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] *Laboratoire de Mathématiques et Modélisation d'Évry (LaMME), Université d'Evry Val d'Essonne, UMR CNRS 8071, ENSIIE, USC INRA.*
[⋆] *Corresponding author: marie.courbariaux@gmail.com*

Abstract

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. The approach is applied to a data set from a cohort of Parkinson's disease patients. Several subtypes of the disease as well as genetic variants potentially having a role in this typology have been identified.

KEYWORDS: Genetic association, Sub-typing, Parkinson's disease, Mixing model with logistic weights, Lasso Penalization.

1 Introduction

It is becoming increasingly difficult to identify new genetic associations in complex non mendelian disease while there is still a significant part of heritability to be explained.

[*] 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.

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 patient follow-up and genotyping data.

1.1 Background

In this section, we review different ways to identify subgroups of interests using both clinical and genomic data, and discuss the pros and cons of each approach.

1.1.1 Clustering of clinical data with a posterior genetic analysis

A first way to proceed would be a two-step approach with i) a subtyping of the disease based on clinical data and then ii) an analysis of the genetic associations in each subtype.

Clustering of clinical data. In a first step, a clustering of clinical data can be performed. 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 in Jacques and Preda (2014a) 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 (classical) clustering step on those descriptors. For instance, Abraham et al. (2003) first represent each input function by calculating its approximation on a fixed truncated basis and then implement the k-means algorithm on the coordinates of the representation. Similarly, Rossi et al. (2004) submit those coordinates to a Self-Organizing Map after an appropriate transformation.
- *Non-parametric methods*, such as k-means, with distance metrics adapted to longitudinal data. For example, Ieva et al. (2012) make recourse of a k-means procedure with a distance involving the wavelet estimate of the first derivative of the curves.
- Finally, *model-based methods* appear to be the most adapted approach to deal with short longitudinal data, as often encountered when dealing with medical follow-ups. Jacques and Preda (2014b) propose a mixture model-based on the assumption of normality of the principal components of a multivariate functional principal components analysis. Alternatively, Schulam and Saria (2015) propose a mixture model involving both basis expansions and a Gaussian Process.

Analysis of clinical clusters with genomics. In a 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 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).

1.1.2 Concomitant clustering of clinical and genomic data

An alternative scheme to cluster clinical data using genomic features would be to gather both kind of information, leading to consider a huge number of variables. In this context, feature or variable selection strategies are mandatory to make the problem tractable computationaly. Also, it may help to identify relevant variables more directly than with dimensionality reduction methods. In the following, we will pay attention to sparse model-based clustering approaches, this latest family being well suited to longitudinal data (Jacques and Preda, 2014a).

Sparse model-based clustering. Provided that clinical and genomic information are merged into a new dataset, we could resort to sparse model-based clustering which has been broadly investigated the past decade. In particular, two recent surveys focus on the problem of variable selection for continuous and categorial data (Bouveyron and Brunet-Saumard, 2014; Fop and Murphy, 2017). 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, 2014a), which may be efficient but less convenient for interpretation.

Sparse multi-view clustering. Multi-view learning is a framework coming from the machine learning community which allows to solve problems considering different feature sets. The survey of Sun (2013) gives an overview of this framework together with theoretical considerations. Especially, we should mention the works of Bickel and Scheffer (2004) and Tzortzis and Likas (2009) relying on (non sparse) model-based approaches. Also, one should complete this survey with the Multiple Kernel Learning framework (Gönen and Alpaydin, 2011) and its clustering counterpart (Zhao et al., 2009). In addition to these references, we can also report recent works dedicated to disease subtyping with clinical and genomic information that fall into the scope of sparse multi-view clustering, though they do not use model-based approaches. Sun et al. (2014) propose a multi-view co-clustering method based on Sparse Singular Values Decomposition (SSVD) (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). 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.

Sparse 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 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 (Shen et al., 2013; Kim et al., 2017).

1.1.3 Limits

To identify subtypes of diseases, one may have clinical data of longitudinal nature on one hand and genomic data which are of categorial nature on the other hand. A well-suited clustering of clinical data followed with a posterior genetic analysis of the clusters obtained may present limitations since the step of clustering does not take benefit from the genomic data. As a consequence, there is no guarantee regarding the connection between the genomic information and the clinical clusters. Integrative-like approaches seem more adapted for this problem. Yet, none of the integrative and multi-view methods explicitly adress 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 unappropriated 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.

1.2 Proposed approach

In the light of the previous considerations, we sketch an alternative path to adress the problem of subtyping when clinical and genomic information are available. Our proposal is to cluster the clinical variables by estimating the weights of a multinomial logistic regression model, with the weights depending on the genetic variables. 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 (Jordan and Jacobs, 1994), in which, however, the variables involved in the logistic weights are the same as those involved in the main regression.

Involving the genetic information leads to problems of large dimension where a variable selection strategy becomes essential, as seen in Section 1.1. Therefore, we propose to discard

the genetic variables that might be unrelevant for the clinical clustering thanks to a sparse constraint on 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 genomic data. Indeed, the clustering of the clinical data is guided by the genetic markers through the weights of the model while being adapted to the nature of the longitudinal data. Also, it leaves the opportunity of studying further potential associations between the genetic markers extracted in the variable selection process and the clinical subtyping.

2 A mixture of regressions model with logistic weights

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

We denote by $Y_{v,i,j}$ the v^{th} clinical variable under consideration for disease subtyping observed during the j^{th} visit of patient i, by V the number of variables under consideration, by K the number of subtypes of the disease, by L the number of genetic elements and by I the size of the patients cohort.

We consider the following regression model with logistic weights:

$$\begin{aligned}(Y_{v,i,j}|Z_{i,k}=1) &= \alpha_{v,0,k}+\alpha_{v,1,k}t_{i,j}+\alpha_{v,2,k}t_{i,j}^2+\cdots+\alpha_{v,N_p,k}t_{i,j}^{N_p} \\ &\quad +\sigma_{v,k}\varepsilon_{v,i,j},\ \varepsilon_{v,i,j}\underset{iid}{\sim}\mathcal{N}(0,1), \qquad (1)\\ Pr(Z_{i,k}=1) &= \frac{e^{\omega_{0,k}+\boldsymbol{\omega}_k^T\mathbf{G}_i}}{\sum_{k'=1}^{K}e^{\omega_{0,k'}+\boldsymbol{\omega}_{k'}^T\mathbf{G}_i}},\end{aligned}$$

where:

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

$\varepsilon_{v,i,j}\underset{iid}{\sim}\mathcal{N}(0,1)$ 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 orthogonal 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 logistic weights allow for concomitant variables, such as genetic data, to guide as well the subtyping. If the Gaussian hypothesis does not apply to the variable v, one may consider the Poisson or the logistic regression instead with no substantial additional cost.

2.1 Lasso and BIC model selection

We combine two model selection strategies in order to select, on the one hand, the variants and, on the other hand, the number of polynomial degrees and the number of clusters that are involved in the mixture.

2.1.1 A Lasso penalty to select the logistic regression parameters

We want to operate a selection on the variables $\mathbf{G}$, since we suspect that many have no influence on the disease phenomenology. Since $\mathbf{G}$ may be very large (about a few millions of elements after genotype imputation in the cases we consider), classical backward-stepwise methods (such as the one described in 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.

2.1.2 A BIC to select the number of clusters and polynomial degrees

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. However, as in the case of mixed models (Delattre et al., 2014), the BIC can not directly apply in our case. Indeed, the effective sample size is not clearly defined: one may consider the number of patients, I, why some other may consider the number of observations, $N = \sum_{i=1}^{I} n_i$. We compute the BIC as follows:

$$BIC = -2LL + (N_p + 2)VK \log(N) + \log(I) \sum_{k=1}^{K} \|\boldsymbol{\omega}_k\|_0, \tag{2}$$

where $\|\boldsymbol{\omega}_k\|_0$ denotes the number of non-null elements of $\boldsymbol{\omega}_k$ including $\omega_{0,k}$ and LL is the log-likelihood. If V_{nl} additional non-longitudinal variables are included in the subtyping, then one can add the term $2V_{nl}K \log(I)$.

3 An 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 the (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:

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

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

$$\mathcal{L}\left(\mathbf{y},\mathbf{Z};\theta\right)=\sum_{i=1}^{i=I}\sum_{k=1}^{k=K}Z_{i,k}\Bigg(\log\left\{Pr\left(Z_{i,k}=1;\boldsymbol{\omega}_k\right)\right\}\\ \sum_{v=1}^{v=V}\sum_{j=1}^{j=n_i}\log\left\{f\left(Y_{v,i,j}=y_{v,i,j}|Z_{i,k}=1;\boldsymbol{\alpha}_{v,k},\sigma_{v,k}\right)\right\}\Bigg),$$

where n_i is the number of follow-up visits of patient i 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 λ 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:

$$\begin{aligned}\tau_{i,k}^{(q+1)}&=\mathbb{E}\left\{Z_{i,k}|\mathbf{Y}=\mathbf{y};\theta^{(q)}\right\}\\&=\frac{Pr\left(Z_{i,k}=1;\boldsymbol{\omega}_k^{(q)}\right)\prod_{v=1}^{v=V}\prod_{j=1}^{j=n_i}f\left(Y_{v,i,j}=y_{v,i,j}|Z_{i,k}=1;\boldsymbol{\alpha}_{v,k}^{(q)},\sigma_{v,k}^{(q)}\right)}{\sum_{k'=1}^{k'=K}Pr\left(Z_{i,k'}=1;\boldsymbol{\omega}_{k'}^{(q)}\right)\prod_{v=1}^{v=V}\prod_{j=1}^{j=n_i}f\left(Y_{v,i,j}=y_{v,i,j}|Z_{i,k'}=1;\boldsymbol{\alpha}_{v,k'}^{(q)},\sigma_{v,k'}^{(q)}\right)}.\end{aligned}$$

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

$$\theta^{(q+1)}=\underset{\theta}{\operatorname{argmax}}\sum_i\sum_k\tau_{i,k}^{(q+1)}\Bigg(\log\left\{Pr\left(Z_{i,k}=1;\boldsymbol{\omega}_k\right)\right\}\\ \sum_{v=1}^{v=V}\sum_{j=1}^{j=n_i}\log\left\{f\left(Y_{v,i,j}=y_{v,i,j}|Z_{i,k}=1;\boldsymbol{\alpha}_{v,k},\sigma_{v,k}\right)\right\}\Bigg)-Pen(\boldsymbol{\omega}).$$

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:

$$\frac{1}{I}\sum_i\sum_k \tau_{i,k}^{(q+1)} \log\left(\frac{e^{\omega_{0,k}+\boldsymbol{\omega}_k^T\mathbf{G}_i}}{\sum_{k'=1}^K e^{\omega_{0,k'}+\boldsymbol{\omega}_{k'}^T\mathbf{G}_i}}\right) - Pen(\boldsymbol{\omega}). \tag{3}$$

When resorting to Classification EM (CEM, Celeux and Govaert (1992)), $\tau_{i,k}^{(q+1)}$ are replaced with the indicator variable of the most likely class for patient i at iteration $(q+1)$, $Z_{i,k}^{(q+1)}$. Equation (3) becomes:

$$\frac{1}{I}\sum_i\sum_k Z_{i,k}^{(q+1)} \log\left(\frac{e^{\omega_{0,k}+\boldsymbol{\omega}_k^T\mathbf{G}_i}}{\sum_{k'=1}^K e^{\omega_{0,k'}+\boldsymbol{\omega}_{k'}^T\mathbf{G}_i}}\right) - Pen(\boldsymbol{\omega}). \tag{4}$$

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.

Implementation 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. We finally proceed to the proposed BIC selection to arbitrate among initializations.

4 Numerical illustrations on artificial data

We first check our estimation methods on artificial data. The goals are to check whether the CEM algorithm gives a good estimation of the parameters and whether our model selection method chooses the appropriate model.

Experimental setting

We simulate artificial data 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 elements, only 10 of which have 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 elements are the ones that will be used in section 5. We choose realistic α and σ parameters with regard to Parkinson's Disease (PD) clinical data.

For each simulation, we launch the proposed CEM algorithm with an imposed number of clusters corresponding to the simulated one ($K = 3$) and a Lasso penalty. This experiment is repeated 100 times. We initialize the estimation with 10 sets of starting values corresponding to 10 random assignments into $K = 3$ clusters and keep the one that leads to the lower BIC.

We compare our results with the one we get with the corresponding two-step method, i.e. without any use of genetic information in the clustering process and having recourse to the Lasso-penalized multinomial logistic regression afterward to get genetic association results. This results in constant weights of the clusters, $\mathbb{P}(Z_{i,k} = 1) = \pi_k$. When possible, we also compare our method to the k-means method. This method corresponds to a simple Gaussian mixture model with identical standard deviations in all clusters. To do so, we make recourse to a k-means method adapted to longitudinal data implemented in the R package `kml3d` (Genolini et al., 2015).

Results

Clustering ability We first compute the Adjusted Rand Index (ARI) for each simulation to check that the estimated clusters are close to the ones that are simulated. The results for our method are illustrated by the boxplot for the "integrative" method of Figure 1. Most clusters are thus adequate. 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 rand indexes results ("oracle integrative"). Those results are obtained by using the true parameters values to predict the clusters the patients 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.

Finally, the k-means algorithm is not able to recover the underlying classification as well: the corresponding rand indexes all range between 0.6 and 0.7. This was expected, since the differences between the clusters partly lies 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 in the main regressions The resulting estimated parameters are illustrated in Figure 3 on the example of the first clinical variable. All the biases are notably close to 0. Accounting for genetic information (integrative method) does

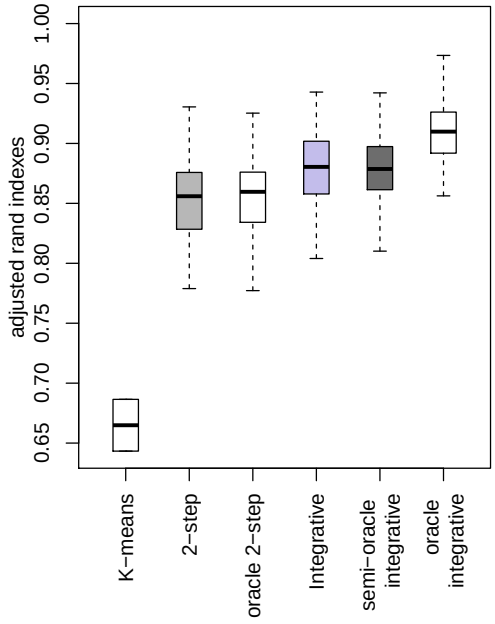

Figure 1: Boxplots of the ARI over 100 simulations with a k-means algorithm, with no use of genetic information (2-step method), with the use of genetic information (integrative method) and corresponding oracles.

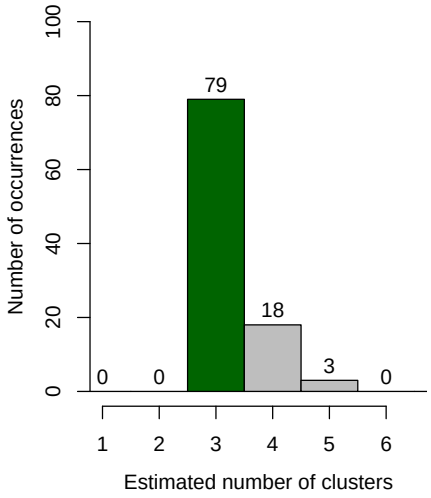

Figure 2: Histogram of the estimated number of clusters according to the BIC over 100 simulations.

not seem to improve notably the parameters estimation. The results are very similar for the 3 other clinical variables (not shown).

Variable selection and parameter estimation ability within the logistic regression Figure 4 illustrates the results of the Lasso selection procedure we propose 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 our selection method. The selection rates obtained using the two-step method are lower, which underlines the interest of the integrative approach we propose. The 2 lasting active elements do not vary between patients and can thus be replaced by any variant with a poor variation. Most of the inactive elements 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 elements is relatively high (2657). This selection performance decreases (as expected) as the ω parameters are set closer to zero (not shown). The sign of the estimated parameters are most of the time adequate for both approaches as illustrated Figure 5.

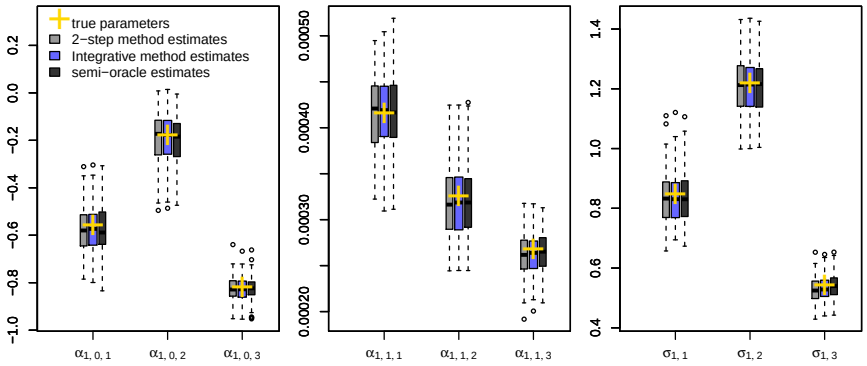

Figure 3: Estimated α and σ parameters over 100 simulations and respective true values for the first simulated clinical variable with no use and with the use of genetic information as well as with the semi-oracle (i.e. using the true $\boldsymbol{\omega}$ parameters).

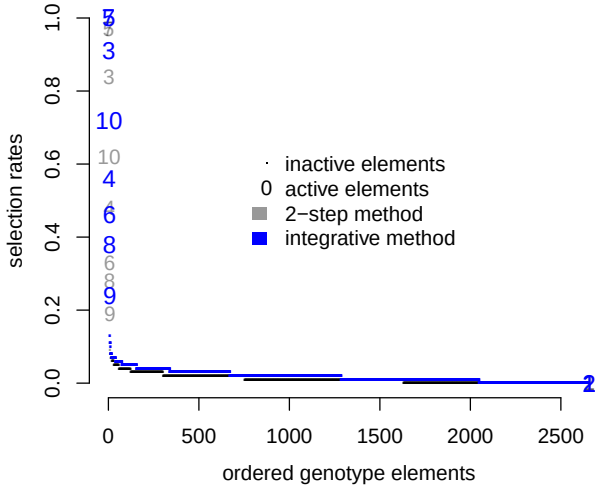

Figure 4: Selection rates of the genetic variables over 100 simulations. The 10 variables that should be selected are represented by their respective number (l).

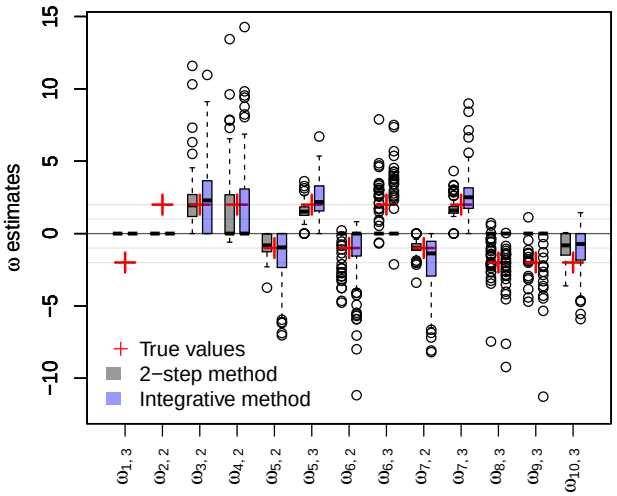

Figure 5: ω non-negative parameters estimates over 100 simulations and respective true values.

Model selection ability We conduct an additional simulation study to evaluate the capacity of the BIC (computed as described Section 2.1) 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. The 21 other times, a higher number of clusters is selected but never more than 5.

5 Application to Parkinson's Disease subtyping

We then apply the proposed method for 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 we apply our method 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 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. We chose to use two of them, ones we assume 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. We chose to use only 2652 of them: 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.

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 Figure 6. Remember that the variables represented here are residuals of a fitted linear model. On the

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

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.

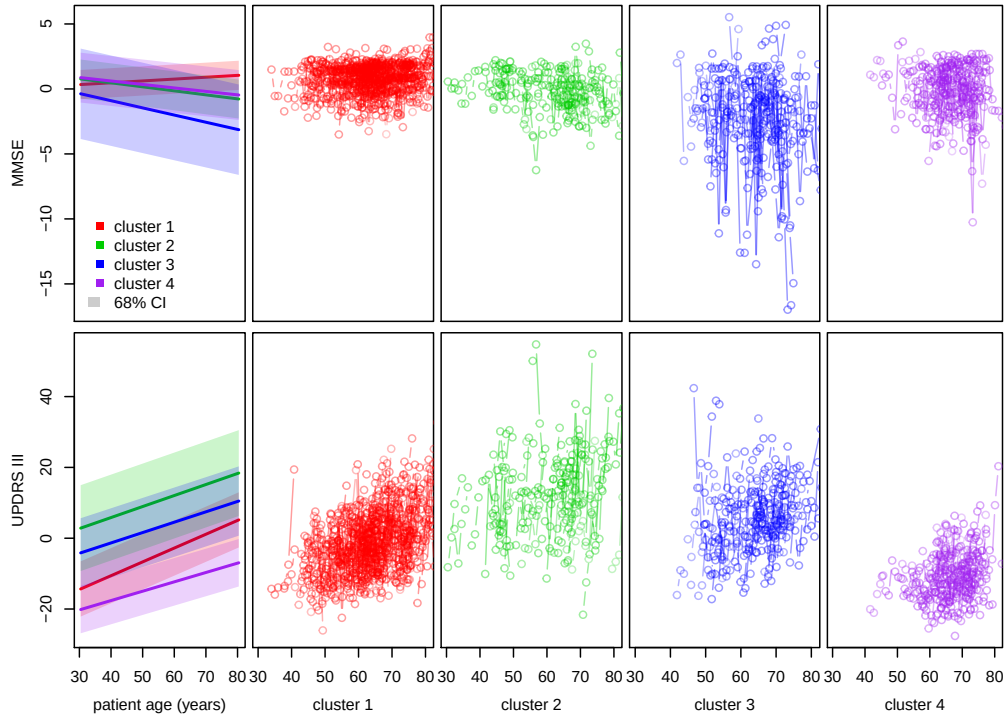

Figure 6: Clustering results with regard to the clinical variables.

Genetic association results Genetic association results are illustrated Figure 7. The 95% confidence intervals associated to the ω parameters are computed from the Hessian matrix (provided by the R `nnet` function of the `nnet` package (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.

15 SNPs were selected, 7 of them are part of genes with a potential role in neurological diseases. The SNP rs13284404 appears to be associated to cluster 1 (23/203 cases with at least one copy of the alternative allele, compared to 3/193 in the other clusters). It belongs to gene semaphorin 4D (SEMA4D), which is expressed in neurons and which expression is diminished for patients with Alzheimer's disease (Villa et al., 2010).

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 we show here with an association with cluster 2 only.

However, an unselected variant doesn't 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.

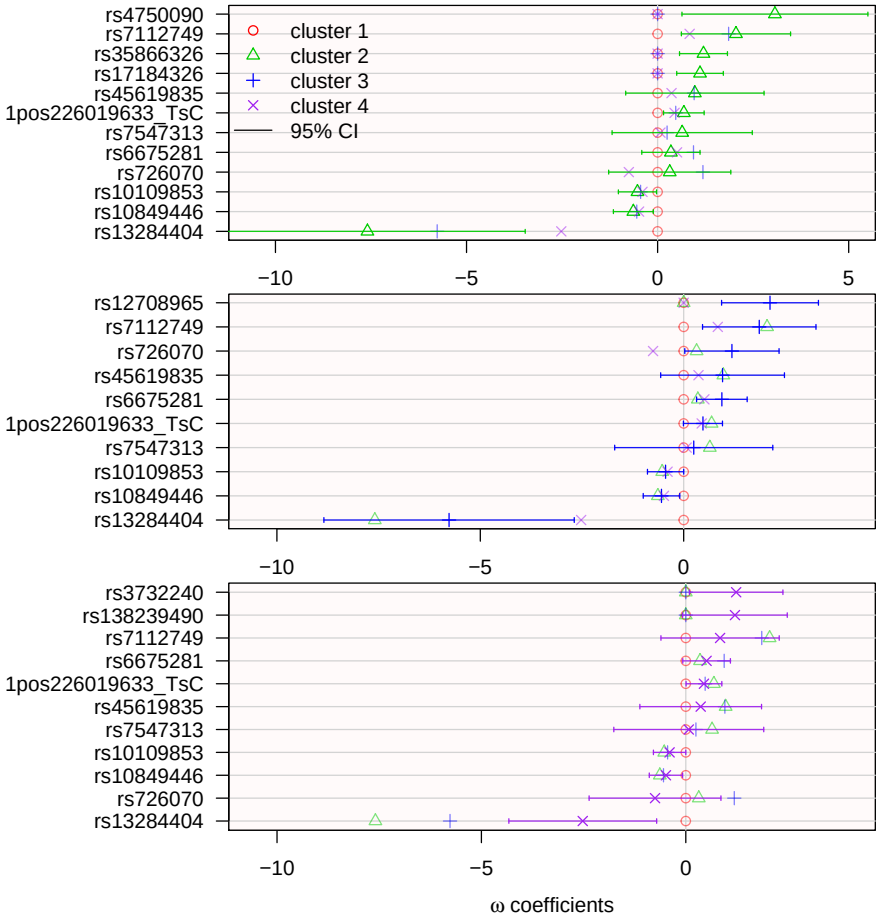

Figure 7: Genetic association results: estimated logistic regression (ω) parameters.

6 Conclusion

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 depending on them. A Lasso penalty on the logistic regression parameters permits to get parsimony and 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) permits to characterize 4 distinct subtypes and 15 genetic factors with a potential impact on the subtyping. Of these 15 SNPs, the one with the most significant role is already associated with PD. Half of the others belong to genes suspected to be implied in neurological diseases.

7 Discussion and perspectives

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 must be performed to asses 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). In short, 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 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).

8 Acknowledgements and funding

This communication concerns work carried out within the MeMoDeep (Methods and Models for Deep Screening of subphenotypes in Parkinson's Disease) project funded by the ANR. The methodological reflections of this communication 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

Abraham, C., Cornillon, P.-A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, 30(3):581–595.

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., Sabach, S., and Teboulle, M. (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., Artaud, F., Cormier-Dequaire, F., Rascol, O., Durif, F., Derkinderen, P., Marques, A.-R., Bourdain, F., Brandel, J.-P., Pico, F., Lacomblez, L., Bonnet, C., Brefel-Courbon, C., Ory-Magne, F., Grabli, D., Klebe, S., Mangone, G., You, H., Mesnage, V., Lee, P.-C., Brice, A., Vidailhet, M., Elbaz, A., and (2018). Longitudinal analysis of impulse control disorders in Parkinson disease. *Neurology*.

Delattre, M., Lavielle, M., and Poursat, M.-A. (2014). A note on BIC in mixed-effects models. *Electronic journal of statistics*, 8(1):456–475.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (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., Haugarvoll, K., Ross, O. A., Stone, J. T., Milkovic, N. M., Cobb, S. A., Whittle, A. J., Lincoln, S. J., Hulihan, M. M., Heckman, M. G., et al. (2006). Genomewide association, Parkinson disease, and PARK10. *American journal of human genetics*, 78(6):1084.

Fop, M. and Murphy, T. B. (2017). Variable selection methods for model-based clustering. *arXiv:1703.0229*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Genolini, C., Alacoque, X., Sentenac, M., Arnaud, 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., Williams-Gray, C., Foltynie, T., Compston, D., Barker, R., and Sawcer, S. (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., Szafranski, M., Ambroise, C., and Samson, F. (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., Chaudhary, K., and Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 8:84.

Ieva, F., Paganoni, A. M., Pigoli, D., and Vitelli, V. (2012). Multivariate functional clustering for the morphological analysis of ECG curves. *Journal of the Royal Statistal Society. Series C (Applied Statistics)*.

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

Jacques, J. and Preda, C. (2014b). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106.

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., Oesterreich, S., Kim, S., Park, Y., and Tseng, G. C. (2017). Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics*, 18(1):165–179.

Kristensen, V. N., Lingjœrde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. (2010). Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095.

Lewis, S., Foltynie, T., Blackwell, A., Robbins, T., Owen, A., and Barker, R. (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., Rowland, C., Schrodi, S., Laird, W., Tacey, K., Ross, D., Leong, D., Catanese, J., Sninsky, J., and Grupe, A. (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., De Andrade, M., Lesnick, T. G., Pant, P. K., Cox, D. R., and Ballinger, D. G. (2006). Response from Maraganore et al. *American journal of human genetics*, 78(6):1092.

Maraganore, D. M., De Andrade, M., Lesnick, T. G., Strain, K. J., Farrer, M. J., Rocca, W. A., Pant, P. K., Frazer, K. A., Cox, D. R., and Ballinger, D. G. (2005). High-resolution whole-genome association study of Parkinson disease. *The American Journal of Human Genetics*, 77(5):685–693.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (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.

Ripley, B., Venables, W., and Ripley, M. B. (2016). Package nnet. *R package version*, pages 7–3.

Rossi, F., Conan-Guez, B., and El Golli, A. (2004). Clustering functional data with the SOM algorithm. In *Proceedings of the 12th European Symposium on Artificial Neural Networks*, pages 305–312.

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., Olshen, A. B., and Ladanyi, M. (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., Olshen, A. B., and Ladanyi, M. (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., Wang, S., and Mo, Q. (2013). Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*, 7(1):269–294.

Sun, J., Bi, J., and Kranzler, H. R. (2014). Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC genetics*, 15(1):73.

Sun, J., Lu, J., Xu, T., and Bi, J. (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.

Villa, C., Venturelli, E., Fenoglio, C., De Riz, M., Scalabrini, D., Cortini, F., Serpente, M., Cantoni, C., Bresolin, N., Scarpini, E., et al. (2010). Candidate gene analysis of semaphorins in patients with Alzheimer's disease. *Neurological sciences*, 31(2):169–173.

Zhao, B., Kwok, J. T., and Zhang, C. (2009). Multiple kernel clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 638–649.