A sparse logistic mixture model for disease subtyping with clinical and genetic data

Marie Courbariaux, Marie Szafranski, Cyril Dalmasso, Fabrice Danjou, Samir Bekadar, Jean-Christophe Corvol, Maria Martinez, Christophe Ambroise

A sparse logistic mixture model for disease subtyping with clinical and genetic data

Marie Courbariaux[2,1], Marie Szafranski[3,1,⋆], Cyril Dalmasso[1], Fabrice Danjou[4,6], Samir Bekadar[4], Jean-Christophe Corvol[4], Maria Martinez[5], and Christophe Ambroise[1]

[1]*Université Paris-Saclay, CNRS, Univ Évry, Laboratoire de Mathématiques et Modélisation d'Évry, 91037 Évry-Courcouronnes, France*

[2]*Sorbonne Université Maison des Modélisations Ingénieries et Technologies (SUMMIT), Sorbonne Université, 75005 Paris, France*

[3]*ENSIIE, 91025 Évry-Courcouronnes, France*

[4]*Institut du Cerveau et de la Moelle épinière, Hôpital Pitié Salpêtrière, 75013 Paris, France*

[5]*Institut de Recherche en Santé Digestive, INSERM, CHU Purpan, 31024 Toulouse Cedex 3, France*

[6]*Agence Technique de l'Information sur l'Hospitalisation, 75012 Paris, France*

[⋆]*Corresponding author: marie.szafranski@math.cnrs.fr*

Abstract

Motivation: Identifying new genetic associations in non-Mendelian complex diseases is an increasingly difficult challenge. Yet, these diseases seem to have a significant part of heritability to explain. This missing heritability could be explained by the existence of subtypes involving different genetic factors.

Taking genetic information into account in clinical trials can therefore be of interest to guide the process of subtyping a complex disease. Most methods dealing with multiple sources of information rely on data transformation, with two main tendencies regarding disease subtyping in that situation: i) the clustering of clinical data followed with posterior genetic analyzes and ii) the clustering of clinical and genetic variables. Both face limitations that we propose to leverage.

Contribution: This work proposes an original method for disease subtyping from both longitudinal clinical variables and high-dimensionnal genetic markers via a sparse mixture of regressions model. The added value of our approach lies in its interpretability regarding two aspects. First, our model links both clinical and genetic data with regard to their respective initial nature (*i. e.* without transformation) and does not need post-processing to come back to the original information to interpret the subtypes. Also, it can adress large-scale problems thanks to a variable selection step to discard genetic variables that may not be relevant for subtyping.

Results: The proposed method is validated on simulations. A dataset from a cohort of Parkinson's disease patients was also analyzed. Several subtypes of the disease as well as genetic variants having potentially a role in this typology have been identified.

Software availability: The `R` code for the proposed method, named `DiSuGen`, and a tutorial are made available at https://github.com/MCour/DiSuGen.

Status: as of march 2021, this preprint has just been submitted to Pattern Recognition Letters.

1 Introduction

Known genetic markers in complex disease usually account for only a part of calculated heritability. A possible interpretation could be that there exists subtypes of those complex diseases involving different genetic factors. In order to identify such subtypes, large heterogeneous datasets are now available, including for example patients follow-up and genotyping data.

Two approaches arise to address the problem of subtyping when clinical and genomic information are available: i) the clustering of clinical data with a posterior genetic analysis and ii) the concomitant clustering of clinical and genomic data. We will discuss the pro and cons of both approaches in Section 2.

Contributions In this work, we sketch an alternative path at the crossroad of the possibilities mentioned above. It consists in clustering the clinical variables by estimating a multinomial logistic regressions model whose weights depend on the genetic variables. The model is shaped for the longitudinal nature of the clinical data and accounts for the high dimensionality of the problem with a sparse constraint on the parameters involved in the logistic weights.

Organization of the paper Section 2 gives an overview of different strategies that may be used for disease subtyping with different sources of information. Section 3 proposes a framework, related to mixture of experts models, for clustering of clinical longitudinal data guided by genetic markers. Section 4 describes the algorithm and its implementation for the high-dimensionality setting. Finally, Section 5 provides an illustration of our approach using numerical simulations, and Section 6 gives an analysis of a cohort of patients with Parkinson's Disease.

2 Disease subtyping with multiple information

In this section, we provide a general picture of approaches that may be used for clustering using different sources of data, with a particular attention on methods dedicated to disease subtyping with multiple information.

2.1 Clustering of clinical data with posterior genetic analyzes

A first attempt would consist in a two-step approach with i) a disease subtyping based on clinical data and then, ii) an analyse of the genetic associations in each subtype.

Clustering of clinical data The data often come from a clinical follow-up in which case they are generally of longitudinal nature. A review of clustering methods adapted to functional data, including longitudinal data, is presented by Jacques and Preda (2014), with the following categorization:

- *Methods with a filtering step* consist in summarizing the curves by a few descriptors such as their slope and intercept, followed by a clustering step on those descriptors.
- *Non-parametric methods*, such as K-means, with distance metrics adapted to longitudinal data.

- Finally, *model-based methods* appear to be the most adapted to deal with short longitudinal data including numerous missing values, as often encountered when dealing with medical follow-ups. An overview of approaches and tools dedicated to *mixture models* for longitudinal data has been proposed by van der Nest *et al.* (2020).

  Remark. In this work, we will focus on *mixtures of experts*, a specific category of *mixture models*, with a dedicated description given in Section 3.1.

Analyse of clinical clusters with genomics In a second step, one might exhibit genetic associations that explain the clusters using them as phenotypes in standart approaches devoted to GWAS which usually involve statistical procedures based on (multiple) hypothesis testing (see for instance (Bush and Moore, 2012) or (Hayes, 2013)). Another way to reveal such associations could be to resort to classical supervised methods, such as (multinomial) logistic regression, with a feature selection procedure (Ma and Huang, 2008, and references therein).

Limitation A well-suited clustering of clinical data followed with posterior genetic analyzes of the clusters obtained does not take benefit from the genomic data in the clustering step. As a consequence, there is no guarantee regarding the connection between the genomic information and the clinical clusters. Also, most sparse model-based clustering methods for functional or longitudinal data in high dimension rely on dimensionality reduction techniques, such as PCA or SVD, which are efficient but much less convenient for interpretation.

2.2 Concomitant clustering of clinical and genomic data

Concomitant clustering gathering both clinical and genomic data represents an attractive alternative although many variables may be involved. In this context, feature or variable selection strategies are mandatory to solve the problem.

Multi-view clustering This framework coming from the machine learning community is popular for solving problems with different feature sets. The survey of Fu *et al.* (2020) divides them into three categories.

- *Graph-based methods* combine different views according to their respective importance and then mainly resort to spectral clustering algorithms.
- *Space-learning-based methods* are designed to construct a new learning space using the most representative characteristic of each view to enhance clustering.
- *Binary-code-learning-based methods* encode original data as binary features using mapping and reduction techniques in order to save computation time and memory.

We must also mention the *Multiple Kernel Learning* framework declinated for clustering (Zhao *et al.*, 2009) as another kind of multi-view learning. In particular, the work of Mariette and Villa-Vialaneix (2017) proposed (consensus) meta-kernels to aggregate the different sources of information while preserving the original topology of the data. The works dedicated to disease subtyping with clinical and genomic information falling into the scope of multi-view clustering use *space-learning-based methods* with dimensionality reduction approaches. Sun *et al.* (2014) propose a multi-view co-clustering method based on Sparse Singular Values Decomposition (Lee *et al.*, 2010). Sun *et al.* (2015) enhance this work providing convergence guarantees using the proximal alternating linearized minimization algorithm of Bolte *et al.* (2014).

Integrative clustering In cancer research, many statistical methodologies have emerged to analyse data coming from different sources, generally multiple omics data, under the concept of *integrative genomics* (Kristensen *et al.*, 2014), with a philosophy closely related to multi-view learning. Huang *et al.* (2017) present a review of multi-omics integration tools. In addition, we should also refer to *mixOmics* (Rohart *et al.*, 2017) which proposes various sparse multivariate methods to explore multiple omics datasets. More specifically, integrative clustering may be built on model-based approaches such as the representative work of Shen *et al.* (2009, 2010). The method *iCluster* uses a latent variable model to connect multiple data types. The optimization of a penalized log-likelihood alternates a process of dimensionality reduction on the representation of original data with a sparse estimation of the corresponding coefficients. Several extensions of *iCluster* using penalties inducing sparsity of different forms have been proposed since (Shen *et al.*, 2013; Kim *et al.*, 2017). Finally, to discover the subtypes across the different views, *PINSPlus* (Nguyen *et al.*, 2018) uses a perturbation scheme applied on each source of data to define stable clusters, before merging results using several algorithms to design a similarity matrix based on the overall connectivity of the patients.

Limitation Concomitant approaches could be adapted for problems dealing with clinical and genomic datasets. However, none of them explicitly adress how to deal with data of different nature nor to take the longitudinal aspect properly into account. Most methods require representations derived from the original space. However, distorting the initial information may significantly complicate the posterior validation of the extracted features. The limitation of methods based on dimensionality reduction has been mentioned above. For methods based on similarity matrices, such as kernel methods which implicitly map the data in a new feature space, an additional difficulty emerges since it requires to solve a pre-image problem to approximate the features and try to interpret them.

3 Mixture of regressions with clinical and genomic data

To take benefit of the clinical and the genomic information, both datasets can be used simultaneously into a mixture model. Mixtures of experts provide an elegant framework to include concomitant variables as a side information to subtype data (Gormley, 2019). This section starts with a description of mixture of experts models in order to lay the foundation and draw the connections of our approach with this framework.

3.1 Background on mixture of experts models

We assume to be given $\mathbf{Y}$, a matrix of N outcomes data represented by variables $v \in \{1 \cdots V\}$ such that $\mathbf{y}_i = (y_{i1}, \cdots, y_{iv}, \cdots, y_{iV})$, for $i \in \{1 \cdots N\}$. These observations come from a population of K components. We denote by $\mathbf{z} = (z_1, \cdots, z_i, \cdots, z_N)$, the component membership vector where $z_i \in \{1 \cdots K\}$, and by $\mathbf{Z}$ the corresponding indicator matrix such that $\mathbf{z}_i \in \{0, 1\}^K$, with $z_{ik} = 1$ if the observation i belongs to the k^{th} component and $z_{ik'} = 0$ otherwise, $\forall k \neq k'$. A matrix $\mathbf{G}$ of N concomitant data represented by variables $\ell \in \{1 \cdots L\}$ is also available, with $\mathbf{g}_i = (g_{i1}, \cdots, g_{i\ell}, \cdots, g_{iL})$, for $i \in \{1 \cdots N\}$. The random vectors associated to these representations are respectively denoted Y, Z and G.

Remark. To lighten notations, the range of indexes will often be omitted, in which case indexes i, v, ℓ and k (or k') will go to the ranges define above.

To refer to the terminology used in Gormley (2019, Section 2.3), we are interested in *simple mixtures of experts models* where the outcome data distribution depends on the latent component membership, which itself depends on the concomitant variables, such as $\mathbb{P}(\mathbf{y}_i, z_i \mid \mathbf{g}_i) = f_{z_i}(\mathbf{y}_i \,;\, \Theta_{z_i}(\mathbf{g}_i))\, \eta_{z_i}(\mathbf{g}_i)$, with

$$\mathbf{y}_i \,|\, \mathbf{g}_i, z_i = k \sim f_k(\mathbf{y}_i \,;\, \Theta_k(\mathbf{g}_i)) \,, \tag{1a}$$

$$\text{and} \quad \mathbb{P}(z_i = k \,|\, \mathbf{g}_i) = \eta_k(\mathbf{g}_i) \,, \tag{1b}$$

where $\Theta_k(\cdot)$ is the set of parameters of the k^{th} component density function $f_k(\cdot; \Theta_k(\cdot))$, *i.e.* the k^{th} expert, and $\eta_k(\cdot)$ the probability weight related to the k^{th} expert.

3.2 Proposed approach

Based on the above described framework, we propose a mixture of regressions model for disease subtyping when patient symptoms are recorded from their follow-up along with genetic markers as concomitant variables.

Specificities Our model is designed to take into account the longitudinal aspect of the clinical data as well as the high dimensional setting of the genetic data. The cohort $\mathbf{Y}$ is observed on clinical variables during several visits indexed by j. The v^{th} clinical variable observed during the j^{th} visit of patient i is denoted $y_{iv_{(j)}}$. Also, the number of variables of genetic data $\mathbf{G}$ may be of the order of a few millions after genotype imputation, so that dedicated metrics (as CADD (Rentzsch *et al.*, 2018), used in our application on Parkison's Disease) or more general elimination techniques such as screening rules (see Ndiaye *et al.* (2017) for instance) may still be required beforehand. Note that even with such a processing, we remain in a configuration where $N \ll L$.

Model In order to connect our proposal with the mixture of experts stated in (1), we characterize the problem as

$$y_{iv_{(j)}} \,|\, \mathbf{g}_i, z_i = k \sim f_k(y_{iv_{(j)}} \,;\, \{\boldsymbol{\alpha}_{vk}, \sigma_{vk}\}) \,, \tag{2a}$$

$$\text{and} \quad \mathbb{P}(z_i = k \,|\, \mathbf{g}_i) = \eta_k(\mathbf{g}_i \,;\, \boldsymbol{\omega}_k) \,, \tag{2b}$$

defining the following regression model with logistic weights

$$f_k(y_{iv_{(j)}} \,;\, \{\boldsymbol{\alpha}_{vk}, \sigma_{vk}\}) = \sum_{p=0}^{P} \alpha_{vkp}\, t_{ij}^{p} + \sigma_{vk}\, \varepsilon_{iv_{(j)}} \,, \tag{3a}$$

$$\text{and} \quad \eta_k(\mathbf{g}_i \,;\, \boldsymbol{\omega}_k) = \frac{\exp\left(\omega_{k0} + \boldsymbol{\omega}_k^{\intercal}\mathbf{g}_i\right)}{\sum_{k'} \exp\left(\omega_{k'0} + \boldsymbol{\omega}_{k'}^{\intercal}\mathbf{g}_i\right)} \,, \tag{3b}$$

where

- t_{ij} is the time, such as the patient age or the time since the beginning of the disease, for the patient i at its j^{th} follow-up visit,
- $p \in \{0 \cdots P\}$ is the polynomial degree considered in the regression ($P = 2$ is generally sufficient),

- $\{\alpha_{vkp}\}$, $\{\sigma_{vk}\}$ and $\{\boldsymbol{\omega}_k\}$ are parameters or vectors to be estimated, with $\{\omega_{1\ell}\} = 0$ for the sake of identifiability,

- $\varepsilon_{iv_{(j)}} \underset{iid}{\sim} \mathcal{N}(0,1)$, implies some conditional independence assumptions between variables, patients and visits when the class is known. The clinical variables are chosen such that they are as independent as possible, correlation between individuals should essentially come from a similar typology of the disease and, finally, the remaining time correlation after the polynomial regression is expected to be poor. If the Gaussian hypothesis does not apply to the variable v, one may consider Poisson or logistic regression instead with no substantial additional cost.

The modeling of posterior probabilities via logistic regression allows concomitant variables, such as genetic data, to subtly influence the subtyping.

Model selection We combine two model selection strategies to select the hyperparameters involved in the mixture. The Bayesian Information Criterion (BIC) is widely used to select K, the most appropriate number of subtypes and P, the polynomial degrees in the main regressions. Also, as discussed above, we suspect that many variables ℓ from $\mathbf{G}$ have an insignificant influence to explain the disease phenomenology. Hence, a Lasso penalization will be applied on the coefficients $\{\boldsymbol{\omega}_k\}$, $\forall k$, to select those which are the most involved in the subtyping. More detail about this aspect will be provided in Section 4.

4 EM algorithm with integrated Lasso inference

The inference of such a model with latent variables, here $\{z_{ik}\}$, can be classically conducted with an Expectation Maximization algorithm (EM algorithm, Dempster *et al.*, 1977). We use a modified version of this algorithm with a Lasso-type penalized likelihood instead of the classical likelihood.

4.1 EM algorithm

At the $(q+1)^{\text{th}}$ iteration of the modified EM algorithm, one maximizes the expected and penalized complete-data log-likelihood $\mathcal{L}(\mathbf{Y} \mid \mathbf{G}, \mathbf{Z}; \Theta = \{\boldsymbol{\alpha}, \boldsymbol{\sigma}, \boldsymbol{\omega}\}) - Pen(\boldsymbol{\omega})$ which reads

$$\sum_i \sum_k z_{ik} \left[\log\left[\eta_k(\mathbf{g}_i\,;\,\boldsymbol{\omega}_k)\right] + \sum_v \sum_j \log\left[f_k(y_{iv_{(j)}}\,;\,\{\boldsymbol{\alpha}_{vk}, \sigma_{vk}\}) \right] \right] - \lambda \sum_k \|\boldsymbol{\omega}_k\|_1 \,,$$

where $\lambda > 0$ controls the amount of sparsity applied on the ℓ_1 norm of $\boldsymbol{\omega}_k$ and where $\eta_k(\cdot;\cdot)$ and $f_k(\cdot;\cdot)$ are defined as in (3).

To maximize the expected and penalized complete-data log-likelihood, each iteration is divided into an expectation step (E) followed by a maximization step (M).

- At step E of the $(q+1)^{\text{th}}$ iteration, posterior weights are updated as follows:

$$\begin{aligned}\tau_{ik}^{(q+1)} &= \mathbb{E}\left[z_{ik} | \left(\mathrm{Y} = \mathbf{y}_i | \mathbf{g}_i\right); \Theta^{(q)}\right] \\ &= \frac{\eta_k\left(\mathbf{g}_i\,;\,\boldsymbol{\omega}_k^{(q)}\right)\prod_v\prod_j f_k\left(y_{iv_{(j)}}\,;\,\{\boldsymbol{\alpha}_{vk}^{(q)}, \sigma_{vk}^{(q)}\}\right)}{\sum_{k'}\eta_{k'}\left(\mathbf{g}_i\,;\,\boldsymbol{\omega}_{k'}^{(q)}\right)\prod_v\prod_j f_{k'}\left(y_{iv_{(j)}}\,;\,\{\boldsymbol{\alpha}_{vk'}^{(q)}, \sigma_{vk'}^{(q)}\}\right)}\,.\end{aligned}$$

- At step M of the $(q+1)^{\text{th}}$ iteration, parameters are updated as follows:

$$\begin{aligned}\Theta^{(q+1)} = \underset{\Theta}{\operatorname{argmax}} \sum_i\sum_k \tau_{ik}^{(q+1)} \Bigg[&\log\left[\eta_k(\mathbf{g}_i\,;\,\boldsymbol{\omega}_k)\right] \\ &+ \sum_v\sum_j \log\left[f_k(y_{iv_{(j)}}\,;\,\{\boldsymbol{\alpha}_{vk}, \sigma_{vk}\})\right]\Bigg] \\ &- \lambda\sum_k \|\boldsymbol{\omega}_k\|_1\,.\end{aligned}$$

The maximization with regard to parameters $\{\boldsymbol{\alpha}, \boldsymbol{\sigma}\}$ presents no difficulty. However, there is no close formula to update the logistic weights parameters. The term to be maximized with respect to $\{\boldsymbol{\omega}\}$ at iteration $(q+1)$ of the EM algorithm is

$$\frac{1}{N}\sum_i\sum_k \tau_{ik}^{(q+1)} \log\left[\eta_k(\mathbf{g}_i\,;\,\boldsymbol{\omega}_k)\right] - \lambda\sum_k \|\boldsymbol{\omega}_k\|_1\,. \tag{4}$$

This maximization problem corresponds to the multinomial logistic regression problem with a ℓ_1 penalty. It can be addressed by a classical partial Newton algorithm [1].

4.2 Variable selection in practice

Common strategies to select the hyperparameter λ are based on adjusted information criterion (see Chen and Chen (2012) for General Linear Models or Fop *et al.* (2018) for a more global overview). In an original approach, Yi and Caramanis (2015) proposed to optimize that hyper-parameter with an iterative scheme through successive M steps and showed local convergence properties in the high dimensional setting.

In this work, we rely on an alternative adopted by Mortier *et al.* (2015) where λ is chosen within the M step by cross-validation so that the likelihood of the multinomial logistic model (4) is maximized. A short simulation study shows that proceeding with that selection at every M step of the EM algorithm does not compromise the convergence of the algorithm.

Finally, to avoid (negative) bias due to the penalization in the parameter estimation, we re-estimate the selected $\{\boldsymbol{\omega}\}$ parameters at the end of the EM algorithm in order to get the maximum likelihood estimates, as usually done (Hastie *et al.*, 2009, p. 91).

[1] For instance, with the `R` package `glmnet` (Friedman *et al.*, 2010).

4.3 Implementation

The implementation of the method proposed in this paper, named `DiSuGen`, as well as an `R` Markdown tutorial are available (Courbariaux *et al.*, 2020). We build on the `FlexMix R` package of Grun and Leisch (2008) which proposes a EM algorithm adapted to multinomial logistic weights mixture models. To implement our method, we developed an adapted concomitant variable driver making use of `glmnet` within `FlexMix`. Finally, for a faster convergence, we resort in practice to a Classification EM (CEM) algorithm in which $\tau_{ik}^{(q+1)}$ are replaced by the indicator variables $z_{ik}^{(q+1)}$ (Celeux and Govaert, 1992).

5 Numerical illustrations on artificial data

The proposed estimation and model selection procedures are first evaluated on artificial data. These simulations are designed to study the ability of the CEM algorithm to produce a good estimation of the parameters on the one hand and to recover the appropriate model on the other hand.

5.1 Data generation

Artificial data are simulated according to the model (3) with $N = 396$ patients, $V = 4$ clinical variables, $K = 3$ clusters, $P = 1$ polynomial degree in the regression, 3 follow-up visits per patient with times t_{ij} randomly ranging from 10 to 410 days for the first visit, from 1800 to 2200 days for the second and from 3600 to 4000 days for the third. Also, $L = 2657$ genetic markers are simulated with only 10 having an influence on the clustering such that $\boldsymbol{\omega}_{k\{\ell\}}$ is $\boldsymbol{\omega}_{2\{2,3,4\}} = \boldsymbol{\omega}_{3\{5,6,7\}} = 2$, $\boldsymbol{\omega}_{2\{5,6,7\}} = -1$ and $\boldsymbol{\omega}_{3\{1,8,9,10\}} = -2$. For the sake of consistency with the study presented in Section 6, the genetic markers come from the Parkinson's disease genetic data and the parameters $\{\boldsymbol{\alpha}, \boldsymbol{\sigma}\}$ are chosen so as to be realistic with regard to the Parkinson's disease clinical data.

5.2 Protocol

For each simulation, the proposed CEM algorithm is launched with a number of $K = 3$ clusters and a Lasso penalty. The estimation is initialized with 10 sets of starting values corresponding to 10 random assignments into $K = 3$ clusters and the one that leads to the lower BIC is kept. This experiment is repeated 100 times. To assess the performance of our method, we propose different settings of comparison:

- The *integrative* method is the one described in this paper, which uses both clinical and genetic data and estimates the parameters $\{\boldsymbol{\alpha}, \boldsymbol{\sigma}\}$ and $\{\boldsymbol{\omega}\}$, and the subtypes $\mathbf{z}$.
- The *oracle integrative* and *semi-oracle integrative* methods also use both type of data to estimate the subtypes $\mathbf{z}$. The *oracle integrative* uses all parameters of the model fixed to their true values. The *semi-oracle integrative* method only fixed the parameters $\{\boldsymbol{\omega}\}$ to their true values. This allows us to check to what extent our method correctly subtypes the data and estimates the parameters related to clinical variables (semi-oracle) and the genetic variables (oracle).
- The *2-step* method does not use genetic information into the clustering process to estimate the parameters $\{\boldsymbol{\alpha}, \boldsymbol{\sigma}\}$, and corresponds to constant weights of the clusters, *i.e.* $\mathbb{P}(z_{ik} =$

$1) = \pi_k$. In this case, the Lasso-penalized multinomial logistic regression is performed afterward to get genetic association results. This allows us to check the advantage of taking into account the genetic information in the clustering process while the clinical parameters are still estimated.

- The *oracle 2-step* method uses the same setting described in the *2-step* method to estimate π_k, except that parameters $\{\boldsymbol{\alpha}, \boldsymbol{\sigma}\}$ are fixed to the true values.
- When possible, the proposed method is also compared to the K-means method which corresponds to a simple Gaussian mixture model with identical proportions and identical standard deviations in all clusters. To do so, recourse is made to a K-means method adapted to longitudinal data implemented in the `R` package `kml3d` (Genolini *et al.*, 2015).

5.3 Results

Clustering ability The Adjusted Rand Index (ARI, Rand (1971); Hubert and Arabie (1985)) is computed for each simulation to check that the estimated clusters are close to the ones that are simulated. The results for the proposed method are illustrated by the boxplot for the *integrative* method of Figure 1(a). Most clusters are well identified. When making no use of genetic information within the clustering, the algorithm globally achieves lower ARI. A better clustering ability can be expected from the algorithm making use of genetic data if more information is provided as illustrated by the *oracle integrative* results. This improvement in cluster prediction does not come from a better estimation of the parameters $\{\boldsymbol{\omega}\}$ as illustrated by the *semi-oracle integrative* results. Finally, the K-means algorithm is not able to recover the underlying classification as well: the corresponding ARI all range between 0.6 and 0.7. This was expected, since the differences between the clusters partly lie in the variances of the variables. Moreover, the K-means method is not able to account for the exact visit times but only for the visit ranks.

Parameters estimation ability The parameters of the the main regressions are estimated accurately and with biases notably close to 0 whatever the considered clinical variable and the chosen approach (*2-step* or *integrative*). Accounting for genetic information does not seem to improve notably the estimation of those parameters. When it comes to the logistic regression parameters, the sign of the estimated parameters are most of the time adequate for both approaches as illustrated Figure 1(b).

Variable selection within the logistic regression Figure 1(d) illustrates the results of the proposed Lasso selection procedure with regard to the genetic variables. The selection rates of 8 of the 10 active genetic variables are notably higher than the selection rates of the other variables, thus showing a good performance of the selection method. The selection rates obtained using the *2-step* method are lower, which underlines the interest of the proposed *integrative* approach. The 2 remaining active markers do not vary between patients and can thus be replaced by any variant with a poor variation. Most of the inactive markers are selected less than 5 times over 100 simulations. However this specificity is good, many false positives are observed for each simulation since the number of genetic markers is relatively high (2657). This selection performance decreases (as expected) as the parameters $\{\boldsymbol{\omega}\}$ are set closer to zero (not shown).

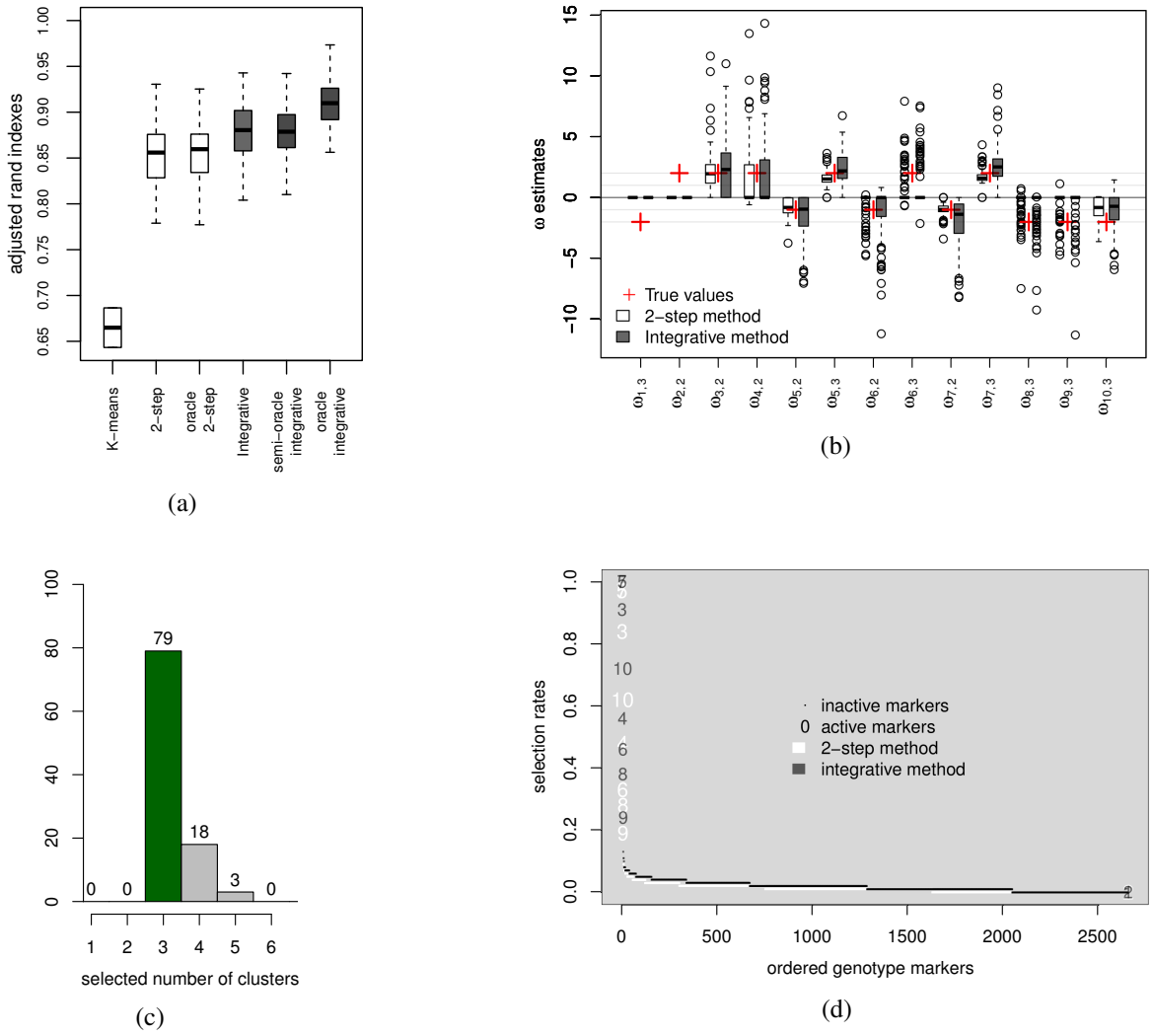

Figure 1: Results on artificial data over 100 simulations. (a): ARI with a K-means algorithm, the *2-step* method (no use of genetic information) and the integrative method (use of genetic information) and the corresponding oracles. (b): Non-negative $\{\boldsymbol{\omega}\}$ parameter estimates and their respective true values. (c): estimated number of clusters according to the BIC. (d): selection rates of the genetic variables; the 10 variables that should be selected are represented by their respective number.

Model selection ability An additional simulation study is conducted to evaluate the capacity of the BIC (computed as described Section 3) to select the correct number of clusters ($K = 3$) on the same 100 simulated datasets. The results are illustrated by the histogram of Figure 1(c). The correct number of clusters is selected 79 times over 100.

6 Application to Parkinson's Disease (PD) subtyping

The proposed method is applied to PD subtyping. This disease is known to have several subtypes. This has given rise to numerous studies among which the one of Lewis *et al.* (2005).

6.1 Data description

The data on which the method is applied are from the DIG-PD cohort (Corvol *et al.*, 2018), composed of 396 genotyped adults with a recent PD onset (diagnosed less than 6 years prior to the beginning of the study).

Clinical Clinical data were collected at inclusion and then at yearly clinical follow up, during one to seven years. They include scores evaluating the progression of the disease. Two of them are assumed to be representative of the evolution of the disease, namely Section III of the Unified PD Rating Scale (UPDRS III, a motor examination) and the Mini-Mental Status Examination toolkit score (MMSE, an evaluation of cognitive impairment). The higher the UPDRS III and the lower the MMSE, the more patients are impaired. These two scores were adjusted for the treatment doses and for the gender effects beforehand, by considering the residuals of the linear regression with gender and treatment doses as (respectively factor and quantitative) predictors. The patient age is taken as the time scale.

Genetic More than 6 million genetic markers were available after imputation for each patient. Only 2652 of them are used: ones that either were associated to PD in previous studies (about 400 of them) or that had an important impact on the gene function (scaled CADD score [2] greater than 25) and an allele frequency greater than 0.01. As done classically, the ones with two copies of the reference allele were encoded -1, the ones with two copies of the alternative allele were encoded 1 and the others (with one copy of each) were encoded 0.

6.2 Results

Model selection results In order to warranty a good interpretability of the results, a maximal number of $K = 4$ clusters was allowed. Also, a maximal number of 2 polynomial degrees was tested. The solution with 4 clusters and 1 polynomial degree resulted is the lowest BIC.

Clinical results Clustering results with regard to the clinical data are illustrated in Figure 6.2. Note that the variables represented here are residuals of a fitted linear model. The cluster attributed to each patient corresponds to the cluster the patient is the most likely to belong to according to the model. Half of the patients are allocated to cluster 1 and about one third of the remaining patients are allocated to each of the 3 remaining clusters. Cluster 3 is characterized

[2] Combined Annotation Dependent Depletion score, this score evaluates the deleteriousness of variants in the human genome (Rentzsch *et al.*, 2018).

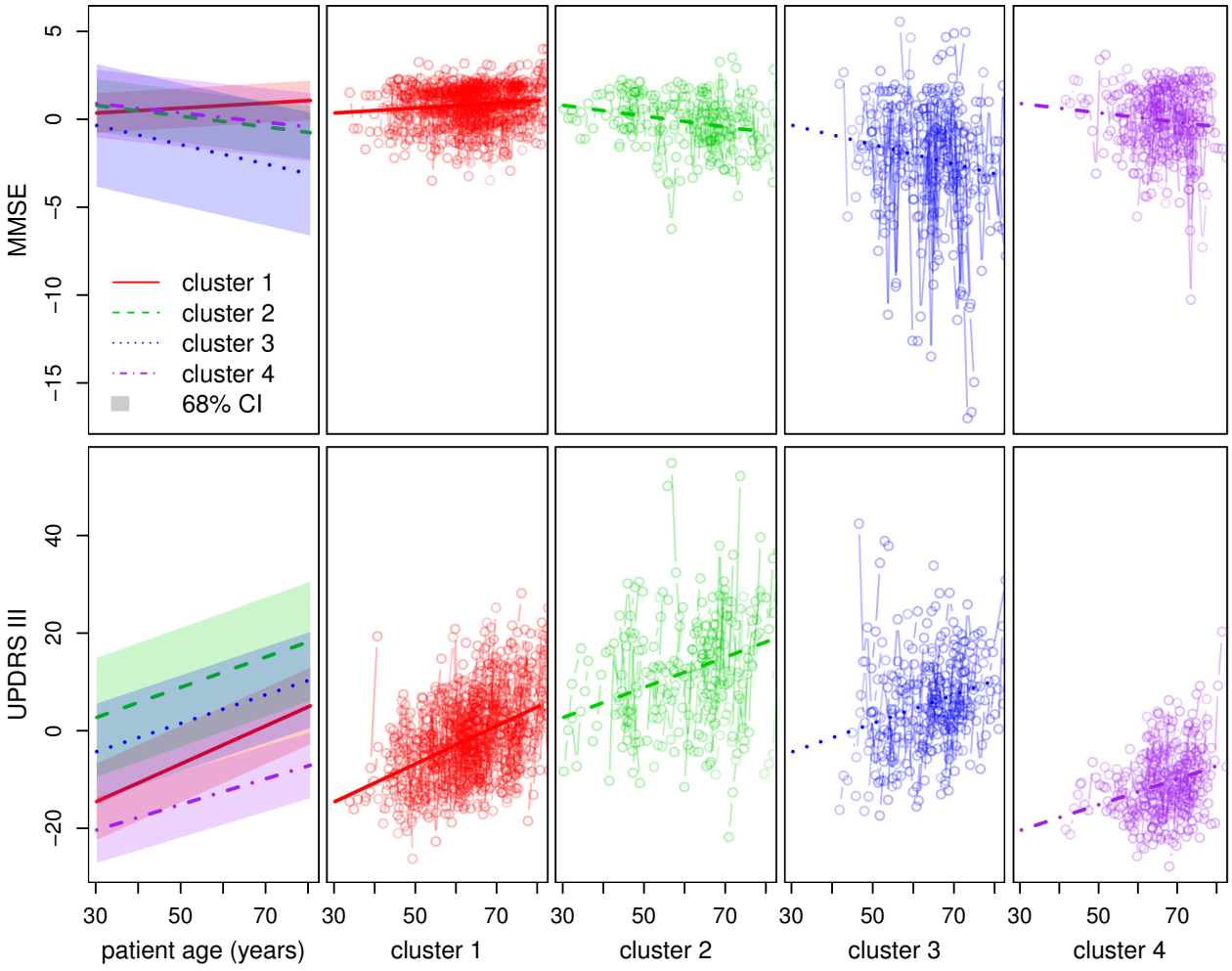

Figure 2: Clustering with regard to the clinical variables.

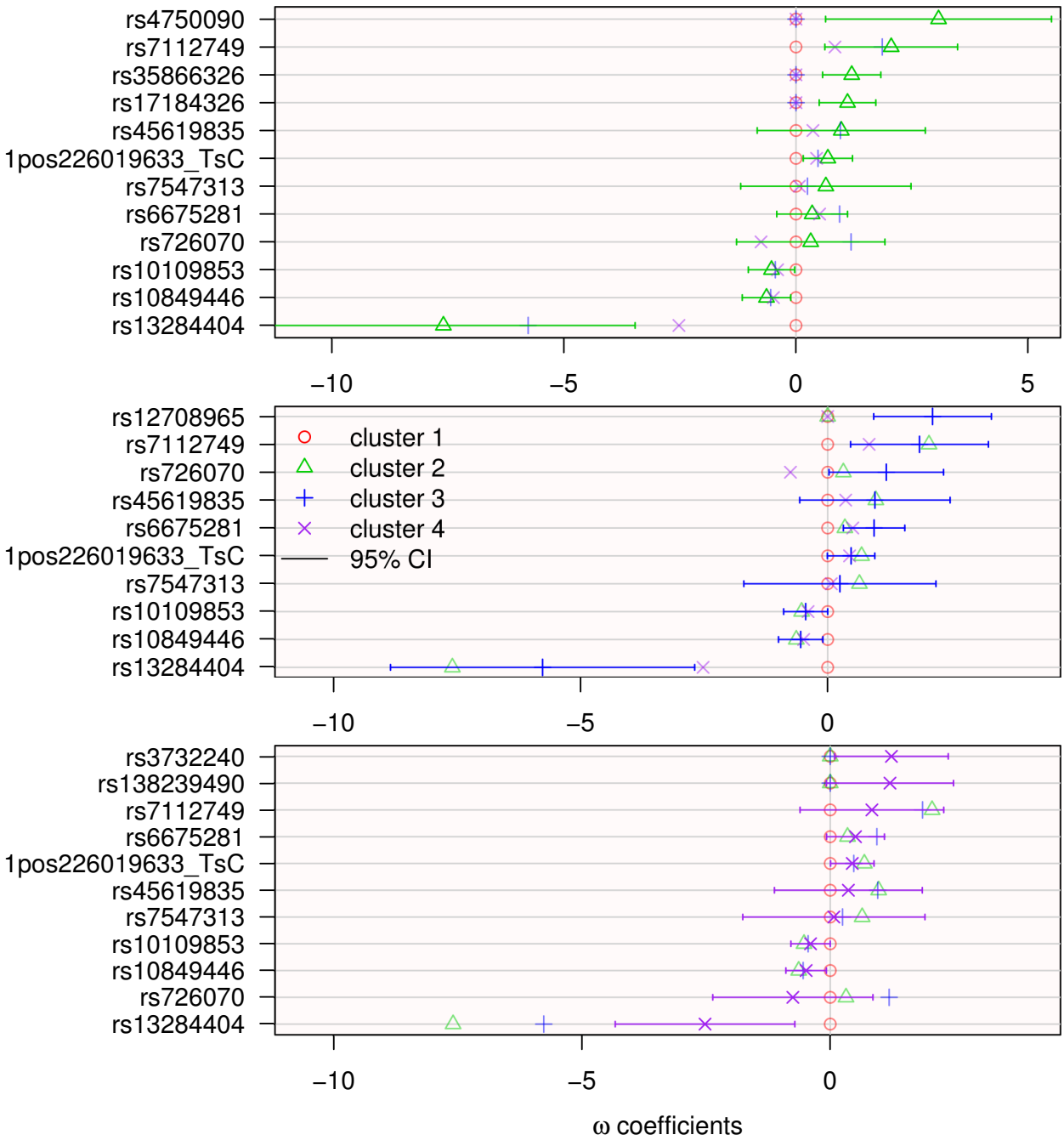

Figure 3: Genetic association: estimated logistic regression parameters $\{\omega\}$.

by late but rapid cognitive and motor decline. Cluster 2 is characterized by a smaller cognitive impairment but by a very significant motor impairment. In contrast, cluster 4 is characterized by a small and late cognitive and motor decline.

Genetic association results Figure 6.2 shows the results with the 95% confidence intervals associated to the parameters $\{\boldsymbol{\omega}\}$. [3] The p-values below 0.05 (i.e. significant association prior to any multiple test correction) correspond to a 0 value outside the 95% confidence interval.

There were 15 SNPs selected, 7 of them are part of genes with a potential role in neurological diseases. The most significantly associated to the clustering is rs35866326, p-value around 10^{-4} for cluster 2 ($19/48$ cases with at least one copy of the alternative allele, compared to $52/348$ in the other clusters). This SNP has been associated with susceptibility to PD (Maraganore *et al.*, 2005, 2006; Goris *et al.*, 2006) although this result has not been replicated by others (Li *et al.*, 2006; Farrer *et al.*, 2006). These controversial results might be due to the fact that this gene is associated with a particular subtype of PD, as is shown here with an association with cluster 2 only. However, an unselected variant does not mean that it may not be associated with the disease subtype. It may be associated but not enough to bring more information relative to the clustering.

7 Conclusion

7.1 Synthesis and results

We proposed a model-based method for disease subtyping where the information comes from both short longitudinal data with varying observation times, as are often clinical follow-up data, and from high-dimensional quantitative data, such as genotyping data. Unlike in most multi-view clustering methods, the data are processed in a non-symmetrical way by integrating genetic data in the clustering via multinomial logistic weights. A Lasso penalty on the logistic regression parameters allows to deal with the high-dimensionality of the genotyping data while exhibiting a short list of genetic factors potentially involved in the typology of the disease.

An experiment on artificial data validates the proposed inference and model selection approach and shows its superiority in finding latent subtypes of the disease as well as influent genetic factors when compared to a clustering on clinical data followed by an association study. Using our method on clinical and genetic data from a cohort of patients with Parkinson's disease allows to characterize 4 distinct subtypes and 15 genetic factors with a potential impact on the subtyping. Of these 15 SNPs, the most significant SNP is already associated with PD. Half of the others belong to genes suspected to be implied in neurological diseases. Being able to recover such results corroborates the relevance of our approach in a real setting.

7.2 Perspectives

Several aspects can be considered in future works.

Replication The statistical analysis presented here uses a relatively small sample size and it may thus be of interest to try a replicate and confirm using independent cohorts.

[3] The confidence intervals are computed from the Hessian matrix provided by the `R` function `nnet::nnet` (Ripley *et al.*, 2016).

Modeling of data If the objective of the subtyping is to predict the evolution of the patient's symptoms and if more data are available for each patient, one can consider taking into account the temporal dynamics specific to each individual in a more refined way, for example by using a Gaussian process as done by Schulam and Saria (2015). In addition, if one chooses to focus on some correlated clinical variables, a multivariate version of the proposed model could be considered, but this is complicated by the functional nature of the data (t_{ij} times are different from one individual i to another). Regarding the genetic data, in order to resort to a lighter elimination preprocessing in a very high-dimensional setting (several millions of SNPs), it may be useful to summarize the data, for instance by aggregating SNPs in linkage disequilibrium blocks (Guinot *et al.*, 2018).

Association study with genetic data Finally, the proposed method does not dispense with the need for a more traditional association study afterward and leaves the opportunity of studying further potential associations between the genetic markers extracted in the variable selection process. In this perspective, a correction for multiple testing can be performed to assess the likelihood that the SNPs identified with our method actually have an impact on the disease typology. This correction should take into account the fact that the Lasso selection is performed on a high number of SNPs and that the tests are performed on a subgroup of those SNPs. In this case post-hoc inference tests may be useful (Goeman *et al.*, 2011).

Acknowledgements and funding

This work was carried out within the MeMoDeeP project funded by the ANR and lead by Maria Martinez. The methodological reflections of this work were fed by sustained exchanges with the members of this project, notably Pierre Neuvial. The data used here are from the DIGPD cohort sponsored by the Assistance Publique Hôpitaux de Paris and funded by the French Ministry of Health (PHRC AOM0810). We would like to thank the DIGPD study group for collecting the data, and the patients who participated to this cohort. The data have been adapted for this specific work by Jean-Christophe Corvol, Samir Bekadar, Fabrice Danjou, Graziela Mangonne, Alexis Elbaz and Fanny Artaud. We would also like to thank Agathe Guilloux for her help and enriching discussions around this work as well as Franck Samson for his work on the web interface of the method (in construction).

Author's contributions

Marie Courbariaux did all the experiments, contributed to the design of the algorithm, contributed to the development of the `R` package, and wrote most of the article. Marie Szafranski and Cyril Dalmasso contributed to the design of the algorithm and contributed to the writing of the paper. Christophe Ambroise contributed to the design of the algorithm, contributed to the developement of the `R` package, coordinated the experiments and the writing of the paper. Regarding the application on Parkinson, Jean-Christophe Corvol, Fabrice Danjou and Samir Bekadar provided the data, Fabrice Danjou filtered the genetic data and Jean-Christophe Corvol interpreted the results.

References

Bolte, J. *et al.* (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, **146**(1-2), 459–494.

Bush, W. S. and Moore, J. H. (2012). Genome-wide association studies. *PLOS Computational Biology*, **8**(12), e1002822.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, **14**(3), 315–332.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica*, **22**(2), 555–574.

Corvol, J.-C. *et al.* (2018). Longitudinal analysis of impulse control disorders in parkinson disease. *Neurology*, **91**(3), e189–e201.

Courbariaux, M. *et al.* (2020). *DiSuGen: Disease Subtyping with integrated Genetic association*. `R` package.

Dempster, A. P. *et al.* (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Farrer, M. J. *et al.* (2006). Genomewide association, parkinson disease, and park10. *The American Journal of Human Genetics*, **78**(6), 1084–1088.

Fop, M. *et al.* (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, **12**, 18–65.

Friedman, J. *et al.* (2010). Regularization paths for GLM via coordinate descent. *Journal of Statistical Software*, **33**(1), 1.

Fu, L. *et al.* (2020). An overview of recent multi-view clustering. *Neurocomputing*, **402**, 148–161.

Genolini, C. *et al.* (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, **65**(4), 1–34.

Goeman, J. J. *et al.* (2011). Multiple testing for exploratory research. *Statistical Science*, **26**(4), 584–597.

Goris, A. *et al.* (2006). No evidence for association with Parkinson disease for 13 single-nucleotide polymorphisms identified by whole-genome association screening. *American journal of human genetics*, **78**(6), 1088.

Gormley, Isobel Claire Früwirth-Schnatter, S. (2019). Mixture of experts models. In S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert, editors, *Handbook of mixture analysis*, chapter 12, pages 271–308. CRC Press.

Grun, B. and Leisch, F. (2008). FlexMix: finite mixtures with concomitant variables and varying and constant parameters. version 2.

Guinot, F. *et al.* (2018). Learning the optimal scale for GWAS through hierarchical SNP aggregation. *BMC bioinformatics*, **19**(1), 1–14.

Hastie, T. *et al.* (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hayes, B. (2013). Overview of statistical methods for Genome-Wide Association Studies (GWAS). In *Genome-wide association studies and genomic prediction*, pages 149–169. Springer.

Huang, S. *et al.* (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in Genetics*, **8**, 84.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**(3), 231–255.

Kim, S. *et al.* (2017). Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics*, **18**(1), 165–179.

Kristensen, V. N. *et al.* (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, **14**(5), 299–313.

Lee, M. *et al.* (2010). Biclustering via sparse singular value decomposition. *Biometrics*, **66**(4), 1087–1095.

Lewis, S. *et al.* (2005). Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry*, **76**(3), 343–348.

Li, Y. *et al.* (2006). A case-control association study of the 12 single-nucleotide polymorphisms implicated in Parkinson disease by a recent genome scan. *The American Journal of Human Genetics*, **78**(6), 1090–1092.

Ma, S. and Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, **9**(5), 392–403.

Maraganore, D. M. *et al.* (2005). High-resolution whole-genome association study of Parkinson disease. *The American Journal of Human Genetics*, **77**(5), 685–693.

Maraganore, D. M. *et al.* (2006). Response from Maraganore et al. *American journal of human genetics*, **78**(6), 1092.

Mariette, J. and Villa-Vialaneix, N. (2017). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, **34**(6), 1009–1015.

Mortier, F. *et al.* (2015). Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics*, **26**(1), 39–51.

Ndiaye, E. *et al.* (2017). Gap safe screening rules for sparsity enforcing penalties. *The Journal of Machine Learning Research*, **18**(1), 4671–4703.

Nguyen, H. *et al.* (2018). Pinsplus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**(336), 846–850.

Rentzsch, P. *et al.* (2018). Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, **47**(D1), D886–D894.

Ripley, B. *et al.* (2016). Package nnet. *R package*, pages 7–3.

Rohart, F. *et al.* (2017). MixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, **13**(11), e1005752.

Schulam, P. and Saria, S. (2015). A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756.

Shen, R. *et al.* (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**(22), 2906–2912.

Shen, R. *et al.* (2010). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis (correction). *Bioinformatics*, **26**(2), 292–293.

Shen, R. *et al.* (2013). Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, **7**(1), 269–294.

Sun, J. *et al.* (2014). Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC genetics*, **15**(1), 73.

Sun, J. *et al.* (2015). Multi-view sparse co-clustering via proximal alternating linearized minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 757–766, Lille, France.

van der Nest, G. *et al.* (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, **43**, 100323.

Yi, X. and Caramanis, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, volume 28, pages 1–9.

Zhao, B. *et al.* (2009). Multiple kernel clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 638–649.