate whether this difference is related only to the BL2 group (who has the lowest survival curve), or to both basal like groups. The groups were projected to the two first axis of the PCA space. The results are shown in Figure 5.4b. As it can be seen in this figure, the Lehmann subtypes are quite well distinguished by these two axis. Indeed, the two large subtypes M and IM are well separated from each other, and a group of MSL can be distinguished.

TNBC clustering

Clustering. The results of the k-means, the HAC-Ward and the GMM-EII cluster algorithms are shown in Figure 5.5. This figure shows the results for the clusterings in $K = 6$ and $K = 2$ groups. First, we clustered the dataset in $K = 6$ groups, since it is the number of Lehamnn subtypes found in this dataset. By doing so, we find (by the HAC-Ward algorithm), a subset of the M samples, a subset of the immunomodulatory (IM) samples, a subset of basal-like 2 (BL2) samples, and two mesenchymal stem–like (MSL) samples, clustered into separate groups. The two basal-like groups and the UNS are clustered together with the remaining samples from the other groups. When the dataset is clustered into $K = 2$ groups, the M and the IM samples are separated from each others, and the rest of the subtypes are mixed.



(a) Survival curves for Lehamann subtypes

(b) 2 first axis of the PCA space

Figure 5.4 – (a) Survival curves estimated by the Kaplan-Meier estimator. p corresponds to the log-rank associated *p.value*. (b) 2 first axes of the PCA space. The color codes indicate Lehmann et al. (2011) subtype.

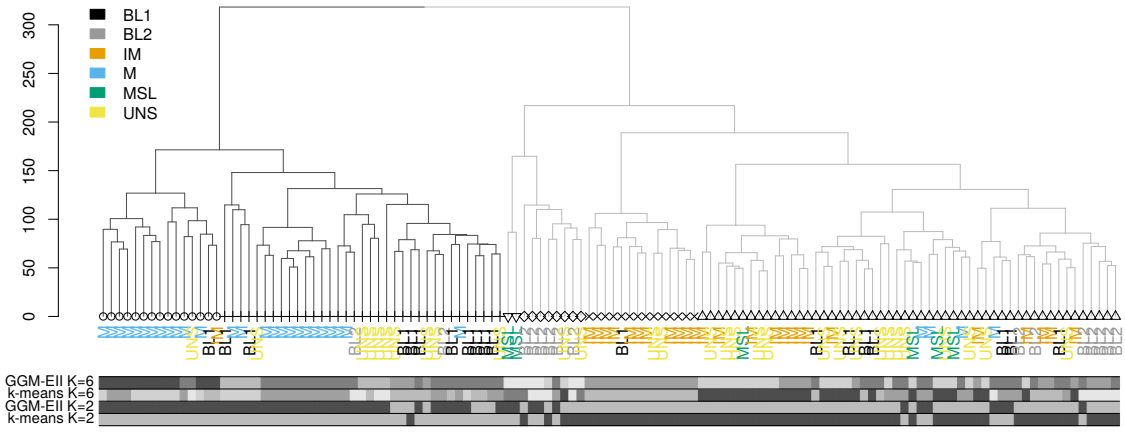


Figure 5.5 – Clustering of TCGA TNBC tumors samples; the dendrogram corresponds to the HAC-Ward algorithm. The classification is based on 2032 genes from the gene signature Lehmann et al. (2011). The three is cut at a level to form $K = 6$ (indicated by the leaves symbols) respective $K = 2$ groups (indicated by the branch colors). The sample labels are colored according to their subgroup of the classification obtained by the TNBCtype tool. Beneath the dendrogram there are four color bars indicating the group belonging for the samples clustered by the GMM-EEI and the k-means algorithm in $K = 6$ (upper bars) and $K = 2$ (lower bars). The Lehmann subtype labels correspond to: basal-like 1 (BL1); basal-like 2 (BL2); immunomodulatory (IM); mesenchymal (M); mesenchymal stem–like (MSL); luminal androgen receptor (LAR); and unstable (UNS).

Selecting the number of groups. In order to estimate the number of groups present in the dataset (without considering the Lehamnn subtypes) several clustering validation criteria were used. The results for these criteria are found in Figure 5.6.

Cluster stability was estimated with the `clustRstab` function using, the k-means algorithm and different proportions of subsampled genes and tumor samples. The cluster comparisons were computed with the $MARI$ score and "random" cluster comparisons were conducted for $nsim = 100$ (this function is described in details in Chapter 4).

The most stable clustering occurs for $K = 2$ groups. This is also the "best" K according to the silhouette average and the Gap statistic. This corresponds to the classification separating the IM samples from the M samples. It can be noticed that these were the two groups that were the most "separated" in the two first axes of the PCA space (see Figure 5.4b). The BIC values of the GMM model are rather indicating $K = 5$ groups and there is a small "peak" in stability for this number of groups too.

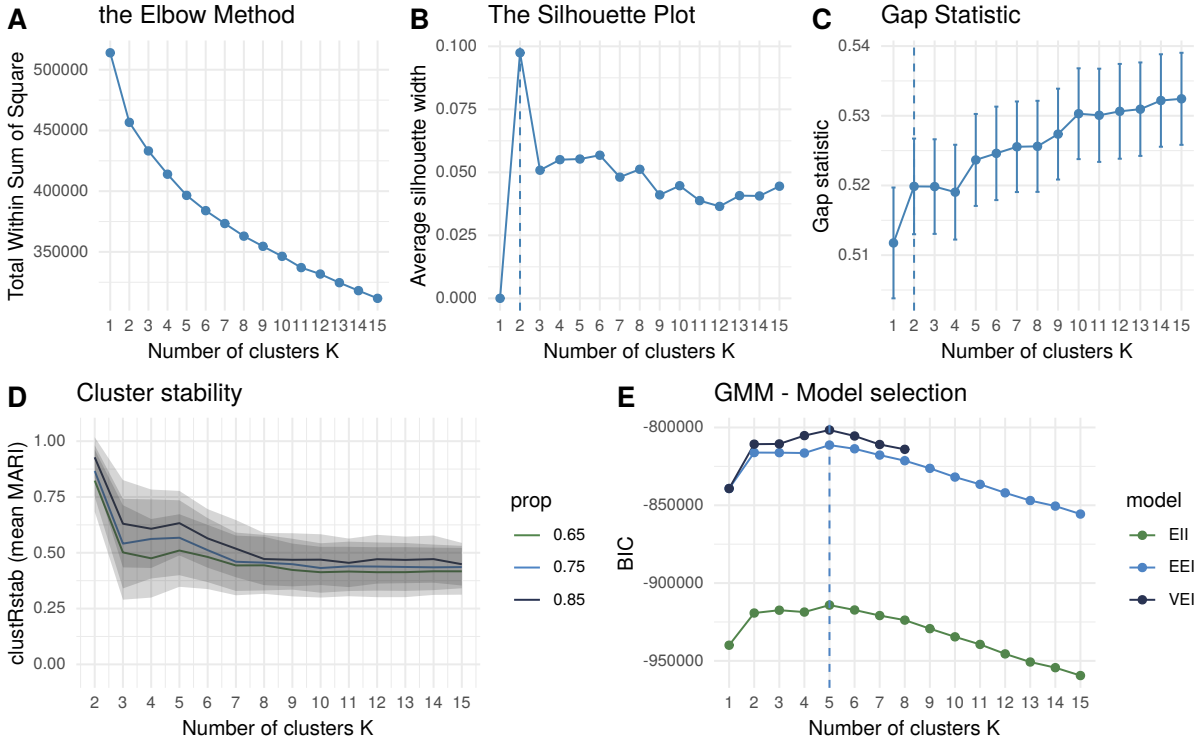


Figure 5.6 – Different methods for selecting the number of groups in the TCGA TNBC dataset. For A, B, C and E: see description of Figure 4.9. D: *prop* corresponds to the proportion of subsampled observations and variables.

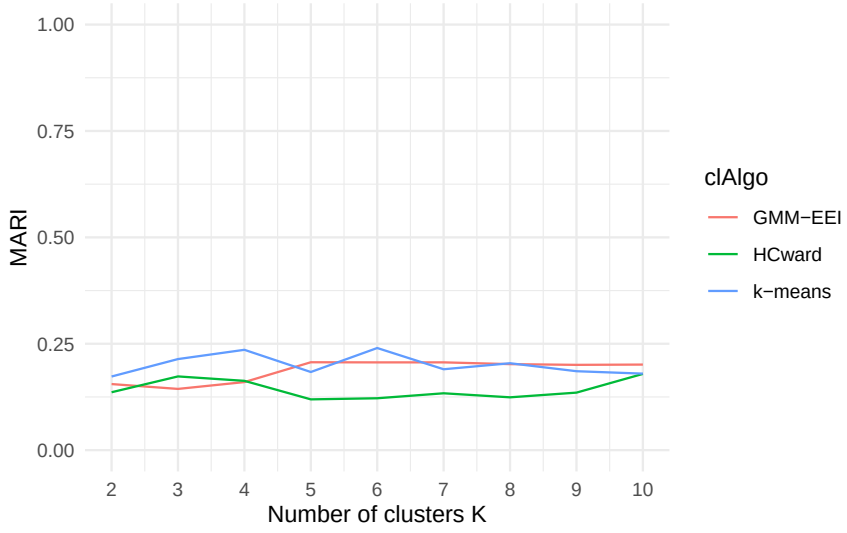


Figure 5.7 – $MARI$ comparison for Lehmann subtypes and different clustering results.

Clustering comparison. In order to see whether $K = 5$ would be a more interesting classification than $K = 2$, I compared the clusterings obtained for different K and the three different clustering algorithms, with the Lehmann subtypes. I did this by computing the $MARI$. For the GMM model, I used the EEI covariance model since it gave the second-best BIC values and converged for all tested K. The results are found in Figure 5.7. As it can be seen, the obtained clustering is quite different from the Lehmann subtypes. Also, it should be noticed that the most stable clustering (obtained for $K = 2$) is not the clustering the most similar to the Lehmann subtypes. Similar results were observed for the NCI60 dataset in described in Chapter 4. This observation, together with the results observed for the clustering criteria, indicates the groups of $K = 5$ or $K = 6$ to be more interesting. Indeed, we saw in Figure 5.5 that the clustering of $K = 6$ groups allowed to distinguish several subsets of different Lehmann subtypes. The contingency tables for the Lehmann subtypes and the k-means clusterings in $K = 2$, $K = 5$ and $K = 6$ groups are shown in Table 5.3.

As can be seen in Table 5.3, the clustering with $K = 6$ groups (obtained with the k-means algorithm) is the one that recovers the most of the Lehmann subtypes. Indeed, two large groups of M samples, as well of IM are found, and a group of BL2 groups are found. This latter was also the group that had a lower survival curve compared to the other groups.

	k-means $K = 5$				
TNtype	1	2	3	4	5
BL1	2	0	3	**12**	2
BL2	1	2	0	2	**10**
IM	**12**	0	0	**14**	0
M	0	0	**11**	0	**19**
MSL	4	2	0	0	0
UNS	**11**	1	4	7	7

	k-means $K = 6$					
TNtype	1	2	3	4	5	6
BL1	2	0	5	8	3	1
BL2	0	1	4	0	0	**10**
IM	**12**	0	0	**14**	0	0
M	0	0	**18**	0	**11**	1
MSL	4	2	0	0	0	0
UNS	**11**	1	8	5	1	4

TNtype		BL1	BL2	IM	M	MSL	UNS
k-means	1	**12**	6	1	**29**	2	11
$K = 2$	2	7	9	**25**	1	4	**19**

Table 5.3 – Contingency Tables for Lehmann subtypes and k-means clustering of TNBC tumor samples. Top: the k-means clustering for $K = 5$ groups (left) and $K = 6$ group (right). Counts > 10 are indicated in bold. The Lehmann subtypes are: basal-like 1 (BL1); basal-like 2 (BL2); immunomodulatory (IM); mesenchymal (M); mesenchymal stem–like (MSL); Unstable (UNS), no luminal androgen receptor (LAR) samples were identified in this dataset.

The clustering in $K = 2$ groups, on the other hand, is also interesting since it identifies withoutsplitting the M and the IM groups. Finally, even though $K = 5$ was the one selected by the BIC criterion (and had a peak in cluster stability), it does not separate as well the Lehamnn subtypes as the clustering in $K = 6$ groups.

Differential analysis should be conducted, both for the Lehmann subtypes and the k-means clusterings, in order to better understand how these different groups are characterized. No significant difference in survival rates was found for any of these k-means clusterings (results not shown).

Conclusion of the TNBC classification study When predicting the Lehmann et al. (2011) subtypes for the $n = 126$ TNBC tumor samples included in this study, 5 of the 6 subtypes were found. No sample was predicted as the LAR subtype. The largest subgroups were M ($n = 30$), UNS ($n = 30$) and IM ($n = 26$). The samples of the BL2 subtype ($n = 15$) had a lower survival rates than the other subgroups and there was a significant difference in survival rates among all the subgroups. Pairwise comparisons should be conducted to determine the origin of this difference. The most stable clustering of the dataset was obtained for $K = 2$ groups. This clustering separated the IM samples from the M samples. However, this clustering was not the most similar clustering compared to the Lehmann subtypes. The clusterings in $K = 5$ and $K = 6$ subgroups were more similar to these subtypes. The clustering in $K = 6$ groups split the M respectively the IM subtype into two different groups each. It also identified a BL2 group.

5.5. Discussion

In this study I classified a large cohort of TNBC tumor samples obtained from the TCGA database based on RNAseq transcriptomic expression. In order to do so, I used two strategies: (1) I predicted the Lehmann et al. (2011) subtypes using the TNBCtype tool (Chen et al., 2012) and (2) I used unsupervised clustering methods and clustering validation criteria that I had developed in this thesis. My goal was to investigate if these two strategies, when using the same data (samples and genes) would result in similar classifications. My hypothesis was that, if that was the case, the observed signal correspond to a biological "robust" signal, present in the TNBC tumors.

The results show that this is the case especially for the IM and the M subtypes that were found both by the TNBCtype tool and by the most stable clustering ($K = 2$). These groups were also the groups that were the most "separated" from the other tumor samples by the first two axes in the PCA space. It should be noticed that these two groups were as well found in the classification of the RATHER-NKI cohort shown in Chapter 2.3.1. It would therefore be interesting to investigate the transcriptomic signal that characterize these groups. A first approach to do this could be to conduct differential analysis by fitting a linear model according to the groups for the different genes. However, Linear Discriminant Analysis (LDA) could also be an alternative in order to find a transcriptomic signature that separates these groups. Also, the $BL2$ subtype was found in the k-means clustering of $K = 6$ groups and showed lower survival rates than the other groups. This is very interesting since it could maybe link this group to a clinical outcome.

Some limits. Some of the limits related to this study are, for example, the fact that I used RNAseq data whereas the TNBCtype tool was based on micro-RNA data. It is possible that this might have induced some noise and explain why the classifications were not "more" similar. Also, the TNBCtype tool is a quite "simple" tool in the sense that it is based on Pearson correlations with the subtypes centroids. It is possible to imagine other more powerful classification methods that would be more suited for the task. For example, using LDA or basing the classification on a probabilistic model (such as the GMM). This latter would make it possibly to predict the outcome by computing the posterior probability (as I did for the RATHER proteomic classification) of the training model. Based on such predicted probability of group belonging, it is easier to see whether the predicted groups make any sense or no. Also, Bonsang-Kitzis et al. (2016) criticized the classification of Lehmann et al. (2011) to be based on a too large number of transcritps ($p = 2188$) and argued that this can induce noise. This research group therefore proposed a transcriptomic signature of $p = 167$ genes which allowed them to propose a TNBC classification of 6 groups. It would be interesting to use their transcriptomic signature for the clustering analysis and see how the results articulate with those obtained in this study. Moreover, it should be noticed that a large number of samples were considered as unspecified ($n = 30$) by the TNBCtype tool. This represents almost a quarter of the 126 TNBC samples included in this study. I believe that this can be an indication that there is other signal among the TNBC samples

that are not captured by the classification of Lehmann et al. (2011). This could in part explain why other groups have had difficulties in reproducing these results. Indeed, they may have captured a biological signal, allowing to classify TNBC tumors, that is different but as important as the one of the Lehmann et al. (2011) classification.

Moreover, these results (as well as those found for RATHER's transcriptomic classification), go against the modification that Lehmann et al. (2016) introduce to their classification. Indeed, they decide to no longer take into account the IM and MSL groups as their transcriptomic singals are linked to the infiltration of white blood cells respective to stromal cells. Their objective was to keep only tumour related signatures. Lehmann et al. (2016) proposed to re-classify the samples of those subgroups by associating them to the subtype with which they have their second-strongest correlation. However, this strategy is not statistically valid. Also, in the classification obtained for the TCGA samples (as for the RATHER classification), this would not make much sense as MI represents one-fifth of the tumors as well as the strongest signal of the classification. Thus, if this group were redistributed to the other subtypes, the resulting classification would no longer correspond to the variation in the dataset. Adding this type of ad-hoc classification criteria will also introduce more noise to the classification, which can induce even more divergences between different studies. For this reason, if one wants to base the classification only on tumour cells, another technique should be used to "remove" these non-tumoral cells. Single cells could be an option, yet, more expensive. However, biologically it is also debatable whether this is a good idea to remove such signals. Indeed, Hida et al. (2019) showed that tumor-infiltrating lymphocytes is a marker for better prognosis and chemotherapeutic effect in TNBC. Li and Dewey (2011) also showed that infiltrating infiltration of cytotoxic T cells into tumors is a critical factor in immunotherapy efficacy. It would therefore be interesting to investigate whether the IM samples identified in this study are linked to such signals.

To conclude, in this study the TNBCtype showed to be a robust tool for TNBC subtyping. Indeed, it found both (potential) biological signal, that was found by the unsupervised clustering methods, especially the identification of the M and the IM group, as well as prognosis prediction, for the BL2 group. This stresses out the utility for such TNBC typing tools and the importance of the classification proposed by Lehmann et al. (2011). However, this method

also has its limits and, for example, a large proportion of the samples were considered as unspecified. In order to fully understand the heterogeneity of the TNBC tumor samples, it would therefore be interesting to compare with other classification results. This will be discussed more in detail in Chapter 6.

Chapter 6

Conclusions and Perspectives

In this thesis I treated the topic of classifying Triple Negative Breast Cancer Tumors (TNBC) from a statistical point of view. In this aim, I mainly focused on clustering and its validation techniques. More precisely, I focused on the use of cluster stability for selecting the number of groups in unsupervised clustering. Indeed, this method has been used in the majority of studies classifying TNBC tumors so far. By doing so, I proposed a proteomic based classification of TNBC and I developed two important methodological contributions that I then applied to a large cohort of TNBC patients. The results of these contributions can be summarized follows:

1. By classifying two large TNBC cohorts based on proteomic RPPA data and refined statistical methods, I identified a classification of two stable groups in the first, but not in the second dataset. There are many potential reasons that can explain why the classification was not validated, but one might be the batch effects that were observed in the RPPA data. These results stress out the importance of conducting rigorous validation procedures when classifying TNBC as well as the importance of using robust normalization procedures. The RPPA technique is still quite "young, and the RPPA community is still quite small, which can explain why tools, allowing to correct these batch effects, do not yet exist. Hopefully, as the RPPA community grows, such normalization issues will no longer be a problem.

2. I improved the Rand Index (Rand, 1971) and its Adjusted version ($(A)RI$, Hubert and Arabie (1985)) by (1) redefining the RI (to MRI, Sundqvist et al. (2020a)) by only counting the consistent pairs by similarity, increasing its interpretability (2) adjusting this score for chance by basing it on a multinomial distribution hypothesis which enables to correctly model the dependent case for the clusterings and conduct statistical inference and (3) propose an efficient algorithm for computing these indices, relying on a sparse representation of the contingency table, implemented in

the `aricode` package (Chiquet et al., 2020).

3. I proposed an R package, `clustRstab`, (Sundqvist et al., 2020b) for easily measuring the stability of a clustering in different parameter settings. This allowed me to conduct a simulation and an application study investigating under which conditions this method can be used as a criterion for selecting the number of groups in unsupervised clustering. The results of these two studies show that (1) cluster stability does not always correctly estimate the number of groups, and this is the case even in simple data settings, and (2) cluster stability does not always allow to identify an interesting clustering. This method was also applied to a large TNBC dataset from the TCGA cohort, and I compared the obtained results with the classification of Lehmann et al. (2011). The results of these studies question the use of cluster stability for clustering such complex data as TNBC tumors and they also stress out the importance of combining cluster stability with other clustering validation criteria.

4. I classified a large TNBC dataset from the TCGA cohort using the clustering validation methods that I developed during my thesis. I compare the results to those obtained by the TNBCtype tool (Chen et al., 2012) corresponding to the classification of Lehmann et al. (2011). These results showed that the Lehmann et al. (2011) classification is important for TNBC subtyping, but not sufficient to take into account all the diversity that exists among the TNBC tumors. Hence, new tools or strategies for TNBC classification would be needed to get a better picture of the TNBC subtypes.

These results will now be discussed in a more global context.

6.1. Is it possible to classify TNBC tumors?

In this thesis I have proposed and discussed the results of different TNBC classifications. Two important observations can be drawn from these results. First, all these classifications have shown a signal in the datasets. This is promising since it probably indicates the presence of a biological signal able to differentiate the TNBC tumors. Also, some gene signatures are found in several of these classifications. For example, the basal like signature was found in both the classification of Lehmann et al. (2011) and the classification of Burstein

et al. (2014), and these two studies, as well as the study of Bonsang-Kitzis et al. (2016), found a subgroup linked to an immunatory gene signature. Second, these classifications differ from each other in the number and the types of groups. As discussed (and shown) in this thesis, this is probably, to a certain degree, due to methodological choices such as the use of different types of omic data, sample cohorts, normalization procedures and statistical methods. However, these divergences are probably also an indication of the complexity and the great heterogeneity among the TNBC tumors. Indeed, it is possible that these different classifications have all captured biological signals present in the TNBC tumors, but not always as apparent. For example, due to the quite small number of samples in TNBC datasets, it might be that some TNBC subgroups are not always represented, or that their gene signature is weaker than the others and therefore, easily gets "hidden" or "embedded" in other, stronger, signals. For example, the LAR subtype of the Lehmann et al. (2011) classification, was not found in the TCGA dataset as shown in Chapter 4. I therefore think that, to understand these classifications of TNBC tumors, it is important to focus both on their similarities but also on their differences.

To do so, the first step would be to understand which of these differences are induced by respectively biological and methodological variations. For this, I think that the best would be to conduct a comprehensive study where the different normalization and classification strategies are tested upon the same dataset and/or conduct the same classification analysis upon all the different classified TNBC cohorts. For this to be possible, all the groups that have proposed a TNBC classification would need to be (willing and) able to share their data and code (which is not always the case due to, for example, patents).

The TNBCtype (Chen et al., 2012) tool is an example of how important it is to enable others to replicate the results of a classification study. Indeed, with this tool, all more recent classification studies have been able to compare their results with the one of Lehmann et al. (2011, 2016). Yet, even though this comparison enables to reveal similarities and differences between classifications, it remains limited since it does not enable to investigate why, or from where, the differences have emerged. Again, having access to the scripts and data of the different studies would help to achieve this understanding.

Also, it would be beneficial for the subtyping of TNBC tumors if negative results could be published. For example, it would be interesting to know whether other groups have tried but, like us, not managed to propose (and validate) a proteomic based classification of the TNBC. Indeed, this could

give some important insights to how proteins and RPPA data can be used for subtyping TNBC and maybe reveal other normalization procedure issues.

Another explanation for the differences in the TNBC classifications might be that non-cancerogen cells induce noise in the tumor sample data. For example, in the study of Lehmann et al. (2016), they showed that two of their six subgroups were correlated with lymphocytes (white blood cells in the immune system) respective with tumor-associated stromal cells. In order to focus on only tumora cells, an idea could be to use single-cell for the classification of TNBC. For example, Wagner et al. (2019) showed promising results for this, by identifying different phenotypic abnormalities and phenotypic dominance in a cohort of 144 breast cancer tumors for which they analyzed transcriptomic single-cell data. They argue that single cell analysis could facilitate the identification of individuals for precision medicine approaches targeting the tumor and its immunoenvironment. Unfortunately, single-cell is an expensive technique and is today mostly used on cell-lines. Also, in view that the TNBC only constitutes approximately 15% of all breast cancers we might have to wait a while before a TNBC single-cell classification can be proposed.

Also, the use of more powerful statistical tools would be beneficial for the understanding of the TNBC classifications. Indeed, in this thesis I showed that cluster stability, the method used so far for classifying TNBC, is a limited method that does not, even in simple settings, allow to select the correct number of groups. Also, I showed that this method is parameter dependent and that a stable clustering is not necessarily interesting. In consequence, if cluster stability shall be used in future TNBC (or other) classification studies, it is important that different stability parameter settings are tested and that the results are compared with other clustering validation criteria, such as the Gap-statistic or the BIC. Moreover, for the further use of cluster stability as a method for selecting the numbers of groups in clustering, it is important to get a better understanding of this method from a theoretical point of view. A first step for this could be to extend the results of Bubeck et al. (2009) on the stability of the k-means algorithm to the finite n case.

Finally, I showed in this thesis that correctly measuring the similarity (or distance) between clusterings is essential for estimating the stability of a clustering. To do so, the cluster comparison score needs to be adjusted for chance. In this thesis I focused on the correction of the Rand Index RI, however, many other different scores exist whereof some might be more suited depending on

the task. For example, the Normalized Information Distance (NID Vinh et al. (2010)), is a score with interesting properties since it is a real distance and metric. However, this score still needs to be adjusted for chance. Nguyen et al. (2009); Vinh et al. (2010) proposed a correction of the Mutual Information (Banerjee et al., 2005) score which, as the NID, is based on the entropy of clusterings. They based this correction on an hypergeometric hypothesis. However, as for the adjustment of the RI this hypothesis is unsatisfying from a statistical modeling point of view. It would therefore be interesting to see if it is possible, as for the $MARI$, to base this correction on a multinomial distribution hypothesis and then extend it to the NID score. Yet, correcting the NID is more difficult than correcting the RI since its expression involve the sum of logarithms for which the expected value is difficult to derive.

6.2. General conclusion

To conclude, clustering is a difficult task for which not one, but several methods exist that yield in more or less different results. Clustering such complex data as TNBC tumors makes the task even harder. Yet, the stakes of this task are highly important. Indeed, identifying subgroups of TNBC tumors could enable the development of specific molecular targeting treatments for the different types of TNBC patients. Therefore, it is important that the research on this topic is efficient and progress quickly. For this reason, I am convinced that transparency about, for example, normalization procedures and the sharing of data and statistical scripts, *i.e.*, "open science", would be beneficial for the advancement of this knowledge.

In this interdisciplinary thesis, I also showed that, studying the TNBC classification from a statistical point can lead to a better understanding of both the biological fundament of the classification as well as a better understanding for statistical methods used. Therefore, I am convinced that the TNBC classification (as well as other biological and statistical issues) would benefit from even more collaborations between statisticians, biologist and computer scientist. For example, if the TNBC data, from the different classification studies, were made public, data challenges could take place where statisticians and computer scientists would be asked to integrate the different datasets and classify the TNBC tumors. This could probably bring in some new creative ideas

about the normalization and statistical analyses of TNBC tumor datasets, and maybe "unblock" the biological results from their computational constraints.

6.2.1. Conclusion générale (français)

Pour conclure, le clustering est une tâche difficile pour laquelle il n'existe pas une, mais plusieurs méthodes qui donnent des résultats plus ou moins différents. Le clustering de données aussi complexes que les tumeurs TNBC rend la tâche encore plus difficile. Pourtant, les enjeux de cette tâche sont très importants. En effet, l'identification de sous-groupes de tumeurs TNBC pourrait permettre le développement de traitements moléculaires ciblés spécifiques pour les différents types de patients de TNBC. Il est donc important que la recherche sur ce sujet soit efficace et progresse rapidement. C'est pourquoi je suis convaincu que la transparence concernant, par exemple, les procédures de normalisation et le partage des données et des scripts statistiques, *i.e.*, "open science", serait bénéfique pour l'avancement de ces connaissances.

Dans cette thèse interdisciplinaire, j'ai également montré que le fait d'étudier la classification TNBC d'un point de vue statistique peut conduire à une meilleure compréhension des fondements biologiques de la classification ainsi qu'à une meilleure compréhension des méthodes statistiques utilisées. Par conséquent, je suis convaincu que la classification TNBC (ainsi que d'autres questions biologiques et statistiques) bénéficierait d'encore plus de collaborations entre statisticiens, biologistes et informaticiens. Par exemple, si les données du TNBC, issues des différentes études de classification, étaient rendues publiques, il pourrait y avoir des data challanges où les statisticiens et les informaticiens seraient invités à intégrer les différents ensembles de données et à classer les tumeurs TNBC. Cela pourrait probablement apporter de nouvelles idées créatives sur la normalisation et les analyses statistiques des jeux de données des tumeurs TNBC, et peut-être "débloquer" les résultats biologiques de leurs contraintes de méthodologiques.

6.3. And a last word...

On ne gagne pas beaucoup à courir le monde.
- Proverbe suisse

Working "in-between" statistics and biology in this thesis has sometimes been challenging. Indeed, the actors of these two domains do not always understand each other. For example, I have noticed that, on one side, biologists sometimes tend to rush towards the biological interpretation of the results (without sufficiently taking into consideration all the necessary statistical adjustments), and on the other side, statisticians sometimes tend to forget the practical use of their models by searching to develop new, preferably more complex and especially faster methods. As a consequence, everything in between, as for example, the preprocessing of the data, easily gets lost, even if it is a critical part of the analysis. Hence, working in between these domains has often required me to be very patient, to use a lot of pedagogy and to question myself and my knowledge. However, by working in this interdisciplinary context, I got the opportunity to study the topic of my thesis from various perspectives. This has been very enriching and rewarding as it, with the help of my supervisors, allowed me to develop my own vision of the topic. As a consequence, I managed to detect pitfalls concerning the classification of TNBC in both the (1) statistical methods and their applications, as well as, in the (2) data normalization procedures and the biological interpretations of the statistical results. My work in this thesis has therefore been more oriented towards the replication and understanding of scientific results (in both biology and statistics) than towards the production of new results or new methods. This might seem quite odd since the scientific world of today is mostly turned towards what is new and by definition publishable. I therefore hope that my thesis can state the example that scientific reproducibility is needed and can, as well as searching for something new, allow research and science to evolve.

Bibliography

Rehan Akbani, Karl-Friedrich Becker, Neil Carragher, Ted Goldstein, Leanne de Koning, Ulrike Korf, Lance Liotta, Gordon B Mills, Satoshi S Nishizuka, Michael Pawlak, et al. Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: A workshop report the rppa (reverse phase protein array) society. *Molecular & cellular proteomics*, 13 (7):1625–1643, 2014.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Sunil Badve, David J Dabbs, Stuart J Schnitt, Frederick L Baehner, Thomas Decker, Vincenzo Eusebi, Stephen B Fox, Shu Ichihara, Jocelyne Jacquemier, Sunil R Lakhani, et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Modern Pathology*, 24(2):157–167, 2011.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005.

Shai Ben-David, Ulrike Von Luxburg, and Dávid Pál. A sober look at clustering stability. In *International Conference on Computational Learning Theory*, pages 5–19. Springer, 2006.

Shai Ben-David, Dávid Pál, and Hans Ulrich Simon. Stability of k-means clustering. In *International conference on computational learning theory*, pages 20–34. Springer, 2007.

Asa Ben-Hur and Isabelle Guyon. Detecting stable clusters using principal component analysis. In *Functional genomics*, pages 159–182. Springer, 2003.

Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Biocomputing 2002*, pages 6–17. World Scientific, 2001.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Marco Bertoletti, Nial Friel, and Riccardo Rastelli. Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron*, 73(2):177–199, 2015.

Giampaolo Bianchini, Justin M Balko, Ingrid A Mayer, Melinda E Sanders, and Luca Gianni. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. *Nature Reviews Clinical Oncology*, 2016.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

Stefanie Boellner and Karl-Friedrich Becker. Reverse phase protein arrays—quantitative assessment of multiple biomarkers in biopsies for clinical use. *Microarrays*, 4(2):98–114, 2015.

H Bonsang-Kitzis, B Sadacca, AS Hamy-Petit, M Moarii, A Pinheiro, C Laurent, and F Reyal. Biological network-driven gene selection identifies a stromal immune module as a key determinant of triple-negative breast carcinoma prognosis. *Oncoimmunology*, 5(1):e1061176, 2016.

Robert L Brennan and Richard J Light. Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 27(2):154–163, 1974.

Sébastien Bubeck, Marina Meila, and Ulrike von Luxburg. How the initialization affects the stability of the k-means algorithm. *arXiv preprint arXiv:0907.5494*, 2009.

Matthew D Burstein, Anna Tsimelzon, Graham M Poage, Kyle R Covington, Alejandro Contreras, Suzanne AW Fuqua, Michelle I Savage, C Kent Osborne, Susan G Hilsenbeck, Jenny C Chang, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 2014.

Adam Byron. Reproducibility and crossplatform validation of reverse-phase protein array data. In *Reverse Phase Protein Arrays*, pages 181–201. Springer, 2019.

Adam Byron, Stephan Bernhardt, Bérèngere Ouine, Aurélie Cartier, Kenneth G Macleod, Neil O Carragher, Vonick Sibut, Ulrike Korf, Bryan Serrels, and Leanne de Koning. Integrative analysis of multi-platform reverse-phase protein array data for the pharmacodynamic assessment of response to targeted therapies. *bioRxiv*, page 769158, 2019.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *arXiv preprint arXiv:1504.04595*, 2015.

Timothy I. Cannings and Richard J. Samworth. *RPEnsemble: Random Projection Ensemble Classification*, 2017. URL `https://CRAN.R-project.org/package=RPEnsemble`. R package version 0.4.

Xi Chen, Jiang Li, William H Gray, Brian D Lehmann, Joshua A Bauer, Yu Shyr, and Jennifer A Pietenpol. Tnbctype: a subtyping tool for triple-negative breast cancer. *Cancer informatics*, 11:CIN–S9983, 2012.

Julien Chiquet, Guillem Rigaill, and Martina Sundqvist. *aricode: Efficient Computations of Standard Clustering Comparison Measures*, 2020. URL `https://CRAN.R-project.org/package=aricode`. R package version 1.0.0.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 2 edition, 2001.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Dirk Eddelbuettel, Romain François, J Allaire, Kevin Ushey, Qiang Kou, N Russel, John Chambers, and D Bates. Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

Aurélie Fischer. On the number of groups in clustering. *Statistics & Probability Letters*, 81(12):1771–1781, 2011.

William D Foulkes, Ian E Smith, and Jorge S Reis-Filho. Triple-negative breast cancer. *New England journal of medicine*, 363(20):1938–1948, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Filippo G Giancotti. Deregulation of cell signaling in cancer. *FEBS letters*, 588(16):2558–2570, 2014.

Philip D Glaves and Jonathan D Tugwood. Generation and analysis of transcriptomics data. In *Drug Safety Evaluation*, pages 167–185. Springer, 2011.

Nadia Harbeck and Michael Gnant. Breast cancer. *Lancet*, 389(10074):1134–1150, 2017.

David P Harrington and Thomas R Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. *impute: impute: Imputation for microarray data*, 2016. R package version 1.46.0.

Bryan T Hennessy, Yiling Lu, Ana Maria Gonzalez-Angulo, Mark S Carey, Simen Myhre, Zhenlin Ju, Michael A Davies, Wenbin Liu, Kevin Coombes, Funda Meric-Bernstam, et al. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clinical proteomics*, 6(4):129, 2010.

Christian Hennig. *fpc: Flexible Procedures for Clustering*, 2020. URL `https://CRAN.R-project.org/package=fpc`. R package version 2.2-8.

Akira I Hida, Takahiro Watanabe, Yasuaki Sagara, Masahiro Kashiwaba, Yoshiaki Sagara, Kenjiro Aogi, Yasuyo Ohi, and Akihide Tanimoto. Diffuse distribution of tumor-infiltrating lymphocytes is a marker for better prognosis and chemotherapeutic effect in triple-negative breast cancer. *Breast cancer research and treatment*, 178(2):283–294, 2019.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Yi-Zhou Jiang, Ding Ma, Chen Suo, Jinxiu Shi, Mengzhu Xue, Xin Hu, Yi Xiao, Ke-Da Yu, Yi-Rong Liu, Ying Yu, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer cell*, 35(3):428–440, 2019.

Louise N Johnson. The regulation of protein phosphorylation. *Biochemical Society Transactions*, 37(4):627–641, 2009.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Alboukadel Kassambara. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. STHDA, 2017.

Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2017. URL `https://CRAN.R-project.org/package=factoextra`. R package version 1.0.5.

Alboukadel Kassambara, Marcin Kosinski, and Przemyslaw Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2020. URL `https://CRAN.R-project.org/package=survminer`. R package version 0.4.8.

E Lebarbier and T Mary-Huard. Classification non supervisée, 2008.

Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750–2767, 2011.

Brian D Lehmann, Bojana Jovanović, Xi Chen, Monica V Estrada, Kimberly N Johnson, Yu Shyr, Harold L Moses, Melinda E Sanders, and Jennifer A Pietenpol. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PLoS One*, 11(6): e0157368, 2016.

Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity. *Neural computation*, 13(11):2573–2593, 2001.

Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1): 323, 2011.

Sabine C Linn and Laura J Van't Veer. Clinical relevance of the triple-negative breast cancer concept: genetic basis and clinical utility of the concept. *European Journal of Cancer*, 45:11–26, 2009.

Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

Minetta C Liu, Brandelyn N Pitcher, Elaine R Mardis, Sherri R Davies, Paula N Friedman, Jacqueline E Snider, Tammi L Vickery, Jerry P Reed, Katherine DeSchryver, Baljit Singh, et al. Pam50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of c9741 (alliance). *NPJ Breast Cancer*, 2(1):1–8, 2016.

Etienne Lord, Matthieu Willems, Francois-Joseph Lapointe, and Vladimir Makarenkov. *ClusterStability: Assessment of Stability of Individual Objects or Clusters in Partitioning Solutions*, 2016. URL `https://CRAN.R-project.org/package=ClusterStability`. R package version 1.0.3.

Etienne Lord, Matthieu Willems, Francois-Joseph Lapointe, and Vladimir Makarenkov. Using the stability of objects to determine the number of clusters in datasets. *Information Sciences*, 393:29–46, 2017.

James W. MacDonald, Debashis Ghosh, and Mark Smolkin. *clusterStab: Compute cluster stability scores for microarray data*, 2018. R package version 1.54.0.

Hiroko Masuda, Yuan Qi, Shuying Liu, Naoki Hayashi, Takahiro Kogawa, Gabriel N Hortobagyi, Debu Tripathy, and Naoto T Ueno. Reverse phase protein array identification of triple-negative breast cancer subtypes and comparison with mrna molecular subtypes. *Oncotarget*, 8(41):70481, 2017.

Sylvie Maubant, Bruno Tesson, Virginie Maire, Mengliang Ye, Guillem Rigaill, David Gentien, Francisco Cruzalegui, Gordon C Tucker, Sergio Roman-Roman, and Thierry Dubois. Transcriptome analysis of wnt3a-treated triple-negative breast cancer cells. *PloS one*, 10(4):e0122333, 2015.

Marina Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003.

Magali Michaut, Suet-Feung Chin, Ian Majewski, Tesa M Severson, Tycho Bismeijer, Leanne de Koning, Justine K Peeters, Philip C Schouten, Oscar M Rueda, Astrid J Bosma, et al. Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific reports*, 6:18517, 2016.

Ulrich Möller and Dörte Radke. Performance of data resampling methods for robust class discovery based on clustering. *Intelligent Data Analysis*, 10(2): 139–162, 2006.

Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

Leslie C Morey and Alan Agresti. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.

Fionn Murtagh and Pierre Legendre. Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm. *arXiv preprint arXiv:1111.6285*, 2011.

E Shannon Neeley, Keith A Baggerly, and Steven M Kornblau. Surface adjustment of reverse phase protein arrays using positive control spots. *Cancer informatics*, 11:CIN–S9055, 2012.

Xuan Vinh Nguyen, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *ICML*, 2009.

Lukasz Nieweglowski. *clv: Cluster Validation Techniques*, 2020. URL `https://CRAN.R-project.org/package=clv`. R package version 0.3-2.2.

Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, 406(6797):747–752, 2000.

Franck Picard. An introduction to mixture models. *Statistics for Systems Biology Group. Research Report*, (7), 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019a. URL `https://www.R-project.org/`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019b. URL `https://www.R-project.org/`.

W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7): e47, 2015. doi: 10.1093/nar/gkv007.

Douglas T Ross, Uwe Scherf, Michael B Eisen, Charles M Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S Jeffrey, Matt Van de Rijn, Mark Waltham, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics*, 24(3):227–235, 2000.

Peter J Rousseeuw and L Kaufman. Finding groups in data. *Hoboken: Wiley Online Library*, 1990.

Dorota Rozmus. Using r packages for comparison of cluster stability. *Acta Universitatis Lodziensis. Folia Oeconomica*, 4(330), 2017.

Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016a. URL `https://doi.org/10.32614/RJ-2016-021`.

Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal*, 8(1):289, 2016b.

Tesa M Severson, Justine Peeters, Ian Majewski, Magali Michaut, Astrid Bosma, Philip C Schouten, Suet-Feung Chin, Bernard Pereira, Mae A Goldgraben, Tycho Bismeijer, et al. Brca1-like signature in triple negative breast cancer: Molecular and clinical characterization reveals subgroups with therapeutic potential. *Molecular oncology*, 9(8):1528–1538, 2015.

Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In *Advances in neural information processing systems*, pages 1297–1304, 2008a.

Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. In *COLT*, pages 367–378. Citeseer, 2008b.

Ohad Shamir and Naftali Tishby. On the reliability of clustering stability in the large sample regime. In *Advances in neural information processing systems*, pages 1465–1472, 2009.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. URL `http://www.jstatsoft.org/v39/i05/`.

Mark Smolkin and Debashis Ghosh. Cluster stability scores for microarray data in cancer studies. *BMC bioinformatics*, 4(1):36, 2003.

Dong Song, Miao Cui, Gang Zhao, Zhimin Fan, Katherine Nolan, Ying Yang, Peng Lee, Fei Ye, and David Y Zhang. Pathway-based analysis of breast cancer. *American journal of translational research*, 6(3):302, 2014.

Therese Sorlie, Robert Tibshirani, Joel Parker, Trevor Hastie, JS Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the united States of America*, 100(14):8418–8423, 2003.

Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386, 2004.

Douglas Steinley and Michael J Brusco. A note on the expected value of the rand index. *British Journal of Mathematical and Statistical Psychology*, 71 (2):287–299, 2018.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

Martina Sundqvist, Julien Chiquet, and Rigaill Guillem. Adjusting the adjusted rand index - a multinomial story. *Computational Statistics*, 2020a. Submitted.

Martina Sundqvist, Julien Chiquet, and Guillem Rigaill. *clustRstab: Flexible estimation of clustering stability for class discovery*, 2020b. URL `https://github.com/MartinaSundqvist/clustRstab`. R package.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL `https://CRAN.R-project.org/package=survival`. version 2.38.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Eleonora Timperi, Mengliang Ye, Thierry Dubois, Didier Meseure, Anne Vincent Salomon, and Emanuela Romano. Wnt/β–catenin pathway activation correlates with the increase of tumor-associated macrophages in triple negative breast cancer (tnbc)., 2020.

Sylvie Troncale, Aurélie Barbet, Lamine Coulibaly, Emilie Henry, Beilei He, Emmanuel Barillot, Thierry Dubois, Philippe Hupé, and Leanne De Koning. Normacurve: a supercurve-based method that simultaneously quantifies and normalizes reverse phase protein array data. *PloS one*, 7(6):e38686, 2012.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct): 2837–2854, 2010.

Ulrike Von Luxburg and Shai Ben-David. Towards a statistical theory of clustering. In *Pascal workshop on statistics and optimization of clustering*, pages 20–26. Citeseer, 2005.

Ulrike Von Luxburg et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274, 2010.

Johanna Wagner, Maria Anna Rapsomaniki, Stéphane Chevrier, Tobias Anzeneder, Claus Langwieder, August Dykgers, Martin Rees, Annette Ramaswamy, Simone Muenst, Savas Deniz Soysal, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, 177(5): 1330–1345, 2019.

Silke Wagner and Dorothea Wagner. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

Wilkerson, Matthew D., Hayes, and D. Neil. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010. URL `http://bioinformatics.oxfordjournals.org/content/26/12/1572.abstract`.

Yasin Şenbabaoğlu, George Michailidis, and Jun Z Li. Critical limitations of consensus clustering in class discovery. *Scientific reports*, 4(1):1–13, 2014.

Appendix A

TTKi for molecular drug discovery in TNBC

A.1. General context

During my thesis I participated in a collaboration with an industrial and the lab of Thierry Dubois at Institut Curie, hereon referred to as the BCBG lab (for Breast Cancer Biology Group). During this project I worked in close collaboration with Amelie Brisson and Clarisse Monchecourt, two biologists from the BCBG lab. The data of this project are confidential and the results will therefore only be discussed in general terms. In this report I will discuss the analysis pipeline I put in place and the linear model I defined in order to conduct these analyses.

A.2. Motivation

This project was conducted with the aim of finding molecules that could be used in combination with a TTK inhibitor (TTKi) to treat TN patients. It was a collaboration with an industrial, currently conducting clinical trials on a TTKi for TN patients.

The kinase TTK plays a role in cell proliferation. Indeed, it verifies whether the chromosomes are aligned before mitosis of a cell. Blocking TTK in cancer could reduce cell proliferation and induce cellular death. In practise, some cancer respond well, but not all. In the later case a subpopulation of cells resists to the treatment, and the proportion of residual cell differ from one patient to the next. Similar results have been observed for cell lines in the BCBG lab and the proportion of their residual cell population, when treated with TTKi, are found in Figure A.1. As it can be seen, for some of these cell lines, almost all cells are dead. These cell lines are indicated by the green bars and are considered as highly sensitive to TTKi. On the contrary, for some of

these cell lines, the proportion of the residual population remains large after TTKi treatment. These cell lines are considered as insensitive to TTKi and are indicated by the orange bars. The questions are then, why do some of the cells resist to TTKi treatment and why are some cell lines less sensitive to the treatment than others? The aim of the present project was therefore to search for molecule candidates that could be combined with TTKi treatment to bypass this resistance and kill the remaining residual population.

Testing genes at random is tedious and extremely costly since there are over 20 000 genes to consider. To reduce this list of genes, a differential gene and protein expression analysis was performed. The idea being that, genes or proteins that are highly over or under expressed in the resistant cell-lines compared to the sensitive cell lines are potential "protectors" for TTKi treatment. Inhibiting these genes or proteins could therefore be a manner to induce a larger effect of TTKi and kill the remaining residual cell population.

In summary, in order to implement this idea, cell-lines were culture three times (to get three biological replicates) and for each replicate some cells were treated with TTKi and some not. An RNA microarray and RPPA experiment was conducted on these samples. I then extracted the RPPA and RNA data and computed a *log-fold change* measure for each cell line, biological replicate and gene respective protein. This *log fold change* corresponded to the difference in gene (or protein) expression observed for a given biological replicate when the cells were treated with TTKi compared to when they were not. Differential analysis were then conducted and different contrasts were tested in order to reveal the genes (or proteins) that were highly, over or under, expressed in the sensitive versus the insensitive cell lines.

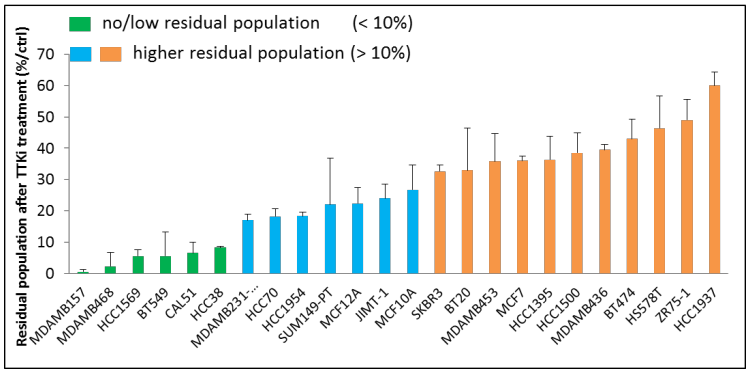


Figure A.1 – Residual population for cell lines when treated with TTK inhibitor. Color code indicates importance of the residual population. Cell lines are ordered according to the size of their residual population. Figure produced at the lab of Thierry Dubois (BCBG-lab).

A.3. Methods

A.3.1. Experimental design

Cell culture. Cell culture was conducted in the BCBG lab for 19 different cell lines. Each cell line was cultured in three biological replicates. Each replicate was split into two samples. One of these samples received a TTKi treatment, diluted in DMSO, whereas the other sample only received the DMSO dilution for control. The treatment was left for 7 days (corresponding to approximately 4 cell cycles depending on the cell line). In total there was hence $n = 114$ samples.

Data extraction. After these 7 days of treatment, transcriptomic and proteomic data were extracted for each sample. Transcriptomic (RNA) data were extracted by Clariom S-Affymetrix at the Genomic platform at Institut Curie. I normalized the data by the following standardized procedures: Robust Multichip Average (RMA) normalization, local – array – summarization, by gene name, and I then log_2 transformed the data. RPPA data were analysed at the RPPA platform at Institut Curie for 180 antibodies measuring proteomic expression and activation (phosphorylation).

A.3.2. Statistical Model

The *log fold change* measure $\Delta_{i,k,g}$. For each biological replicate and gene (or protein) a *log fold change* measure was defined as follows

$$\Delta_{i,k,g} = log_2\left(\frac{x_{i,k,g}^{t=2}}{x_{i,k,g}^{t=1}}\right),$$

where $x_{i,k,g}^{t}$ is the gene expression value for a given gene g (or protein), a given cell line i, with $i = 1, \ldots, 19$, and a given replicate k, with k = 1,2,3, t indicates whether the sample has received TTKi treatment, $t = 2$ or not $t = 1$. The higher the $\Delta_{i,k,g}$, the more is a given gene (or protein) over expressed in a biological replicate when this latter is treated with TTKi. Similarly, the smaller the $\Delta_{i,k,g}$, the more is a given gene (or protein) under expressed in a biological replicate when this latter is treated by TTKi.

Linear model. The *log fold change* value, $\Delta_{i,k,g}$, for each biological replicate, cell line and gene (or protein) was then modeled as a linear combination of the cell line effect as follows,

$$\Delta_{i,k,g} = \mu_i + \epsilon_{i,k,g}$$

where μ_i indicates the mean value of the log_2 transformed values of the cell line i, $\epsilon_{i,k,g}$ indicates the residual error terms with $\epsilon_{i,k,g} \sim \mathcal{N}(0,\sigma^2)$ and σ^2 the variance of $\Delta_{i,k,g}$. The intercept of this model was fixed to 0.

This linear model allowed us test different contrasts $C' = \sum_{i=1}^{19} a_i\mu_i$, where a_i is a constant defined such that $\sum_{i=1}^{19} a_i = 0$. The values of these constants were then modified according to the contrast C' that we wanted to test. For example, when testing the difference in gene (or protein) expression for the sensitive versus the insensitive cell lines the values of a_i were defined as follows:

- $a_i = \frac{1}{n_{is}}$ for $\forall i \in \{insensitive\ cell\ lines\}$, where n_{is} is the cardinal of the subset of insensitive cell lines,
- $a_i = \frac{-1}{n_s}$ for $\forall i \in \{sensitive\ cell\ lines\}$, where n_s is the cardinal of the subset of sensitive cell lines,
- all other a_i are set to 0.

The contrast was then tested by the `limma` package.

Statistical analysis. To implement the linear model and conduct the differential analysis we used the R-package `limma` (Ritchie et al., 2015). This package is specially designed for linear models and differential expression for microarray data.

The significance level was set at $\alpha = 0.05$ and p-values were corrected for multiple testing using the Benjamin-Hochberg false discovery rate correction Benjamini and Hochberg (1995).

Pathway analyses were conducted for the most significantly expressed genes. I did this by conducting hypergeometric tests, testing gene enrichment, for different pathways using different pathway databases (GO, Kegg, Reactome).

A.4. Results

As already said in the introduction, the results of this project are confidential and will only be discussed in general terms.

A.5. RPPA

Using the `limma` package few proteins were found to be differentially expressed for the different tested contrasts.

The BCBG lab decided to test these candidates. Clarisse and Amelie seeked to validate the differential expression between sensitive and resistant cell-lines using Western Blot. Their Western Blot results were discordant with the RPPA results. We thus tried to understand the reason for this discordance. Our in-depth analysis of the data revealed some issues in the normalization procedure commonly used for RPPA data. As there was no quick or easy fix to this we decided to disregard protein data and move on to the analyses of the transcriptome.

A.5.1. RNA

Exploratory analysis Figure A.2 shows the observations, before the $\Delta_{i,k,g}$ transformation, projected into the two first axes of the PCA space of this dataset. As it can be seen on the right (coloring observation according to the cell-line), in the upper panel, a strong cell line effect was observed on the data. Hence, no treatment effect was revealed for the raw data when projected on these two axes. The bottom panel shows the data projected into the two first

axes of the PCA space when the gene expression values of each replicate had been corrected by the mean effect of its cell line, that is $log_2(x^t_{i,k,g}) - \mu_i$, where $x^t_{i,k,g}$ is the gene expression of a given sample, and μ_i is the log value of the mean-value of the cell line of the sample. As can be seen, once the data were corrected for this cell line effect, the experimental condition became visible.

I then studied the correlation matrix of the $\Delta_{i,k,g}$ values (measuring the difference between TTKi and control for each replicate) in order to verify that the effect of the cell lines was corrected. Ideally, the $\Delta_{i,k,g}$ for the differnet replicates of a cell line should no longer be correlated.

As can be seen in Figure A.3, for some cell lines, all or some, replicates remained strongly correlated. After discussing with Thierry, Amélie and Clarisse, we realized that all these replicates had proliferation issues during the cell culture. We believe this is the reason for this high correlation and we excluded these samples from the analysis.

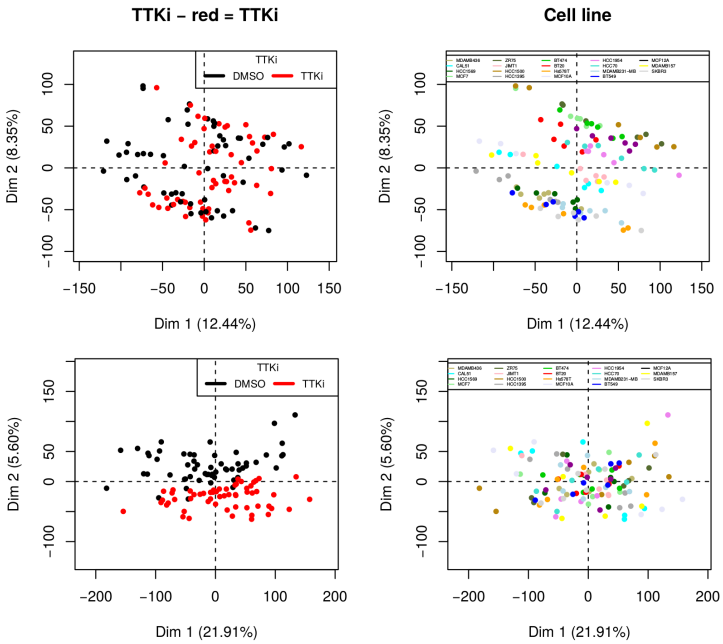


Figure A.2 – First two axes of RNA PCA space. Upper panel: Raw data. Bottom panel: data corrected for the cell line effect. Color code for left panel: DMSO/TTKi treatment. Color code for right panel: Cell line.

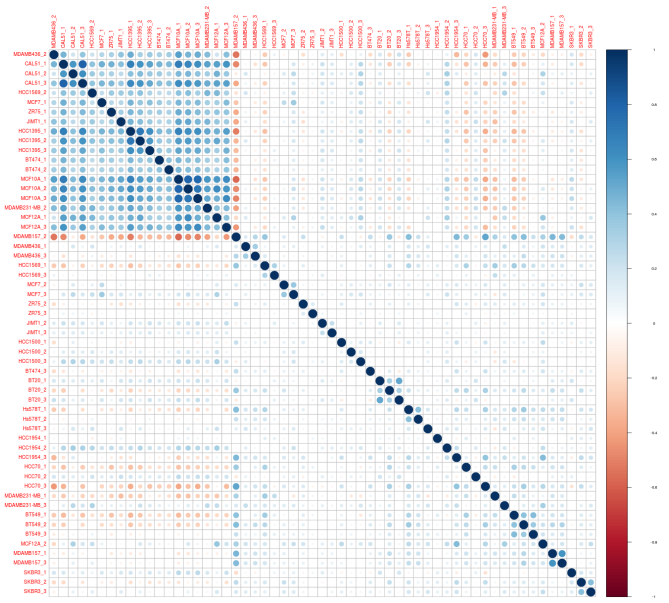


Figure A.3 – Pearson Correlation Coefficient matrix for RNA $\Delta_{i,k,g}$ values of cell line replicates. The replicates of each cell-line are organized after one another. The color code indicates the strength and the sign of the correlation coefficient.

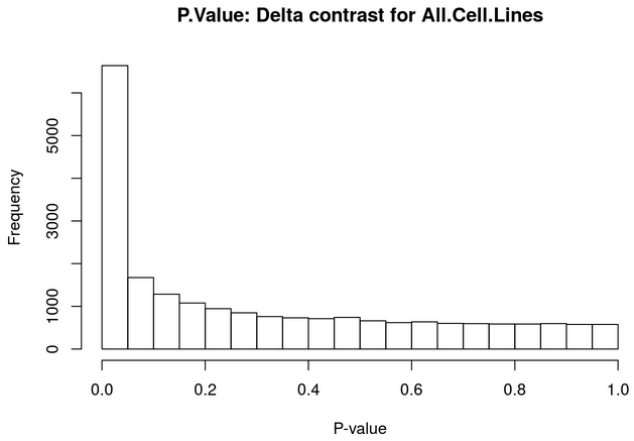


Figure A.4 – Uncorrected *p.values* for $\Delta_{i,k,g}$ values associated to the differential analysis including all cell lines.

Results for the linear model. The uncorrected *p.values* associated to the contrast for all cell lines are shown in Figure A.4. As it can be seen, a peak of $p - values$ were found for $p.values \in [0 : 0.05]$, indicating that there is an enrichment for differentially expressed genes. Also, we observed for all tested

contrast that the histogram is flat for larger p-values which is reassuring from a modeling perspective. We observed that the genes with the smallest $p-values$ were similar for the different tested contrasts. This indicates that the results are revealing a more general biological process that is involved. The differential analysis (not-described here) pinpointed two molecules that have since then been tested biologically in the BCBG lab in combination, or not, with TTKi.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Stabilité et sélection du nombre de groupes en clustering non-supervisé: application à laclassification des cancers du sein triple négatifs

Mots clés: Cancer du sein triple négatif, Classification non supervisée, Omique, Stabilité des clusters, Rand Index

Résumé: Dans cette thèse, je traite, d'un point de vue statistique, le sujet de la classification des tumeurs du cancer du sein triple négatif (TNBC). Je me concentre principalement sur l'utilisation de la stabilité des clusters pour sélectionner le nombre de groupes dans le clustering, la méthode généralement utilisée pour la classification des TNBC. L'objectif de cette méthode est d'obtenir une classification robuste, c'est-à-dire facilement reproductible sur des données similaires. Malgré sa popularité, on sait encore peu de choses sur la façon dont cette méthode fonctionne. Pour cette raison, je propose deux contributions méthodologiques importantes : (1) un package R, `clustRstab`, qui permet d'estimer, de manière flexible, la stabilité d'un clustering avec différents paramètres. Ce package est accompagné d'une étude de simulation et d'une étude d'application qui examine sous quelles conditions cette méthode fonctionne. (2) Une version modifiée de la version Ajusté du Rand Index (ARI), un score populaire pour les comparaisons de clusters, étape cruciale pour estimer la stabilité d'un clustering. Je corrige ce score en le basant sur une hypothèse de distribution multinomiale qui lui permet de prendre en compte la dépendance entre les clusters et de faire des inférences statistiques. Ce ARI modifié ($MARI$) est implémenté dans le package R `aricode`. Ces deux méthodes sont ensuite appliquées à une large cohorte de tumeurs TNBC et les résultats sont discutés en relation avec des résultats des classification du TNBC de la littérature.

Title: Stability and selection of the number of groups in unsupervised clustering: application to the classification of triple negative breast cancers

Keywords: Triple Negative Breast Cancer, Omics, Unsupervised classification, Cluster stability, Rand Index

Abstract: In this thesis, I treat the topic of classifying Triple Negative Breast Cancer (TNBC) tumors from a statistical point of view. After proposing a classification of TNBC based on proteins, I mainly focus on the use of cluster stability for selecting the number of groups in unsupervised clustering. Indeed, this is the method generally employed when classifying TNBC. The aim of this method is to obtain a classification that is robust, that is, easily replicable on similar data. This is measured by its sensibility to small changes, such as subsamplig of the dataset. Despite the popularity of this method, little is still known about how or when it works. For this reason, I propose two important methodological contributions, increasing the usability and interpretability of this method: (1) an R-package, `clustRstab`, that easily enables to estimate the stability of a clustering in different parameter settings. This package is accompanied by a simulation and an application study investigating when and how this method works. (2) A Modified version of the Adjusted Rand Index (ARI), a popular score for cluster comparisons which is a crucial step for estimating the stability of a clustering. I correct this score by basing it on a multinomial distribution hypothesis which enables it to take into account dependence between clusterings and conduct statistical inference. This Modified ARI ($MARI$) is implemented in the R package `aricode`. These two methods are then applied to a large cohort of TNBC tumors and the results are discussed in relation to earlier classification results of TNBC.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France