Cluster stability for more robust classification in Triple-Negative Breast Cancer

Martina Sundqvist[1,4], Leanne de Koning[3], Guillem Rigaill[2,5], Thierry Dubois [4] & Julien Chiquet[1]

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France
[2] Institute of Plant Sciences Paris-Saclay, UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité
[3] Institut Curie - PSL Research University, Translational Research Department, Plateforme RPPA, 26 rue d'Ulm, 75005 Paris, France
[4] Institut Curie - PSL Research University, Translational Research Department, Breast Cancer Biology Group, 26 rue d'Ulm, 75005 Paris, France
[5] Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), Université d'Evry Val dEssonne, UMR CNRS 8071, ENSIIE, USC INRA

Introduction

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 tumors by using unsupervised classification methods.**

Selecting the number of groups in unsupervised classification is an open question in statistics and the number of TN tumor groups is also unknown. For this reason, **we investigated an approach of cluster stability** [3] in order to identify TN tumor groups that were robust, i.e., insensitive to variable variation such as subsampling.

Also, genomic and transcriptomic based classifications have, so far, not allowed to improve patient outcome. Proteins are the effector molecules in the cell and the targets of therapies. Therefore, **we used proteomic and not transcriptomic data for the classification.**

Study Goals

The aim of this work was to **classify** TN breast cancer tumors based on proteomic data, **characterize** the obtained groups and **validate** the classification in a supplementary dataset.

Data

Two datasets of TN breast cancer that were both part of the Rational Therapy for Breast Cancer RATHER (described in [1]) consortium are at our disposal.

- **Training set:**
  Samples collected at
  Netherlands Cancer Institute (NKI) $(n = 67)$
  Protein data performed at Institut Curie $(p = 116)$
- **Validation set:**
  Samples collected at
  Addenbrookes Hospital University, Cambridge $(n = 31)$
  Protein data performed at Institut Curie $(p = 116)$

Cluster Comparison

Clustering stability is based on measures of cluster comparison, *i.e.*, measuring to which extent two clusterings $U = \{U_1, U_2, \ldots, U_R\}$ and $V = \{V_1, V_2, \ldots, V_C\}$ are similar.

We used the **normalized information distance (NID)** which measures to which extent knowing the clustering V would help to predict the clustering U and vice versa [2]. The NID is normalized within the range $[0, 1]$, equaling 0 when the two clusterings are identical, and 1 when they are independent, *i.e.*, sharing no information with each other. Mathematically:

$$NID(U,V) = 1 - \frac{H(U) - H(U|V)}{max\{H(U), H(V)\}},$$

where $H(V)$, $H(U)$ and $H(U|V)$ are the entropy and conditional entropy for U and V.

Cluster Stability

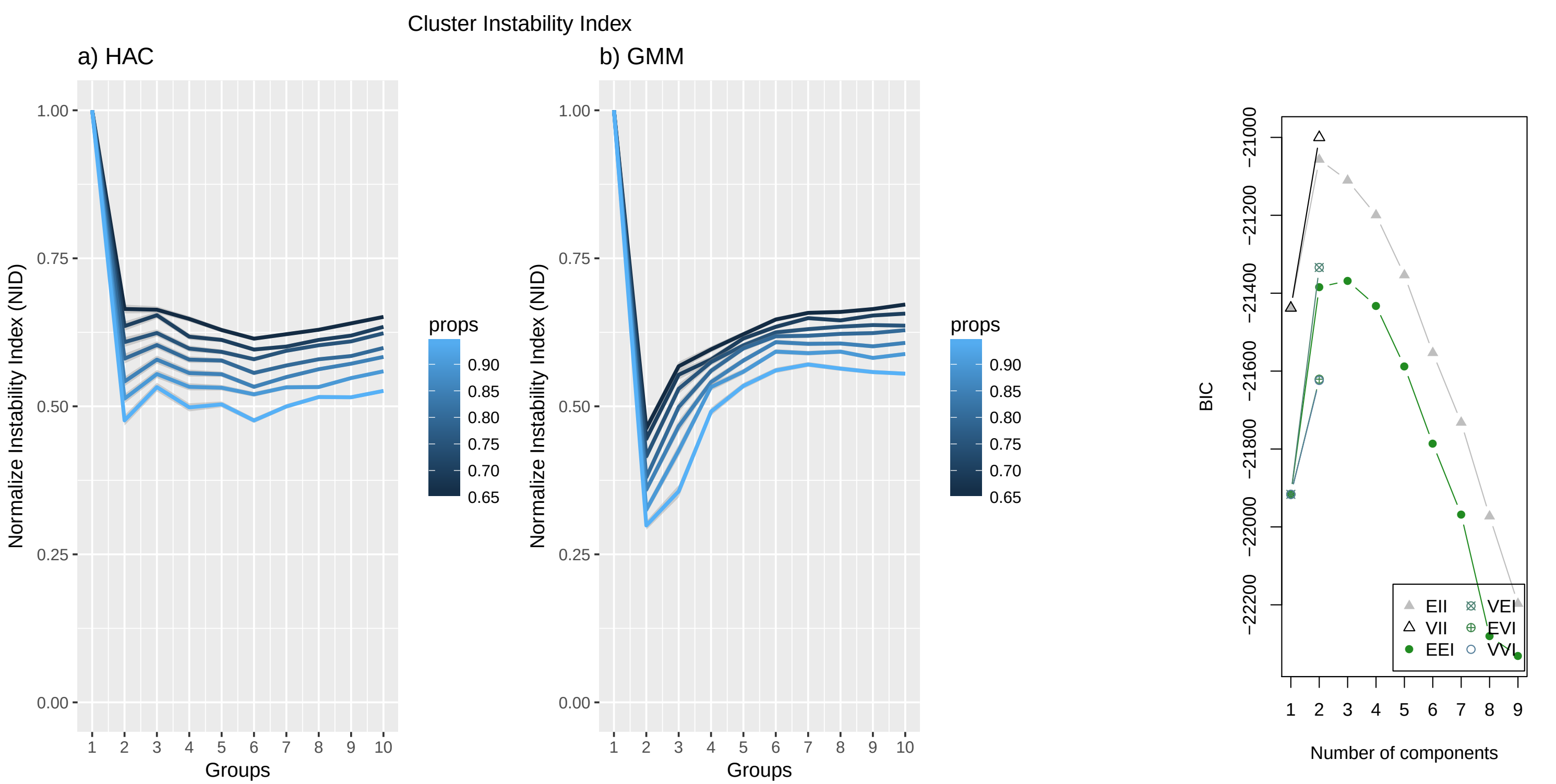
Algorithm 1 Clustering Stability

```
Generate perturbed versions $S_b$ $(b = 1, \ldots, b_{max})$ of the original data set by subsampling a certain proportion of $p$ variables or $n$ observations
Cluster the data set $S_b$ with algorithm $\mathcal{A}$ into $K$ clusters to obtain clustering $C_b$
Compute pairwise comparisons $NID_d(C_b, C_{b'})$ between all possible pairs of clusterings
Permute the labels in $C_b$ to obtain $C_{b,perm}$, compute pairwise comparisons between these clusterings to obtain $NID_{d,perm}(C_{b,perm}, C_{b'})$
Compute normalized instability $\widehat{Instab}(K,p,n)$ as the mean of $NID_d(C_b, C_{b'})$ normalized by $NID_{d,perm}(C_{b,perm}, C_{b'})$
```

Choose the parameter K that gives the best stability:

$$\widehat{K} = \underset{K=1,\ldots,n}{Argmin} \widehat{Instab}(K,p,n)$$

A stable classification with $\widehat{K} = 2$ groups was identified



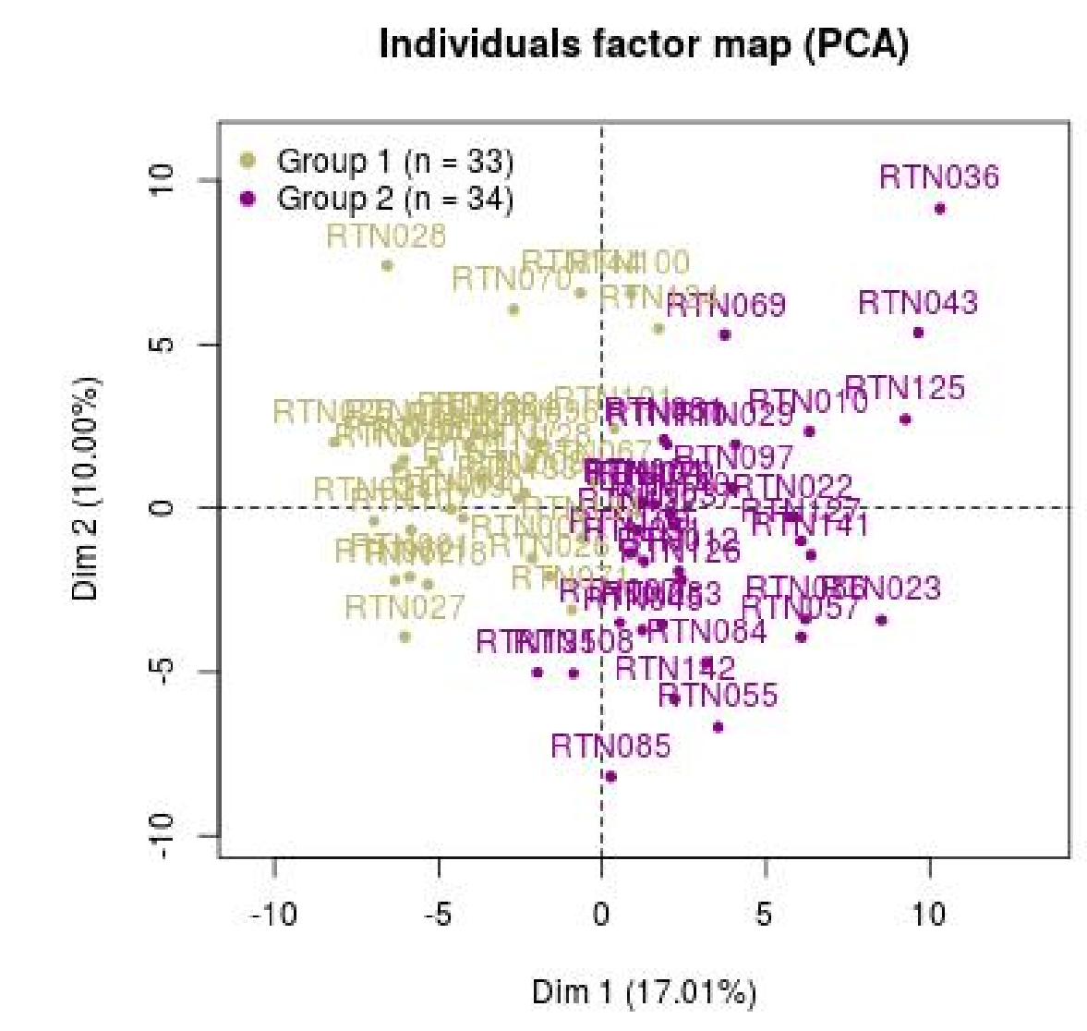
(a) Cluster Stability

(b) BIC values for GMM

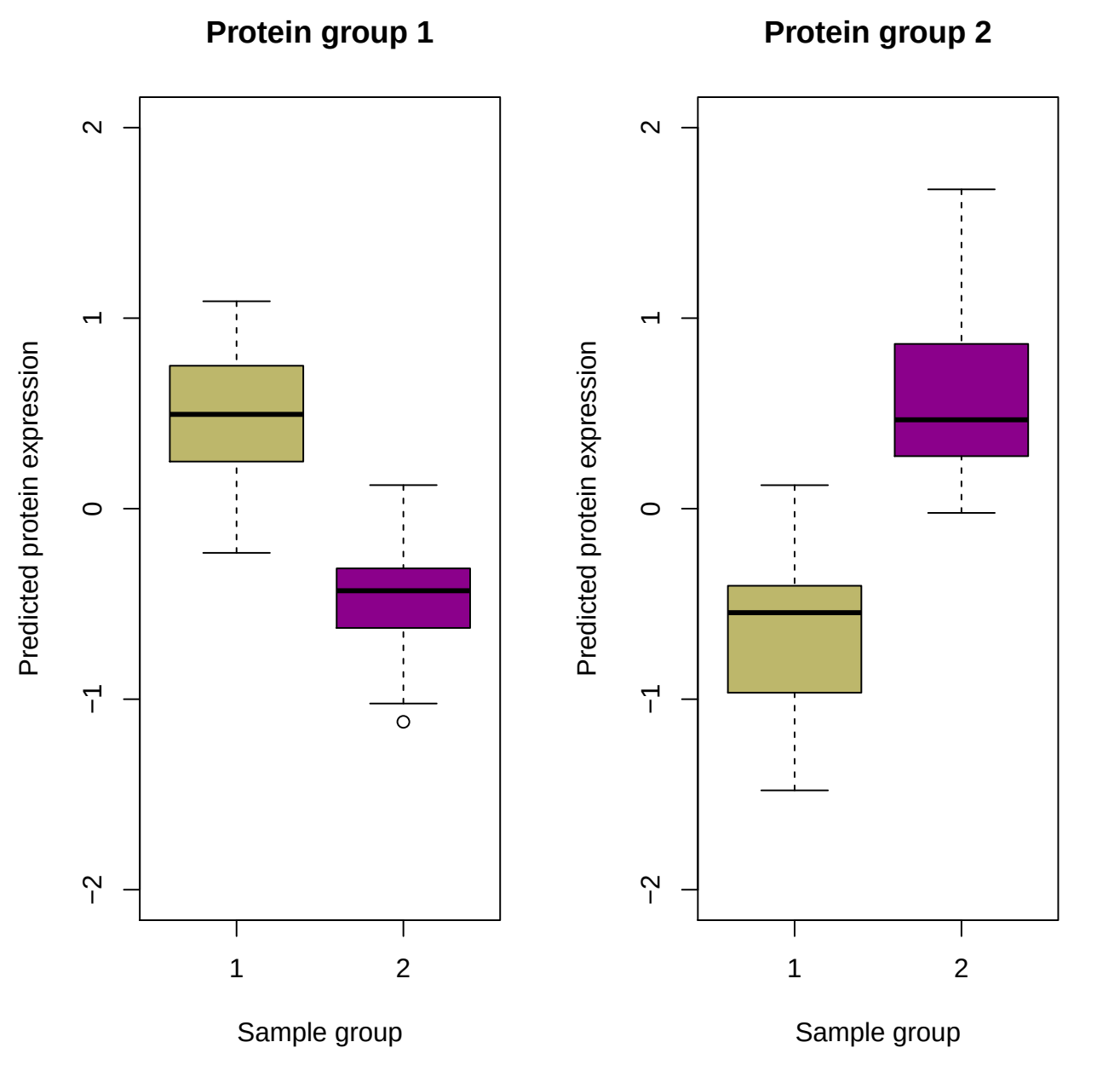
Conclusion: According to the **BIC** and to **clustering stability** $\hat{K} = 2$. The spherical, equal volume (EII) GMM is retained.

The two sample groups showed contrary protein expression patterns

Among the 2 groups 38 proteins are differentially expressed $(p < 0.05)$. These proteins were later classified using GMM methods.



The 2 TN groups projected in two first axis of PCA space.

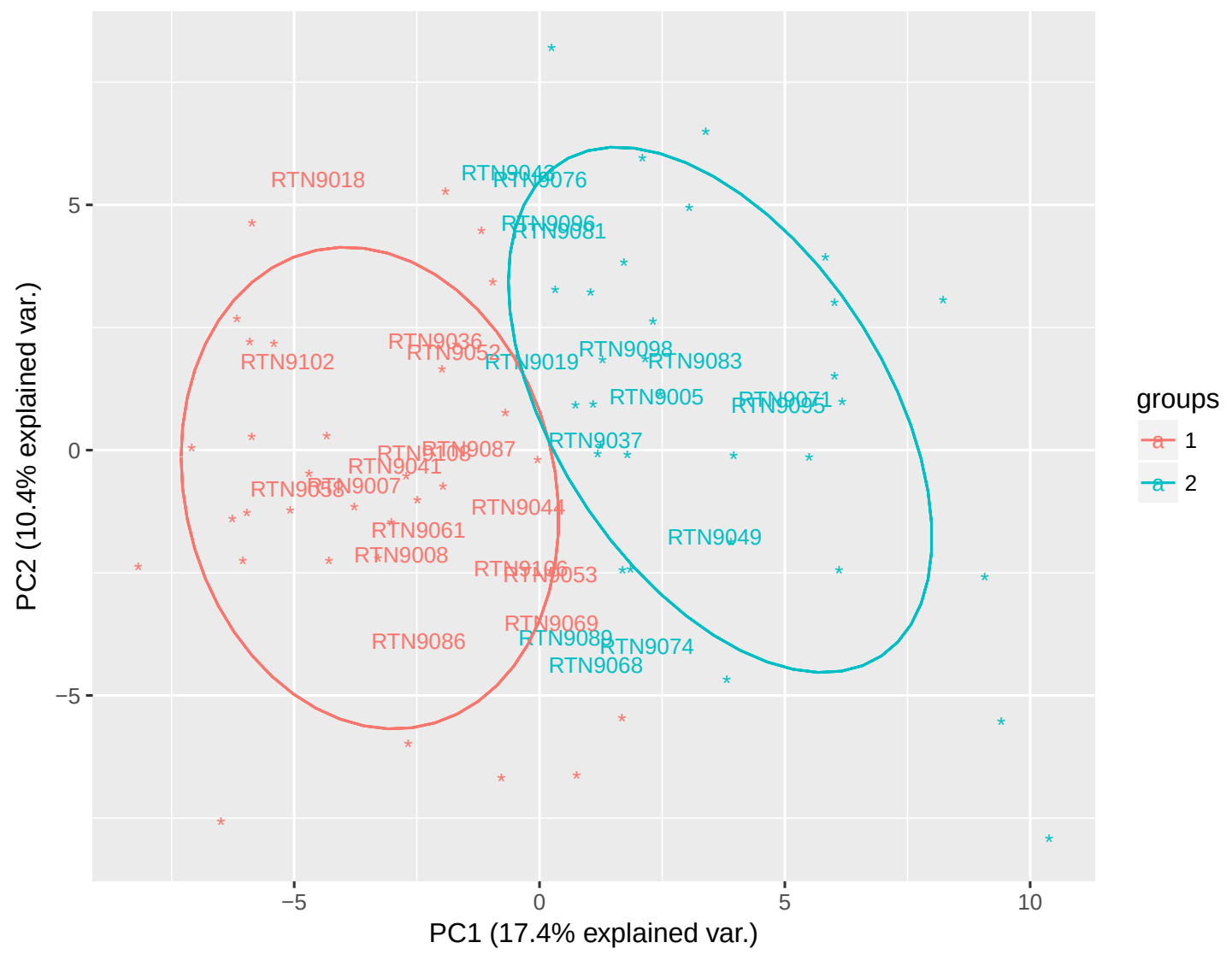


Boxplots for each protein group and each TN sample group.

Conclusion: Each TN group is characterized by a higher expression of a certain group of proteins compared to the other group.

The classification was validated on a second dataset

- The posterior probability of the GMM training model was computed:



Conclusion: The two same groups were found in the validation set.

Discussion et conclusion

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, **confirming the biological relevance of the two groups**. Also, 18 proteins were found to be crucial for identifying the two groups, encouraging for further investigations.

Thus, **the use of clusters stability seems to be a promising approach in selecting number of groups in unsupervised classification.**

References

[1] M. Michaut, S.-F. Chin, I. Majewski, T. M. Severson, T. Bismeijer, L. de Koning, J. K. Peeters, P. C. Schouten, O. M. Rueda, A. 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.

[2] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[3] U. Von Luxburg et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274, 2010.