An adapted linear discriminant analysis for the classification in high-dimension, and an application to medical data

Khuyen T Le, Caroline Chaux, Frédéric Jp Richard, Eric Guedj

An adapted linear discriminant analysis for the classification in high-dimension, and an application to medical data.

Khuyen T. Le[a], Caroline Chaux[a], Frédéric J.P. Richard[a], Eric Guedj[b]

[a] *Aix Marseille Université, CNRS, Centrale Marseille, I2M, Marseille, France*
[b] *APHM, Hôpital de la Timone, Service de Biophysique et Médecine Nucléaire & Aix-Marseille Université, CNRS, Ecole Centrale Marseille, UMR 7249, Institut Fresnel, Marseille, France*

Abstract

In this paper, we deal with the issue of classifying normally distributed data in a high-dimensional setting when variables are more numerous than observations. Under a sparsity assumption on terms of the inverse covariance matrix (the precision matrix), we adapt the method of the linear discriminant analysis (LDA) by including a sparse estimate of the precision matrix over all populations. Furthermore, we develop a variable selection procedure based on the graph associated to the estimated precision matrix. For that, we define a discriminant capacity for each connected components of the graph, and keep variables of the most discriminant components. The adapted LDA and its selection procedure are both evaluated on synthetic data, and applied to real data from PET brain images for the classification of patients with Alzheimer's disease.

Keywords: Classification, linear discriminant analysis, Graphical LASSO, precision matrix estimation, variable selection, PET imaging, Alzheimer's disease.

1. Introduction

In this paper, we focus on supervised classification. This issue consists of identifying the category of an individual using a model whose parameters are learned from a pre-classified population. We tackle this issue in a high-dimensional setting when the number p of model parameters is larger than the number N of observations available to learn the model.

Assume that observations $X_j^{(k)}$ of a class k are sampled from a Gaussian distribution $\mathcal{N}(\mu^{(k)}, \Sigma^{(k)})$. Using a Bayes approach, a new individual x can be classified into the class k^* reaching the maximum of the posterior density function

$$k^* = \underset{k}{\operatorname{argmax}} \left(f_k(x)\pi_k\right) = \underset{k}{\operatorname{argmax}}(\log(f_k(x)\pi_k)), \tag{1}$$

Email addresses: lekhuyen.maths@gmail.com (Khuyen T. Le), caroline.chaux@univ-amu.fr (Caroline Chaux), frederic.richard@univ-amu.fr (Frédéric J.P. Richard), eric.guedj@ap-hm.fr (Eric Guedj)

Preprint

where $f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma^{(k)}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu^{(k)})(\Sigma^{(k)})^{-1}(x-\mu^{(k)})^T\right\}$ is the density function of a Gaussian vector and π_k is the probability to belong to the class k. This method is known as Quadratic Discriminant Analysis (QDA), or Linear Discriminant Analysis (LDA) when the covariance matrices $\Sigma^{(k)}$ are assumed to be the same for all classes. The estimation of inverse covariance matrices $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ (also called precision matrices) are required for these methods. When $p < N$, the estimation can be done by solving a maximum-likelihood problem

$$\hat{\Theta}^{(k)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \quad (-\log\det(\Theta) + \operatorname{trace}(S^{(k)}\Theta)), \tag{2}$$

where $\Theta \succ 0$ stands for a set of positive definite matrices and

$$S^{(k)} = \frac{1}{N_k}\sum_{j=1}^{N_k}(X_j^{(k)} - \mu_k)(X_j^{(k)} - \mu_k)^T$$

is the empirical covariance matrix of the class k. However, in a high-dimensional setting, this optimization problem is no longer well-posed and can not be used anymore to estimate the precision matrix. To fix this issue, Friedman proposed a regularized discriminant analysis [4] which uses a sparsity assumption for the covariance matrix estimation. This method was further improved in [7, 16]. In [2], Cai and Liu proposed another method which directly gives an estimate of the product between the precision matrix and the difference of mean vectors. This product is directly used to apply the decision rule of the LDA method.

In this paper, we propose to extend the use of the LDA in high dimension using an estimate of a common precision matrix obtained as a solution of the Graphical LASSO [19]:

$$\hat{\Theta}(\lambda) = \underset{\Theta \succ 0}{\operatorname{argmin}} \quad (-\log\det(\Theta) + \operatorname{trace}(S\Theta) + \lambda\|\Theta\|_1)\,, \tag{3}$$

where S is the empirical covariance matrix within groups

$$S = \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{N_k}(X_j^{(k)} - \mu^{(k)})(X_j^{(k)} - \mu^{(k)})^T, \tag{4}$$

$||\Theta||_1 = \sum_{m,n=1}^{p}|\Theta_{mn}|$ is a l_1-norm on Θ, and $\lambda > 0$. By adding this norm to the log-likelihood term, the precision matrix estimate $\hat{\Theta}(\lambda)$ becomes sparse. The sparsity level of $\hat{\Theta}(\lambda)$ depends on the value of λ: the larger λ is, the sparser is the solution.

Besides, it is well-known that the estimated precision matrix accounts for the conditional dependence of variables [15]: two variables X_i and X_j are dependent conditionally to the other variables if and only if $\hat{\Theta}_{ij}(\lambda) \neq 0$. From this information, it is possible to build a dependency graph known as the Graphical LASSO model: in this graph, two nodes i and j are connected if $\hat{\Theta}_{ij} \neq 0$. We can further extract the connected components of this graph and re-order variable indices so that the estimated precision matrix becomes block-diagonal.

The shrunk estimate of the precision matrix allows us to apply the LDA in a high-dimensional setting. By canceling some terms of the matrix, this method also reduces the complexity of the classification model. This can help improving the generalization

performances of the classification in high dimension. However, it might not be sufficient to deal with the generalization issue when the complexity is very large. So we propose an original selection method to further reduce the model complexity and improve the classification performance. We define a discriminant capacity for each block of variables in the estimated precision matrix. Then we rank blocks according to their discriminant capacity and remove variables of the least discriminant blocks.

We apply our method to data extracted from images acquired using positron emission tomography with [18F]-fluorodeoxyglucose (FDG-PET) [14]. Measuring the local glucose consumption, this imaging modality enables to observe the neural activity. It is used to study the so-called metabolic connectivity, which is defined as the coherence of the neural activity within the brain [8]. For instance, it was used to investigate cognitive functions [18, 20] and disease impairments [9]. In this paper, we focus on the classification of healthy control people and patients with Alzheimer's disease. The interest of our classification approach is to take into account an information about the metabolic connectivity through partial correlations encoded within precision matrix terms.

The rest of the paper is organized as follows. In Section 2, we present an algorithm for solving the GLASSO and estimating the precision matrix. In Section 3, we define the discriminant capacity of precision matrix blocks and describe our variable selection procedure. In Section 4, we evaluate our classification method on synthetic data. In Section 5, we present the application.

2. Estimation of a sparse precision matrix

2.1. Estimation of the precision matrix structure

In this section, we present an algorithm for choosing a value of the parameter λ^* for which the block-structure of the estimated precision matrix $\hat{\Theta}(\lambda^*)$ corresponds to the one of the target matrix Θ of the model. The algorithm relies upon a necessary and sufficient condition characterizing the structure of the estimated precision matrix [17]:

Theorem 1. *The solution $\hat{\Theta}(\lambda)$ of Eq.(3) is block diagonal with L blocks $B_1, \ldots, B_L$ if and only if $|S_{ij}| \leq \lambda$, for all $i \in B_l, j \in B_{l'}$, $l \neq l'$.*

This theorem enable the extraction from $\{|S_{ij}|, i \neq j\}$ of a finite ordered subset of critical parameter values $\Lambda = \{(\lambda_i)_{i=\overline{1,M}} : \lambda_1 > \lambda_2 > \cdots > \lambda_M\}$, called the GLASSO path, where the block structure of the precision matrix changes. Then, following [6], we can apply statistical tests to check at successive critical values λ_k the hypothesis $\mathcal{H}_k$: *"Each connected component of the graphical model $\hat{\mathcal{G}}(\lambda)$ derived from $\hat{\Theta}(\lambda)$ contains a connected component of the graphical model $\mathcal{G}$ associated to Θ, for all $\lambda < \lambda_k$"*. The set of hypotheses $\mathcal{H}_k$ has a nested structure: if $\mathcal{H}_k$ holds then $\mathcal{H}_{k'}$ is also true for any $k' > k$. According to [6], the hypothesis $\mathcal{H}_k$ can be tested using a statistic

$$T_k = N\lambda_k(\lambda_k - \lambda_{k+1}). \tag{5}$$

Assuming that k^* is the lowest index (i.e λ_{k^*} is the largest value) for which $\mathcal{H}_{k^*}$ is true, the probability distributions of these statistics tend to some exponential distributions

$$\begin{cases} T_{k^*} \xrightarrow{d} \text{Exp}(1), \\ T_{k'} \xrightarrow{d} \text{Exp}\left(\dfrac{1}{k' - k^* + 1}\right) & \text{for } k' > k^* \end{cases} \tag{6}$$

as $N, p \to \infty$ and $\dfrac{\log p}{N} \to 0$. In the test procedure, the hypothesis $\mathcal{H}_k$ is rejected when the statistic T_k is above a threshold τ. Using Eq. (5), the threshold is set to ensure that the risk of the first type error is below some $\alpha \in (0, 1)$. Starting from $k = 1$, we apply successively the test until the hypothesis $\mathcal{H}_k$ is not rejected.

Due to the asymptotic approximation of the distribution of T_k, the test may be inaccurate when the number of observations is not large enough. In this case, the risk of the second type error might be large. To compensate this, we can stop the iterative test procedure when the number of connected components in the graph of the estimated precision matrix is lower than a predefined bound $c_{\min}$.

The whole procedure is summarized in Algorithm 1.

2.2. Estimation of the precision matrix by GLASSO

Here, we assume that the precision matrix has a block diagonal form

$$\Theta = \begin{pmatrix} \Theta^{(1)} & 0 & \dots & 0 \\ 0 & \Theta^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Theta^{(L)} \end{pmatrix}, \tag{7}$$

This block structure may be estimated using the algorithm described in the previous section. We now present an algorithm to estimate terms of the matrix blocks knowing the block structure. Since the matrix is block diagonal, solving the GLASSO problem on the whole matrix reduces to solving separate GLASSO problems on each block

$$\hat{\Theta}^{(l)} = \underset{\Theta^{(l)} \succ 0}{\operatorname{argmin}} \quad \left[-\log\det(\Theta^{(l)}) + \operatorname{trace}(S^{(l)}\Theta^{(l)}) + \lambda \|\Theta^{(l)}\|_1 \right], \tag{8}$$

where $\Theta^{(l)}$ and $S^{(l)}$ are the sub-matrices extracted from the l^{th} block of Θ and S, respectively [17]. This problem separation is convenient in high dimension as it reduces the number of parameters to be estimated to $\sum_{l=1}^{L} p_l^2$, where p_l is the variable number on the l^{th} block.

There are many algorithms for solving the GLASSO among which the block-coordinate descent method [19, 5, 11] and the Alternating Direction Method of Multipliers (ADMM) [1]. In this paper, we use the ADMM algorithm which is both simple and efficient.

Introducing a matrix Z, problem (3) is equivalent to the minimization problem

$$(\hat{\Theta}, \hat{Z}) = \underset{\Theta = Z;\ \Theta, Z \succ 0}{\operatorname{argmin}} \left[F(\Theta) + \lambda ||Z||_1 \right], \tag{9}$$

where $F(\Theta) = -\log\det(\Theta) + \operatorname{trace}(S\Theta) + \lambda\|\Theta\|_1$. This problem can be solved by minimizing over (Θ, Z, U) the scaled augmented Lagrangian

$$L_\rho(\Theta, Z, U) = F(\Theta) + \lambda ||Z||_1 + \frac{\rho}{2}||\Theta - Z + U||_F^2 - \frac{\rho}{2}||U||_F^2, \tag{10}$$

where U is a dual variable and ρ serves as a penalty parameter. The ADMM is an iterative algorithm which alternates the minimization over variables Θ, Z and U [1]. It is described in Algorithm 2.

Algorithm 1 Estimation of the precision matrix structure.

Input: the sample covariance matrix S, the observation number N, the significance level α and the minimal number $c_{\min}$ of precision matrix blocks.
Output: an optimal value λ^* and the corresponding block structure of $\hat{\Theta}(\lambda^*)$.
The block structure of the estimated precision matrix $\hat{\Theta}(\lambda^*)$ associated to the optimal knot λ^*.
$\Gamma = \left\{(\gamma_l)_{l=\overline{1,Q}} : \gamma_1 \geq \cdots \geq \gamma_Q\right\}$ is obtained by sorting all the absolute values of off-diagonal elements of S in descending order.
$c_0 = p$ is the initial number of connected components (each node is a connected component).
Compute the threshold $\tau = F^{-1}(1-\alpha)$ where F is the cumulative distribution function of an exponential distribution of parameter 1.
$k = 1$, $\lambda_k = \gamma_1$.
for $l = 2, \ldots, Q$ **do**
    Find an adjacency matrix A representing all connected nodes where $A_{ij} = 1$ if $|S_{ij}| > \gamma_l$, and $A_{ij} = 0$ otherwise.
    Find the connected component number c of graph $\mathcal{G}$ associated to A.
    if $c \geq c_{\min}$ **then**
        if $c < c_0$ **then**
            Compute the statistic test T: $T_k = N\lambda_k(\lambda_k - \gamma_l)$.
            if $T_k > \tau$, (the hypothesis $\mathcal{H}_k$ is rejected) **then**
                $k := k + 1$;
                $\lambda_k = \gamma_l$;
                $c_0 = c$;
            else
                $\lambda^* = \gamma_l$;
                Stop iterations;
            end if
        end if
    else
        Stop iterations;
    end if
end for
From $\mathcal{G}$, infer the dependency structure of $\hat{\Theta}(\lambda^*)$. Reorder indices of the matrix $\hat{\Theta}(\lambda^*)$ so that it becomes block diagonal.

Algorithm 2 The ADMM algorithm to solve Problem (3).

Input: Sample covariance matrix S, parameters λ and ρ.
Output: Estimated precision matrix $\hat{\Theta}$.
Initialize: $\Theta^{(1)} = I, Z^{(1)} = I, U^{(1)} = 0$,
Select a scalar $\rho > 0$
for $n = 1, 2, 3, \ldots$ **do**
  $\Theta^{(n+1)} = \underset{\Theta \succ 0}{\operatorname{argmin}}\ L_\rho(\Theta, Z^{(n)}, U^{(n)})$
  $Z^{(n+1)} = \underset{Z \succ 0}{\operatorname{argmin}}\ L_\rho(\Theta^{(n+1)}, Z, U^{(n)})$
  $U^{(n+1)} = U^{(n)} + (\Theta^{(n+1)} - Z^{(n+1)})$
end for
Convergence condition: $||\Theta^{(n+1)} - \Theta^{(n)}|| \leq \varepsilon$

3. A method to select connected components

In this section, we design a method to select connected components that are the most discriminant. For that, we first define a criterion to assess the discriminant capacity of a component. This definition is inspired from the principles of the factorial discriminant analysis (FDA) which we recall next.

The FDA is a dimensionality reduction method which consists of finding a low dimensional projection of variables maximizing a dispersion criterion. In dimension $q \leq p$, variable projections are given as

$$Z_\Phi^{(k)} = \Phi^T X^{(k)}, \tag{11}$$

where Φ is in a set $\mathcal{M}_{p,q}$ of matrices of size $p \times q$ such that the covariance matrix of $Z_\Phi^{(k)}$ is equal to the identity matrix I_q. This condition means that components of $Z_\Phi^{(k)}$ are normalized and uncorrelated and is fulfilled if and only if $\Phi^T \Sigma \Phi = I_q$. Denoting

$$E^{(q)} = \{\Phi \in \mathcal{M}_{p,q}, \Phi^T \Sigma \Phi = I_q\}, \tag{12}$$

the FDA problem consists of finding $\Phi^{(q)}$ which maximizes over $E^{(q)}$ the function

$$\mathcal{J}(\Phi) = \sum_{k=1}^{K} \pi_k \mathbb{E}\left(Z_\Phi^{(k)} - \mathbb{E}(\sum_{j=1}^{K} \pi_j Z_\Phi^{(j)}) \right)^2, \tag{13}$$

where π_k is the probability for an observation to be in the k^{th} class. The function $\mathcal{J}$ represents a variance of the class means of projected variables. It can also be written as

$$\mathcal{J}(\Phi) = \Phi^T B \Phi, \tag{14}$$

where B is the inter-class covariance matrix defined by

$$B = \sum_{k=1}^{K} \pi_k (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T, \tag{15}$$

with $\mu = \sum_{k=1}^{K} \pi_k \mu^{(k)}$.
Let $\Phi^{(q)} = \underset{\Phi \in E^{(q)}}{\operatorname{argmax}} \mathcal{J}(\Phi)$. We define the discriminant capacity of a subspace of dimension q as

$$\Delta^{(q)} = \mathcal{J}(\Phi^{(q)}), \tag{16}$$

and the relative discriminant capacity as

$$\Delta_r^{(q)} = \frac{\Delta^{(q)}}{\Delta^{(p)}}\%. \tag{17}$$

The FDA consists of finding $\Phi^{(q)}$ for the lowest dimension q such that the $\Delta_r^{(q)}$ exceeds a predefined percent. This can be done in practice using the following proposition.

Proposition 1. Let $\phi_1, \cdots, \phi_p$ be eigenvectors of the matrix ΘB associated to its ordered eigenvalues $\psi_1 \geq \cdots \geq \psi_p$ such that $\phi_i^T \Sigma \phi_j = 1$ if $i = j$ and 0 otherwise. Then

$$\Phi^{(q)} = (\phi_1 | \cdots | \phi_q), \tag{18}$$

$$\Delta^{(q)} = \sum_{j=1}^{q} \psi_j, \quad \text{and} \quad \Delta_r^{(q)} = \frac{\sum_{j=1}^{q} \psi_j}{\operatorname{trace}(\Theta B)}. \tag{19}$$

Due to this proposition, the FDA algorithm reduces to the problem of finding the eigenvalues of ΘB.

We adapt the FDA to take into account the block structure of the matrix. Consider a block l indexed in a subset $I^{(l)}$ of $\overline{1, p}$ of size q_l. We now focus on projections in dimension q_l which only involve variables indexed in $I^{(l)}$. Let

$$\tilde{E}^{(l)} = \{\Phi \in \mathcal{M}_{p,q_l}, \Phi^T \Sigma \Phi = I_{q_l}, \Phi_{i,j} = 0, \forall i \notin I^{(l)}\}. \tag{20}$$

We define the discriminant capacity of block l as

$$\tilde{\Delta}^{(l)} = \max_{\Phi \in \tilde{E}^{(l)}} \mathcal{J}(\Phi), \tag{21}$$

and the relative one as

$$\tilde{\Delta}_r^{(l)} = \frac{\tilde{\Delta}^{(l)}}{\sum_{m=1}^{L} \tilde{\Delta}^{(m)}}. \tag{22}$$

To compute these discriminant capacities, we can use the following proposition.

Proposition 2. Assume that Θ is block diagonal. Let $\Theta^{(l)}$ and $B^{(l)}$ be sub-matrices extracted on the l^{th} block of Θ and B, respectively. Then

$$\tilde{\Delta}^{(l)} = \operatorname{trace}(\Theta^{(l)} B^{(l)}). \tag{23}$$

and

$$\tilde{\Delta}_r^{(l)} = \frac{\operatorname{trace}(\Theta^{(l)} B^{(l)})}{\operatorname{trace}(\Theta B)}. \tag{24}$$

In order to select blocks, we compute the relative discriminant capacity for each block, and rank them according to their capacity. We then select the L_0 most discriminant blocks where L_0 is chosen so that $\sum_{l=1}^{L_0} \tilde{\Delta}_r^{(l)}$ is above a predefined percent γ. For the classification, we eventually keep the variables on the selected blocks. In practice, the discriminant capacities are evaluated using estimates of Θ and B. The estimate of Θ and its block structures are obtained using Algorithms 1 and 2. The estimate of B is given by

$$\hat{B} = \sum_{j=1}^{K} \hat{\pi}_j (\hat{\mu}^{(j)} - \hat{\mu})(\hat{\mu}^{(j)} - \hat{\mu})^T, \tag{25}$$

where $\hat{\mu}^{(j)} = \sum_{i=1}^{N_j} X_i^{(j)}$ and $\hat{\mu} = \sum_{j=1}^{K} \pi_k \hat{\mu}^{(j)}$.

The block discriminant capacity may tend to foster large blocks even though they do not contain very discriminant variables. To attenuate this effect, we can also use a normalized discriminant capacity defined as

$$\tilde{C}_r^{(l)} = \frac{\tilde{\Delta}_r^{(l)}}{p_l} \tag{26}$$

The whole selection scheme is presented in Algorithm 3.

Algorithm 3 Connected component selection procedure

1: **Input:** the estimated precision matrix $\hat{\Theta}$, the estimated inter-class covariance matrix $\hat{B}$, the maximum ratio of discriminant capacity $0 < \gamma < 1$.
2: **Output:** The connected components whose discriminant capacities are the largest.
3: Compute the relative discriminant capacity of l^{th} block $\tilde{\Delta}_r^{(l)} = \dfrac{\text{trace}(\Theta^{(l)} B^{(l)})}{\text{trace}(\Theta B)}$, for $l = \overline{1, L}$.
4: Normalize these values by block size: $\tilde{C}_r^{(l)} = \dfrac{\tilde{\Delta}_r^{(l)}}{p_l}$, for $l = \overline{1, L}$.
5: Reorder these values by descending order $\overline{C}_r^{(1)} \geq \overline{C}_r^{(2)} \geq \cdots \geq \overline{C}_r^{(L)}$.
6: Choose the L_0 largest connected component with the largest L_0 satisfying $\sum_{l=1}^{L_0} \overline{C}_r^{(l)} \leq \gamma$.

4. Numerical study

We evaluated the A-LDA and the connected component selection method on synthetic data. We repeated 500 experiments. For each experiment, we generated a data set composed of two normally distributed populations of size $N = 300$: $X^{(k)} = \{X_1^{(k)}, X_2^{(k)}, \ldots, X_N^{(k)}\}$ where $X_i^{(k)} \sim \mathcal{N}(\mu_k, \Sigma)$ for $i = \overline{1, N}$ and $k = 1, 2$. The observation vectors $X_i^{(k)}$ had values in $\mathbb{R}^p$ where p was chosen in $\{150, 300, 500\}$. The covariance matrix Σ had block

diagonal form:

$$\Sigma = \begin{pmatrix} \Sigma^{(1)} & 0 & \dots & 0 \\ 0 & \Sigma^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma^{(L)} \end{pmatrix}. \tag{27}$$

For each experiment, this matrix was randomly generated as follows. Its number of blocks L was set to $\frac{p}{10}$. Setting $p_0 = 0$, the size p_l of each block l was chosen randomly as follows.

$$p_l \sim \mathcal{U}\left([\![1, p - \sum_{i=1}^{l-1} p_i - (L-l)]\!] \right), \forall 1 \leq l \leq L. \tag{28}$$

The mean vector μ_1 of the first population was sampled from a multivariate normal distribution $\mu_1 \sim \mathcal{N}(0, I)$. The one of the second population was set so as to have only $L_0 = \left\lfloor \frac{25}{100} L \right\rfloor$ discriminant blocks: $\mu_2 = \mu_1 + \delta$ where $\delta \in \mathbb{R}^p$ is given by:

$$\delta_i = \begin{cases} 0.3(1 + 0.9\omega) & \text{if } i \in V_l, l = \overline{1, L_0}, \\ 0 & \text{otherwise,} \end{cases} \tag{29}$$

where V_l is the index set of variables of the l^{th} block and $\omega \sim \mathcal{U}([0, 1])$.

In order to evaluate the different parts of the method, we applied it in three different situations of increasing complexity:

1. The terms and the block structure of covariance matrix Σ and its inverse Θ are completely known. They are not estimated.

2. Only the block structure of matrices Σ and Θ is known. The terms of the precision matrix are estimated using Algorithm 2.

3. The precision matrix is completely unknown, and fully estimated using both Algorithms 1 and 2.

The method was evaluated with and without connected component selection. The selection algorithm 3 was applied with $\gamma = 0.8$. In each situation, the classification model was trained on the subset containing 200 observations from each population. A classification error was computed on the 100 remaining observations of each population. We compared our method to other state-of-the-art methods implemented in the Statistic and Machine learning Matlab Toolbox: Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN) and Ensemble classification (ENS). We also compared it to a Naive-Bayes classifier (NB) consisting of a LDA decision rule with a pseudo-inverse of the sample covariance matrix S as an estimation of Σ.

p	C	A-LDA		NB		SVM		DT		ENS	
		A	B	A	B	A	B	A	B	A	B
150	1	7.33	6.31	12.21	7.21	13.51	8.99	36.86	34.11	17.73	15.02
	2	7.62	6.97	12.21	7.44	13.51	8.93	36.86	33.81	17.73	15.24
	3	8.13	7.78	12.21	8.76	13.51	11.16	36.86	34.99	17.73	16.33
300	1	12.57	9.80	28.02	12.28	21.19	15.37	43.85	41.41	28.03	23.22
	2	12.82	10.59	28.02	13.21	21.19	16.73	43.85	41.68	28.03	23.92
	3	13.91	13.05	28.02	16.13	21.19	18.29	43.85	42.49	28.03	25.55
500	1	19.67	14.59	37.92	19.26	29.35	22.68	46.78	44.93	36.97	30.83
	2	20.00	17.70	37.92	27.18	29.35	25.67	46.78	46.19	36.97	34.47
	3	21.71	19.84	37.92	29.78	29.35	26.38	46.78	46.15	36.97	34.69

Table 1: Classification error (in %) of all methods applied without (A) and with (B) variable selection. C refers to the experimental situation.

Variable number p	150	300	500
Estimation of the connected components (Algorithm 1)	6.58	70.10	519.78
Estimation of the precision matrix (Algorithm 2)	7.82	70.62	254.80
Selection of connected components (Algorithm 3)	0.002	0.009	0.02

Table 2: Computational times (in seconds) for the complete learning of the classification model A-LDA for different variable number p.

Classification errors are reported in Table 1 for all methods. Results of the A-LDA were significantly better than the ones of the NB, showing the importance of taking into account the precision matrix structure in the decision rule. Moreover, comparing A-LDA results in different situations, we observe that the estimation of the precision matrix had only a slight effect on the classification performances. In the case when $p = 150$, the classification error obtained in the situation $C = 2$ (resp. $C = 3$) where the precision matrix is partly (resp. fully) estimated was only 0.3% (resp. 0.8%) higher than the one in the situation $C = 1$ where it is known. We had similar results in the high dimension setting when $p = 300$ or $p = 500$.

Besides, the connected component selection method improved the performance of A-LDA. It slightly reduced the classification error by 0.7% in the setting when $p = 150$. The error was more importantly reduced in high dimensional settings. The selection decreased by 0.95% (resp. 3.08%) when $p = 300$ (resp. $p = 500$).

A-LDA outperformed all the other classification methods. Among the other methods, SVM achieved the lowest error (13.51% for $p = 150$), which is much higher than the one of A-LDA. Let us outline that other methods do not use any information from the precision matrix. Hence, their comparison to A-LDA suggests that information from the precision matrix are critical for classification. Besides, performances of the other methods were significantly improved using the component selection method. For instance, the error of SVM was reduced about 2% to 3% when selecting components.

Computational time for learning the classification model A-LDA is presented in Table 2 for different variable numbers p. The estimation of connected components was the most time-consuming part of the method. Its computation time particularly increased as p got larger.

5. Application to medical data

In this section, we aim at discriminating patients with the Alzheimer's Disease (AD) from Healthy Control (HC) people and investigate impairments of the neural coherence due to Alzheimer's disease.

In collaboration with La Timone Hospital (Marseille, France), a PET image was acquired for each individual of a cohort composed of 38 patients with AD and 56 HC. Each image was automatically segmented into 116 predefined anatomical regions. Then the mean image intensity was computed on each region. For an individual i of the class k (1 for HC and 2 for AD), we formed an observation vector $X_i^{(k)}$ composed of the mean intensity on the brain regions.

We estimated the connected components of the precision matrix using Algorithm 1 with $\alpha = 0.05$, and different values of $c_{\min} \in \{1, 5, 7, 10, 15, 20, 30\}$, the parameter monitoring the minimal number of components. Then the precision matrix was estimated using Algorithm 2, with a value of λ corresponding to the critical value of the GLASSO path corresponding to the number of connected components. Eventually, we applied Algorithm 3 for selecting connected components with different threshold values γ. We evaluated classification errors by cross-validation. Errors are given in Table 3. When

L \ $\gamma(\%)$	< 70	$[70, 80)$	$[80, 90)$	$[90, 100)$	100
1	-	-	-	-	3.52 (116)
5	13.72 (2)	-	13.25 (4)	3.52 (114)	3.52 (116)
7	13.72 (2)	13.25 (3)	4.84 (111)	3.10 (113)	4.42 (116)
10	13.72 (2)	3.52 (101)	-	4.42 (103)	4.84 (116)
15	12.83 (5)	15.04 (6)	2.21 (99)	3.52 (109)	6.63 (116)
20	12.83 (5)	11.89 (40)	4.84 (97)	7.05 (108)	9.26 (116)
30	4.84 (31)	6.63 (32)	8.79 (60)	9.68 (74)	10.57 (116)
116	9.68 (20)	11.00 (29)	11.47 (38)	11.04 (58)	12.36 (116)

Table 3: Classification errors (in %) of the A-LDA applied to the classification of AD versus HC. L is the number of estimated connected component. The parameter γ is the threshold used for selecting these components. The values in parentheses give the number of selected variables in each case.

applied with $L = 116$, A-LDA does not take into account signal correlations between regions. The method then corresponds to a reference method used in clinical routines. This method had 12.36% of classification error. At the opposite, when $L = 1$, the method takes into account all correlations. In this case, the error was only 3.52%, which highlights the importance of correlations for the classification. In this case, all regions belongs to a unique connected component. Splitting regions into several connected components and

selecting the most discriminant ones, we could further reduce the error. For $L = 15$ and $\gamma \in [80, 90)$, we obtained only 2.21% of error. This improvement is due to model simplifications which result from removing non-discriminant components and neglecting correlations between the discriminant ones. On this data, our method outperformed SVM whose errors were 6.38% and 4.26% when applied without and with our selection method, respectively.

In Table 4, we detail the composition of the connected components found by our method in the best classification case ($L = 15$ and $\gamma \in [80, 90)$). There was a large discriminant connected component. Among the most discriminant brain regions identified, the posterior cingulate cortex is known to be the first metabolically involved in AD [12]. By contrast, non-discriminant regions are indeed supposed to be only involved in the last stages of the disease (cerebellum, thalamus, Rolandic region, putamen, pallidum). Their presence in a discriminant connected component is explained by their interactions with impaired regions. The non-discriminant connected components gathered regions which are known to be little affected by the disease.

CC	NR	Discriminant brain regions
1	2	Cingulum-Post (L,R)
2	1	Caudate (L)
3	2	Cingulum-Mid (L,R)
4	1	Heschl (R)
5	92	Precentral (L,R), Frontal-Sup (L,R), Frontal-Sup-Orb (L,R), Frontal-Mid (L,R), Frontal-Mid-Orb (L,R), Frontal-Inf-Oper (L,R), Frontal-Inf-Tri (L,R), Frontal-Inf-Orb (L), Supp-Motor-Area (L,R), Olfactory (L,R), Frontal-Sup-Medial (L,R), Frontal-Med-Orb (L,R), Rectus (L,R), Cingulum-Ant (L,R), Hippocampus (L,R), ParaHippocampal (L,R), Amygdala (L,R), Calcarine (L,R), Cuneus (L,R), Lingual (L,R), Occipital-Sup (L,R), Occipital-Mid (L,R), Occipital-Inf (L,R), Fusiform (R), Postcentral (L,R), Parietal-Sup (L,R), Parietal-Inf (L,R), SupraMarginal (L,R), Angular (L,R), Precuneus (L,R), Paracentral-Lobule (L,R), Temporal-Sup (R), Temporal-Pole-Sup (L,R), Temporal-Mid (L,R), Temporal-Pole-Mid (L,R), Temporal-Inf (R), Cerebelum-Crus1 (L,R), Cerebelum-Crus2 (L,R), Cerebelum-3 (L,R), Cerebelum-4-5 (L,R), Cerebelum-6 (L,R), Cerebelum-7b (L,R), Cerebelum-8 (L,R), Cerebelum-9 (L,R), Vermis-1-2, Vermis-3, Vermis-4-5, Vermis-6, Vermis-7, Vermis-8, Vermis-9, Vermis-10
6	1	Caudate (R)
CC	NR	Non-discriminant brain regions.
7	1	Cerebelum-10-L
8	1	Thalamus (L)
9	1	Temporal-Inf (L)
10	7	Rolandic-Oper (L), Insula (L), Putamen (L,R), Pallidum (L), Heschl (L), Temporal-Sup (L)
11	3	Frontal-Inf-Orb (R) Rolandic-Oper (R) Insula (R)
12	1	Thalamus (R)
13	1	Fusiform (L)
14	1	Cerebelum-10 (R)
15	1	Pallidum (R)

Table 4: Estimated connected components of brain regions, CC and NR stand for the connected component and the number of regions respectively.

6. Discussion

So as to classify normally distributed data in a high-dimensional setting, we proposed to adapt the linear discriminant analysis (LDA). Under the assumption that the precision matrix of the model was sparse, we included in the LDA decision rule a sparse estimate of the precision matrix obtained as a solution of the GLASSO problem. We further developed a variable selection procedure based on the graph associated to the estimated precision matrix. This procedure relied upon the definition of a discriminant capacity of a connected component of the graph. It consisted of keeping variables of the most

discriminant components.

The assumption that the K groups share the same precision matrix helps to limit the number of parameters to be estimated by GLASSO problem. That further improves the classification performance. However, it would be interesting to have different precision matrices for the K groups which would lead to the use of a quadratic discriminant analysis for the classification. In a future work, we plan to adapt the quadratic discriminant analysis to our context.

The connected component selection method helps to reduce the classification error. Nevertheless, the number of selected variables is still large. This can be explained by the fact that there exist some large connected components with high discriminant capacities. However, some variables of these components may not be so important for classification purposes. Hence, one can think about further selecting the most discriminating variables in each connected component in order to reduce as much as possible the number of useless variables.

The adapted LDA was applied to data extracted from PET images for discriminating patients with Alzheimer Disease (AD) from Healthy Control (HC) people. In [9], authors used a quadratic discriminant analysis with sparse estimates of class precision matrices to classify a set composed of 49 AD and 67 HC. From PET data, they manually selected 42 anatomical regions within four lobes (Frontal, Parietal, Occipital and Temporal). They reported a sensitivity of 88% and a specificity of 88%, which are both lower than those we obtained (sensitivity of 97.37% and specificity of 98.21% when 99 brain regions are selected). Our study further showed that brain regions in other lobes (Limbic and Cerebellum) could play an important role for classifying these groups leading up to 10 % improvement.

Besides, some studies have been conducted on other functional image data. They all reported classification errors higher than ours. In [13], authors classified 114 AD and 114 HC using EEG data of size 132. Applying a regularized linear discriminant analysis, they obtained 26% of mis-classification. In [3], a pattern classification approach was used to classify 200 AD and 200 HC based on MRI data. This method obtained 5.7% of mis-classifications. In [10], authors made four different trials on MRI dataset. Results varied between 5% and 19% of errors.

References

[1] S. Boyd, N. Parikh, E. Chu, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[2] T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

[3] Y. Fan, N. Batmanghelich, C. Clark, C. Davatzikos, et al. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, 39(4):1731–1743, 2008.

[4] J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, 9(3):432–441, 2008.

[6] M. G'Sell, J. Taylor, and R. Tibshirani. Adaptive testing for the graphical LASSO. *arXiv preprint arXiv:1307.4765*, 2013.

[7] J. Hoffbeck and D. Landgrebe. Covariance matrix estimation and classification with limited training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):763–767, 1996.

[8] B. Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19(2):466–470, 2003.
[9] S. Huang, J. Li, L. Sun, et al. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.
[10] S. Klöppel, C. Stonnington, C. Chu, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008.
[11] R. Mazumder and T. Hastie. The graphical LASSO: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125, 2012.
[12] L. Mosconi. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. FDG-PET studies in MCI and AD. *European Journal Nuclear Medicine and Molecular Imaging*, 32(4):486–510, 2005.
[13] E. Neto, F. Biessmann, H. Aurlien, et al. Regularized linear discriminant analysis of EEG features in dementia patients. *Frontiers in Aging Neuroscience*, 8:273, 2016.
[14] M. Phelps, H. Schelbert, and J. Mazziotta. Positron computed tomography for studies of myocardial and cerebral function. *Annals of Internal Medicine*, 98(3):339–359, 1983.
[15] J. Wang and M. Kolar. Inference for sparse conditional precision matrices. *arXiv preprint arXiv:1412.7638*, 2014.
[16] D. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71 (3):615–636, 2009.
[17] D. Witten, J. Friedman, and N. Simon. New insights and faster computations for the graphical LASSO. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.
[18] I. Yakushev, G. Chetelat, F. Fischer, et al. Connectivity within the default mode network relates to working memory performance in young healthy subjects. *Journal of Nuclear Medicine*, 53 (supplement 1):304–304, 2012.
[19] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, pages 19–35, 2007.
[20] N. Zou, G. Chetelat, M. Baydogan, et al. Metabolic connectivity as index of verbal working memory. *Journal of Cerebral Blood Flow & Metabolism*, 35(7):1122–1126, 2015.

Appendix A. Proof of Proposition 1

Proof. The FDA aims to find $(K-1)$ axes where the projection of data on these axes maximizes the variance between classes and minimize the variance within classes. These axes satisfy the following problem:

$$\phi^* = \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \quad \phi^T B \phi, \text{ subject to } \phi^T \Sigma \phi = 1. \tag{A.1}$$

It is equivalent to:

$$\phi^* = \underset{\phi \in \mathbb{R}^p}{\operatorname{argmax}} \quad \left[\phi^T B \phi - \psi\left(\phi^T \Sigma \phi - 1\right)\right]. \tag{A.2}$$

We obtain:

$$\Theta B \phi^* = \psi^* \phi^* \tag{A.3}$$

where $\Theta = \Sigma^{-1}$. Hence, ϕ^* is an eigenvector of ΘB corresponding to the eigenvalue λ^*. The most important axis is the eigenvector ϕ_1 corresponding to the largest eigenvalue ψ_1 of ΘB. The second important axis (follows ϕ_1) for discriminating the data is the eigenvector corresponding to the second largest eigenvalue ψ_2 of ΘB ($\psi_2 < \psi_1$). Therefore, we can evaluate the discriminant capacity of all eigenvectors of ΘB through their corresponding eigenvalues. Hence, the matrix $\Phi^{(q)}$ contains q eigenvectors corresponding to the q biggest eigenvalues of ΘB, $\Phi^{(q)} = (\phi_1 | \ldots | \phi_q)$. Following Equation (16), the

discriminant capacity of the subspace generated by these q vectors is given by:

$$\begin{aligned}\Delta^{(q)} &= \text{trace}((\Phi^{(q)})^T B\Phi^{(q)}) = \text{trace}((\Phi^{(q)})^T \Sigma\Theta B\Phi^{(q)}) = \text{trace}((\Phi^{(q)})^T \Sigma\Phi^{(q)}) \\ &= \text{trace}((\Phi^{(q)})^T \Sigma\Phi^{(q)}(\Phi^{(q)})^T \Theta B\Phi^{(q)}) = \text{trace}((\Phi^{(q)})^T \Theta B\Phi^{(q)}) \\ &= \sum_{i=1}^{q} \psi_i\end{aligned}$$

If $q = p$ then $\Phi^{(p)} = (\phi_1|\cdots|\phi_p)$ contains all eigenvectors of ΘB corresponding to eigenvalues $\psi_1 \geq \cdots \geq \psi_p$. Therefore, the discriminant capacity of $\Phi^{(p)}$ is given by:

$$\Delta^{(p)} = \text{trace}((\Phi^{(p)})^T \Theta B\Phi^{(p)}) = \sum_{i=1}^{p} \psi_i = \text{trace}(\Theta B).$$

Hence, the relative discriminant capacity of $\Phi^{(q)}$ is rewritten as:

$$\Delta_r^{(q)} = \frac{\sum_{i=1}^{q} \psi_i}{\text{trace}(\Theta B)} = \frac{\sum_{i=1}^{q} \psi_i}{\sum_{i=1}^{p} \psi_i}. \tag{A.4}$$

□

Appendix B. Proof of Proposition 2

Proof. Following Equation (21), the discriminant capacity of l^{th}-block is computed by:

$$\begin{aligned}\tilde{\Delta}^{(l)} &= \max_{\Phi\in\tilde{E}^{(l)}} \mathcal{J}(\Phi) = \max_{\Phi\in\tilde{E}^{(l)}} \text{trace}(\Phi^T B\Phi) \\ &= \max_{\Phi\in\tilde{E}^{(l)}} \sum_{k=\overline{1,q_l},i,m=\overline{1,p}} \Phi_{ik} B_{km} \Phi_{mi} \\ &= \max_{\Phi\in\tilde{E}^{(l)}} \sum_{i,k,m=\in I^{(l)}} \Phi_{ik} B_{km} \Phi_{mi} \\ &= \max_{\tilde{\Phi}^{(l)}} \text{trace}((\tilde{\Phi}^{(l)})^T B^{(l)} \tilde{\Phi}^{(l)}) \quad \text{where } \tilde{\Phi}^{(l)} = (\Phi_{ij})_{i,j\in I^{(l)}} \\ &= \max_{\tilde{\Phi}^{(l)}} \text{trace}((\tilde{\Phi}^{(l)})^T \Sigma^{(l)} \tilde{\Phi}^{(l)} (\tilde{\Phi}^{(l)})^T \Theta^{(l)} B^{(l)} \tilde{\Phi}^{(l)}) \\ &= \max_{\tilde{\Phi}^{(l)}} \text{trace}((\tilde{\Phi}^{(l)})^T \Theta^{(l)} B^{(l)} \tilde{\Phi}^{(l)})\end{aligned}$$

Let denote $\hat{\Phi}^{(l)} = (\phi_1^{(l)}|\ldots,|\phi_{p_l}^{(l)}) \in \mathcal{M}_{q_l,q_l}$ contains q_l eigenvectors $\Theta^{(l)} B^{(l)}$ corresponding to q_l eigenvalues $\psi_1^{(l)} \geq \cdots \geq \psi_{p_l}^{(l)}$. Then

$$\tilde{\Delta}^{(l)} = \text{trace}((\hat{\Phi}^{(l)})^T \Theta^{(l)} B^{(l)} \hat{\Phi}^{(l)}) = \text{trace}(\Theta^{(l)} B^{(l)}) \tag{B.1}$$

The relative discriminant capacity of the l^{th}-block is given by

$$\tilde{\Delta}_r^{(l)} = \frac{\tilde{\Delta}^{(l)}}{\sum_{l=1}^{L} \tilde{\Delta}^{(l)}} = \frac{\text{trace}(\Theta^{(l)} B^{(l)})}{\sum_{l=1}^{L} \text{trace}(\Theta^{(l)} B^{(l)})} = \frac{\text{trace}(\Theta^{(l)} B^{(l)})}{\text{trace}(\Theta B)} \tag{B.2}$$

□