A mixture model with logistic weights for disease subtyping with integrated genome association study

Marie Courbariaux, Christophe Ambroise, Cyril Dalmasso, Marie Szafranski, Memodeep Consortium

A mixture model with logistic weights for disease subtyping with integrated genome association study

Marie Courbariaux[1⋆], Christophe Ambroise[1], Cyril Dalmasso[1], Marie Szafranski[1], and the MeMoDeep Consortium[*] [1,2,3,4]

[1] *Laboratoire de Mathématiques et Modélisation d'Évry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA.*

[2] *Institut du Cerveau et de la Moelle épinière (ICM), Paris, France*

[3] *INSERM U563, Toulouse, France*

[4] *Institut de Mathématiques de Toulouse; UMR5219 Université de Toulouse; CNRS UPS IMT, Toulouse, France.*

[⋆] *Corresponding author: marie.courbariaux@gmail.com*

Abstract

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.

This work proposes an original method for disease subtyping from both longitudinal clinical variables and genetic markers via a mixture of regressions model, with logistic weights function of a potentially large number of genetic variables. In order to address these large-scale problems, variable selection is an essential step. We thus propose to discard genetic variables that may not be relevant for clustering by maximizing a penalized likelihood via a Classification Expectation Maximization algorithm.

The proposed method is validated on simulations. A data set 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.

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

[*] MeMoDeep: Methods and Models for Deep Screening of subphenotypes in Parkinson's Disease. Christophe Ambroise, Samir Bekadar, Jean-Christophe Corvol, Marie Courbariaux, Cyril Dalmasso, Fabrice Danjou, Maria Martinez, Pierre Neuvial, Marie Szafranski.

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 data sets are now available, including for example patients follow-up and genotyping data.

A simple way to find genetically meaningful subtypes would be a two-step approach: i) a subtyping of the disease based on clinical data followed by ii) an analysis of the genetic associations in each subtype. The first step consists in clustering of clinical data. Such 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 in Jacques and Preda (2014). In the second step, one might exhibit genetic associations that explain the clusters using them as phenotypes in standard approaches devoted to GWAS which usually involve statistical procedures based on hypothesis testing (see for instance (Bush and Moore, 2012) or (Hayes, 2013)). Another way to reveal such genetic associations resorts to supervised learning methods, such as (multinomial) logistic regression, with a feature selection procedure (Ma and Huang, 2008, and references therein).

Concomitant clustering of clinical and genomic data represents an attractive alternative to the two-step scheme. In this context, feature or variable selection strategies are mandatory to make the problem computationally tractable. Also, it may help to identify relevant variables more directly than with dimensionality reduction methods. Two recent surveys focus on the problem of variable selection for continuous and categorical data (Bouveyron and Brunet-Saumard, 2014; Fop *et al.*, 2018). For functional or longitudinal data, James and Sugar (2003) introduced the method *fclust* which enforces sparsity on the cluster means. More recently, McNicholas and Murphy (2010) developed a framework based on Gaussian mixtures using a constrained modified Cholesky decomposition of the group covariance matrices to achieve sparsity. Note that there is less work devoted to sparse model-based clustering for functional or longitudinal data, the usual approach relying on dimensionality reduction (Jacques and Preda, 2014), which may be efficient but less convenient for interpretation.

Multi-view learning is a framework coming from the machine learning community which allows to solve problems by considering feature sets of different nature. The survey of Sun (2013) gives an overview of this framework together with theoretical considerations. The works of Bickel and Scheffer (2004) and Tzortzis and Likas (2009) relying on (non sparse) model-based approaches is representative of this framework. Muliple Kernel Learning framework (Gönen and Alpaydin, 2011) and its clustering counterpart (Zhao *et al.*, 2009) are another avatar of multi-view clustering approach. Sun *et al.* (2014) propose a sparse multi-view co-clustering method based on Sparse Singular Values Decomposition (SSVD) (Lee *et al.*, 2010). This recent work is dedicated to disease subtyping with clinical and genomic information. Sun *et al.* (2015) build on this work providing convergence guarantees using the proximal alternating linearized minimization algorithm of Bolte *et al.* (2014). Despite a wide range of existing methods in various applications, there is no sparse multi-view model-based clustering method available to our knowledge in this community.

In cancer research, many statistical methodologies have emerged to analyze 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 recent review of multi-omics integration tools. 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 the work of Shen *et al.* (2013) and Kim *et al.* (2017).

When dealing with both genetic and clinical longitudinal data, a two-step approach represents a straightforward solution to the problem of genetic subtyping. Yet, the clustering step of such method does not take benefit from the genetic data in devising the groups. As a consequence, there is no guarantee regarding the connection between the genetic information and the clinical clusters. Integrative-like approaches seem more adapted. However, none of the integrative and multi-view methods explicitly address how to deal with multiple data of different nature (such as longitudinal data and counting data for instance). Indeed, merging the datasets without a clever pretreatment to smooth the differences between the nature of variables is certainly not appropriate in the sense that only one kind of information may influence all the clustering. To get around this aspect, most methods use representations of data obtained from projections on subspaces (PCA or SVD for example). However, distorting the initial information may significantly complicate the posterior validation of the extracted features since they correspond to a composition of several variables.

In the light of the previous considerations, our proposal is to cluster the clinical data by estimating a mixture of regressions model, with the weights depending on the genetic variables. Involving the genetic information leads to problems of large dimension where a variable selection strategy becomes essential. Therefore, we propose to discard the genetic variables that might be irrelevant for the clinical clustering thanks to a sparse constraint on the parameters involved in the logistic weights. The proposed scheme is at the crossroad of the possibilities presented above, namely i) the clustering of clinical data with a posterior genetic analysis and ii) the concomitant clustering of clinical and genetic data. Indeed, the clustering of the clinical data is guided by the genetic markers through the weights of the mixture model while being adapted to the longitudinal nature of the data.

2 Mixture of regressions model with logistic weights

We propose a general model that may apply when the disease has to be subtyped from several longitudinal variables, such as patient symptoms recorded from their follow-up.

Thereafter Y_{vij} stands for the v^{th} clinical variable under consideration for disease subtyping observed during the j^{th} visit of patient i, V is the number of variables under consideration, K is the number of disease subtypes, L is the number of genetic markers and I is the size of the cohort.

The following regression model with logistic weights is considered:

$$\begin{aligned}(Y_{vij}|Z_{ik}=1) &= \alpha_{v0k}+\alpha_{v1k}t_{ij}+\cdots+\alpha_{vN_pk}t_{ij}^{N_p}\\ &\qquad + \sigma_{vk}\varepsilon_{vij},\ \varepsilon_{vij}\underset{iid}{\sim}\mathcal{N}\left(0,1\right), \\ Pr(Z_{ik}=1) &= \frac{e^{\omega_{0k}+\boldsymbol{\omega}_k^T\mathbf{G}_i}}{\sum_{k'=1}^{K}e^{\omega_{0k'}+\boldsymbol{\omega}_{k'}^T\mathbf{G}_i}},\end{aligned} \tag{1}$$

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,
- N_p is the maximum polynomial degree we consider in the regression ($N_p = 2$ is generally sufficient),
- $\mathbf{G}_i$ is its genotype vector,
- Z_{ik} is the indicator variable of patient i belonging to the class k,
- $(\boldsymbol{\omega}_k)_{k\in\{1,\ldots,K\}}$, $(\alpha_{vpk})_{v\in\{1,\ldots,V\},p\in\{0,\ldots,N_p\},k\in\{1,\ldots,K\}}$ and $(\sigma_{vk})_{v\in\{1,\ldots,V\},k\in\{1,\ldots,K\}}$ are parameters or vectors of parameters to be estimated.
- $\omega_{l1} = 0$, $\forall l \in \{1, ..., L\}$ for the sake of identifiability.

Also, $\varepsilon_{vij}\underset{iid}{\sim}\mathcal{N}\left(0,1\right)$ implies some conditional independence assumptions between variables, patients and visits when the class is known. Indeed, 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.

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

If the Gaussian hypothesis does not apply to the variable v, one may consider Poisson or logistic regression instead with no substantial additional cost.

Note that a similar model is used for disease subtyping, though without genotyping data, by Schulam and Saria (2015). This model could also be viewed as a Mixture Of Experts (MOE) close to the work of Jordan and Jacobs (1994), in which however, the variables involved in the logistic weights are the same as those involved in the main regression.

3 Model selection

We combine two model selection strategies in order to select the number of polynomial degrees, the number of clusters and the genetic markers that are involved in the mixture.

In order to select the most appropriate number of subtypes (K) and of polynomial degrees in the main regressions (N_p), the Bayesian Information Criterion (BIC) seems convenient.

A selection on the variables in $\mathbf{G}$ is needed, since we suspect that many have no influence on the disease phenomenology. Since $\mathbf{G}$ may be very large (about a few millions of markers after genotype imputation in the cases we consider), classical backward-stepwise methods (such as the one described by Maugis *et al.* (2009)) would result in non affordable computational time. In the case where there are a lot of possible regressors, and one suspects that many are not necessary to describe the phenomenon, one may use a Lasso-type penalization. Here, as Khalili (2010) in the case of MOE models, this selection takes action within an Expectation-Maximization (EM) algorithm.

4 EM algorithm with integrated Lasso inference

The inference of such a model with latent variables (here, the Zs) can be classically conducted with an Expectation Maximization algorithm (EM algorithm, Dempster *et al.*, 1977). We used a modified version of this algorithm in order to maximize a (Lasso-type) penalized likelihood instead of the likelihood. This modification does not compromise the convergence of the EM algorithm (Green, 1990).

At the $(q+1)^{th}$ iteration of the modified EM algorithm, one maximizes the expected and penalized complete-data log-likelihood:

$$\mathbb{E}_{\left(\mathbf{Z}|\theta^{(q)},\mathbf{Y}=\mathbf{y}\right)}\left\{\mathcal{L}\left(\mathbf{y},\mathbf{Z};\theta\right)-Pen(\boldsymbol{\omega})\right\},$$

where $\theta=(\boldsymbol{\omega},\boldsymbol{\alpha},\boldsymbol{\sigma})$ is the vector of all model parameters, $Pen(\boldsymbol{\omega})$ is the penalty and $\mathcal{L}\left(\mathbf{y},\mathbf{Z};\theta\right)$ is the complete-data log-likelihood:

$$\mathcal{L}\left(\mathbf{y},\mathbf{Z};\theta\right)=\sum_i\sum_k Z_{ik}\left(\log\left\{Pr\left(Z_{ik}=1;\boldsymbol{\omega}_k\right)\right\}\sum_v\sum_j\log\left\{\phi_{vijk}(y_{vij};\theta)\right\}\right),$$

where $\phi_{vijk}(y;\theta)=f\left(Y_{vij}=y|Z_{ik}=1;\boldsymbol{\alpha}_{vk},\sigma_{vk}\right)$ and $f\left(X=x;\zeta\right)$ is the probability density function of X in x with parameters ζ. The penalty, $Pen(\boldsymbol{\omega})$ can for instance have the following form: $Pen(\boldsymbol{\omega})=\lambda\sum_k\|\boldsymbol{\omega}_k\|_1$, the Lasso penalty, where $\|\cdot\|_1$ is the l_1-norm and $\lambda>0$ is a parameter to choose.

To maximize the expected and penalized complete-data log-likelihood, each iteration is divided into 2 steps: an expectation step (E) followed by a maximization step (M).

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

$$\begin{aligned}\tau_{ik}^{(q+1)} &= \mathbb{E}\left\{Z_{ik}|\mathbf{Y}=\mathbf{y};\theta^{(q)}\right\} \\ &= \frac{Pr\left(Z_{ik}=1;\boldsymbol{\omega}_k^{(q)}\right)\prod_v\prod_j\phi_{vijk}(y_{vij};\theta^{(q)})}{\sum_{k'=1}^{k'=K} Pr\left(Z_{ik'}=1;\boldsymbol{\omega}_{k'}^{(q)}\right)\prod_v\prod_j\phi_{vijk'}(y_{vij};\theta^{(q)})}.\end{aligned}$$

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

$$\begin{aligned}\theta^{(q+1)} = \underset{\theta}{\text{argmax}} \sum_i \sum_k \tau_{ik}^{(q+1)} \Bigg(& \log\left\{Pr\left(Z_{ik}=1;\boldsymbol{\omega}_k\right)\right\} \\ & \sum_v \sum_j \log\left\{\phi_{vijk}(y_{vij};\theta)\right\}\Bigg) - Pen(\boldsymbol{\omega}).\end{aligned}$$

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

$$\frac{1}{I}\sum_i\sum_k \tau_{ik}^{(q+1)} \log\left(\frac{e^{\omega_{0k}+\boldsymbol{\omega}_k^T\mathbf{G}_i}}{\sum_{k'=1}^{K} e^{\omega_{0k'}+\boldsymbol{\omega}_{k'}^T\mathbf{G}_i}}\right) - Pen(\boldsymbol{\omega}). \tag{2}$$

When resorting to Classification EM (CEM, Celeux and Govaert (1992)), $\tau_{ik}^{(q+1)}$ are replaced by the indicator variable of the most likely class for patient i at iteration $(q+1)$. This maximization problem corresponds to the multinomial logistic regression problem with a penalty. It can be addressed by a partial Newton algorithm as for instance in the `glmnet` R package (Friedman *et al.*, 2010). The λ parameter is chosen by cross-validation so that the likelihood of the multinomial logistic model is maximized.

To implement our method, we build an adapted concomitant variable driver making use of `glmnet` within the `FlexMix` R package (Grun and Leisch, 2008). The `FlexMix` package proposes a (C)EM algorithm adapted to multinomial logistic weights mixture models.

To avoid (negative) bias due to the penalization in the parameter estimation, we re-estimate the selected ω parameters at the end of the EM algorithm in order to get the maximum likelihood estimates. A BIC selection finally arbitrates among initializations.

This implementation of the proposed method (named `DiSuGen`) as well as an R Markdown tutorial can be downloaded at https://github.com/MCour/DiSuGen.

5 Numerical illustrations on artificial data

The proposed estimation and model selection procedures are first evaluated on artificial data. The goals are to check whether the CEM algorithm produces a good estimation of the parameters and whether the model selection method chooses the appropriate model.

Experimental setting

Artificial data are simulated according to the model (1) with $K = 3$ clusters, $V = 4$ clinical variables, $I = 396$ patients, times $t_{i,j}$ 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, $N_p = 1$ polynomial degree in the regression, $n_i = 3$ observations per patient and $L = 2657$ genetic markers, only 10 of which having an influence on the clustering: $\boldsymbol{\omega}_{(2,3,4),2} = \boldsymbol{\omega}_{(5,6,7),3} = 2$, $\boldsymbol{\omega}_{(5,6,7),2} = -1$ and $\boldsymbol{\omega}_{(1,8,9,10),3} = -2$. Those genetic markers are the ones that will be used in section 6. The parameters α and σ are chosen so as to be realistic with regard to Parkinson's Disease (PD) clinical data.

For each simulation, the proposed CEM algorithm is launched with a number of clusters corresponding to the simulated one ($K = 3$) and a Lasso penalty. This experiment is repeated 100 times. 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.

The results are compared with the ones got with the corresponding two-step method, i.e. without any use of genetic information in the clustering process, what corresponds to constant weights of the clusters, $\mathbb{P}(Z_{ik} = 1) = \pi_k$. In this case, the Lasso-penalized multinomial logistic regression is performed afterward to get genetic association results. When possible, the proposed method is also compared to the k-means method. This algorithm 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).

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. Most clusters are well identified. When making no use of genetic information within the clustering, the algorithm globally achieves lower ARI.

In theory, a better clustering ability can be expected from the algorithm making use of genetic data as illustrated by the corresponding oracle results ("oracle integrative"). Those results are obtained by using the true parameters values to predict the clusters the patients

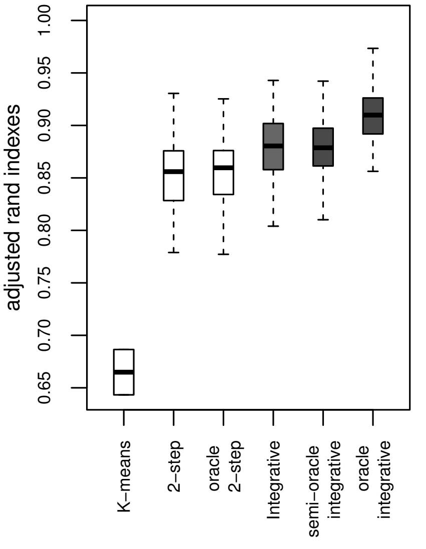

Figure 1: Boxplots of the ARI over 100 simulations with a k-means algorithm, with no use of genetic information (2-step method), with the use of genetic information (integrative method) and corresponding oracles.

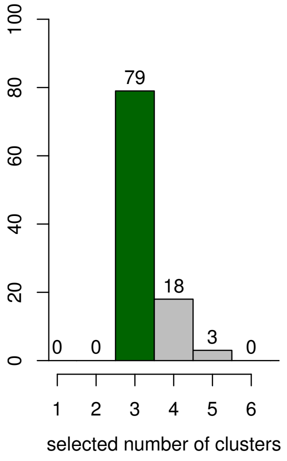

Figure 2: Histogram of the estimated number of clusters according to the BIC over 100 simulations.

belong to. This improvement in cluster prediction does not come from a better estimation of the $\boldsymbol{\omega}$ parameters as illustrated by the "semi-oracle integrative" results. Those results are obtained using the true $\boldsymbol{\omega}$ parameters and estimating the other parameters.

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 (two-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 3.

Variable selection within the logistic regression. Figure 4 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

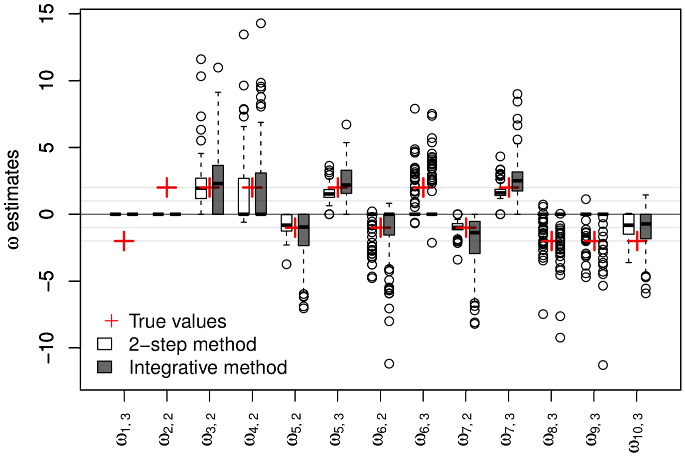

Figure 3: Non-negative ω parameters estimates over 100 simulations and respective true values.

selection rates obtained using the two-step method are lower, which underlines the interest of the proposed integrative approach. The 2 lasting 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 ω parameters are set closer to zero (not shown).

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 2. The correct number of clusters is most of the time selected: 79 times over 100.

6 Application to Parkinson's Disease subtyping

The proposed method is then applied to Parkinson's Disease (PD) subtyping. This disease is known to have several subtypes. This has given rise to numerous studies among which (Lewis *et al.*, 2005).

Data description

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

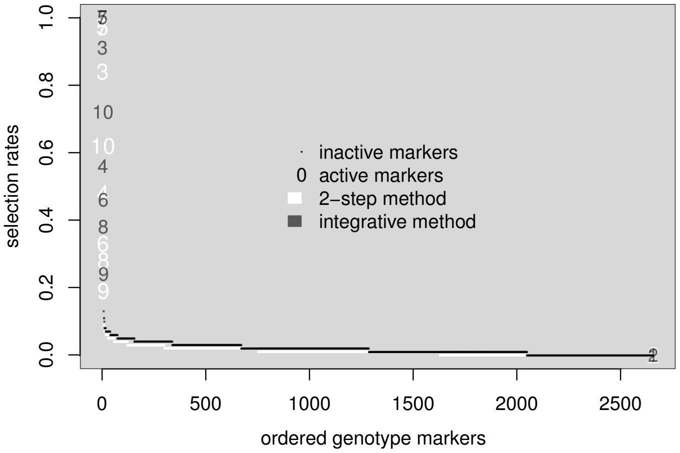

Figure 4: Selection rates of the genetic variables over 100 simulations. The 10 variables that should be selected are represented by their respective number (l).

than 6 years prior to the beginning of the study).

Clinical data were collected at inclusion and then at yearly clinical follow up (during one to seven years). They include a number of scores evaluating the progression of the disease. Two of them are used, ones that 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. The 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.

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[1] 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.

[1] Combined Annotation Dependent Depletion score, this score evaluates the deleteriousness of variants in the human genome.

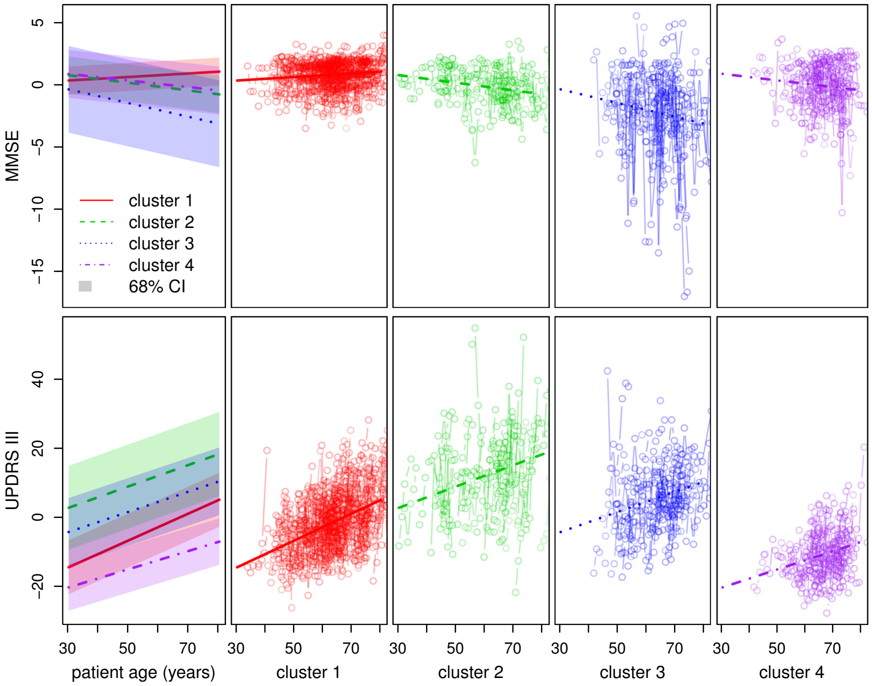

Figure 5: Clustering results with regard to the clinical variables.

Results

Model selection results. In order to warranty a good interpretability of the results, a maximal number of $K = 4$ clusters was allowed. Moreover, 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 5. Remember that the variables represented here are residuals of a fitted linear model. On the figure, 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 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. Genetic association results are illustrated Figure 6. The 95% confidence intervals associated to the ω parameters are computed from the Hessian

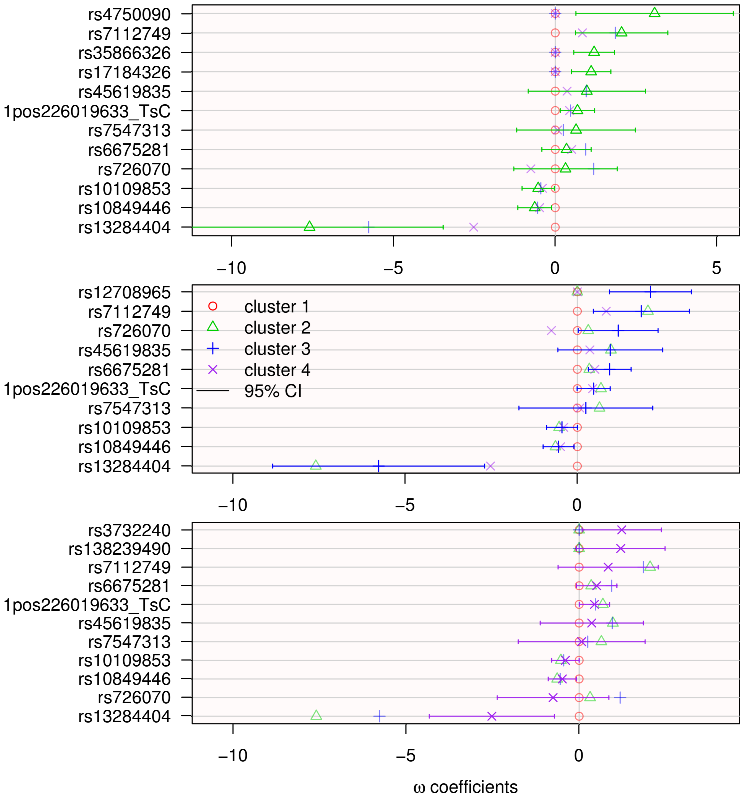

Figure 6: Genetic association results: estimated logistic regression (ω) parameters.

matrix (provided by the R nnet::nnet function (Ripley *et al.*, 2016)). Note that 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. rs35866326 is the most significantly associated to the clustering, with a 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 Parkinson's disease (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 and outlook

We propose a model-based method for disease subtyping 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 two types of 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 permits to get sparsity by 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) method and shows its superiority in finding latent subtypes of the disease and in finding the influent genetic factors when compared to the corresponding two-step method (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 (PD) 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.

The statistical analysis presented here may be underpowered due to the relatively small sample size of the dataset. It may thus be of interest to try a replication in independent cohorts. Moreover, a correction for multiple testing has to 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. Moreover, in order to use this method with a very high-dimension genetic dataset (several millions of SNPs), it may be necessary to summarize the data, for instance by aggregating SNPs in linkage disequilibrium blocks (Guinot *et al.*, 2017; Dehman *et al.*, 2015). In short, it leaves the opportunity of studying further potential associations between the genetic markers extracted in the variable selection process. The proposed method does not dispense with the need for a more traditional association study afterward.

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).

Finally, 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).

Acknowledgements and funding

This work was carried out within the MeMoDeep (Methods and Models for Deep Screening of subphenotypes in Parkinson's Disease) project funded by the ANR (ANR-16-CE37-0008). The methodological reflections of this work were fed by sustained exchanges with the members of this project, notably Samir Bekadar, Jean-Christophe Corvol, Fabrice Danjou, Maria Martinez and 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).

References

Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, pages 19–26, Washington, DC, USA.

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

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, **71**, 52–78.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, **8**(12), 1–11.

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

Corvol, J.-C. *et al.* (2018). Longitudinal analysis of impulse control disorders in Parkinson disease. *Neurology*.

Dehman, A. *et al.* (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC bioinformatics*, **16**(1), 148.

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. *American journal of human genetics*, **78**(6), 1084.

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 generalized linear models via coordinate descent. *Journal of statistical software*, **33**(1), 1.

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

Gönen, M. and Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, **12**, 2211–2268.

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.

Green, P. J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452.

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

Guinot, F. *et al.* (2017). Learning the optimal scale for GWAS through hierarchical snp aggregation. *arXiv preprint arXiv:1710.01085*.

Hayes, B. (2013). *Overview of Statistical Methods for Genome-Wide Association Studies (GWAS)*, pages 149–169. Humana Press.

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.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**(462), 397–408.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, **6**(2), 181–214.

Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, **38**(4), 519–539.

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.

Maugis, C. *et al.* (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**(3), 701–709.

McNicholas, P. D. and Murphy, B. T. (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics*, **38**(1), 153–168.

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

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

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. *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.

Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*, **23**(7-8), 2031–2038.

Tzortzis, G. and Likas, A. (2009). Convex mixture models for multi-view clustering. In *Proceedinsg of the 19th International Conference of Artificial Neural Networks*, pages 205–214, Berlin, Heidelberg.

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