A greedy great approach to learn with complementary structured datasets

Christophe Ambroise, Julien Chiquet, Marie Szafranski

A greedy great approach to learn with complementary structured datasets

Christophe Ambroise CHRISTOPHE.AMBROISE@GENEPOLE.CNRS.FR
UÉVE & LaMME, UMR CNRS, USC INRA

Julien Chiquet JULIEN.CHIQUET@GENEPOLE.CNRS.FR
INRA & MMIP, UMR AgroParisTech – UÉVE & LaMME, UMR CNRS, USC INRA

Marie Szafranski MARIE.SZAFRANSKI@MATH.CNRS.FR
ENSIIE & LaMME, UMR CNRS, USC INRA

1. Introduction

We are interested in problems where two hierarchically structured datasets may be complementary for a learning process. This case may arise in biological applications where genomic and metagenomic analyses may be collected for studying the genomic features of an organism along with its environnement.

In this work, we propose a model to assess the relevant interactions between the two datasets. We use a compressed representation of the original data to cope with the high dimensionality of the problem. We show that the collection of models, characterized through the hierarchical structures, forms a partially ordered set and take advantage of this organization to define a greedy approach to solve the problem more efficiently. Finally, we illustrate the behavior of the resulting algorithm on numerical simulations.

2. Learning with complementary datasets

The proposed model relies on observations stemming from two different views, G and M, which are gathered into a training set $\mathcal{S} = \{(\mathbf{x}_i^G, \mathbf{x}_i^M, y_i)\}_{i=1}^N$, where $(\mathbf{x}_i^G, \mathbf{x}_i^M, y_i) \in \mathbb{R}^{D_G} \times \mathbb{R}^{D_M} \times \mathbb{R}$. We consider that the variables of each view are organized into a tree structure.

2.1. Notations

We use the following notations and conventions. Bold lowercase letters will usually denote vectors and bold capital letters will usually denote matrices. The symbol $\cdot$ is used for the classical product, $\times$ for the cartesian product, $\otimes$ for the Kronecker product and $|S|$ for the cardinal of a set S.

The tree structure over G is of depth H_G and is defined at height k, with $0 \leq k \leq H_G$, by the N_G^k groups of variables $\mathcal{G}^k = \{\mathcal{G}_g^k\}_{g=1}^{N_G^k}$. We denote $\mathbf{x}_i^g \in \mathbb{R}^{D_g^k}$, the sample i restricted to the variables of G from group $g \in \mathcal{G}^k$.

Similarly, the tree structure over M is defined at height ℓ, with $0 \leq \ell \leq H_M$, by the N_M^ℓ groups of variables $\mathcal{M}^\ell = \{\mathcal{M}_m^\ell\}_{m=1}^{N_M^\ell}$ and $\mathbf{x}_i^m \in \mathbb{R}^{D_m^\ell}$ denotes the sample i restricted to the variables of M from group $m \in \mathcal{M}^\ell$.

We also introduce $N_I^{k\ell} = N_G^k \cdot N_M^\ell$, the number of possible interactions and $D_I = D_G \cdot D_M$, the number of variables that interact, which remains the same whatever k and ℓ.

2.2. Full representation

For a given height k of G and a given height ℓ of M, we define the model

$$y_i = \sum_{g \in \mathcal{G}^k} \sum_{m \in \mathcal{M}^\ell} \mathbf{x}_i^g \mathbf{\Theta}_{gm}^{k\ell} (\mathbf{x}_i^m)^T + \epsilon_i \,, \qquad (1)$$

where the matrix $\mathbf{\Theta}_{gm}^{k\ell} \in \mathbb{R}^{D_g^k \times D_m^\ell}$ contains the coefficients of interactions between the variables of groups g and m at heights k and ℓ and $\epsilon_i \in \mathbb{R}$ is a residual error. [1]

The dimension D_I involved in Problem (1) to estimate $\mathbf{\Theta}_{gm}^{k\ell}$, for all $g \in \mathcal{G}^k$ and $m \in \mathcal{M}^\ell$, may be huge especially for applications with an important number of variables such as in biology with genomic and metagenomic data.

To reduce to $N_I^{k\ell} \ll D_I$ the dimension of the optimisation problem, we might compress the data at given levels of the hierarchies by averaging the variables of each group or using Hierarchical Multiple Factor Analysis (HMFA, Escofier and Pagès, 2008; Lê et al., 2008).

Aside from the benefit of reducing the dimension of Problem (1), we will show in Section 3 that such a compact representation will be convenient to explore the hierarchies.

[1] Note that in the absence of a regularization term to enforce the hierarchical structure on $\mathbf{\Theta}_{gm}^{k\ell}$, all the models are equivalents whatever k or ℓ here.

2.3. Compact representations

Averaged variables on groups. A first simple manner to compress the data is to consider averaged variables on groups at a given level. For view G and group $g \in \mathcal{G}^k$, and for view M and group $m \in \mathcal{M}^\ell$, we can define $\tilde{\mathbf{x}}^g \in \mathbb{R}^N$ and $\tilde{\mathbf{x}}^m \in \mathbb{R}^N$ as

$$\tilde{\mathbf{x}}^g = \frac{1}{D_g^k} \sum_{j \in \mathcal{G}_g^k} \mathbf{x}^j \,, \qquad \mathbf{x}^j \subseteq \mathbf{x}^G \,, \qquad \mathbf{x}^j \in \mathbb{R}^N \,,$$

$$\tilde{\mathbf{x}}^m = \frac{1}{D_m^\ell} \sum_{j \in \mathcal{M}_m^\ell} \mathbf{x}^j \,, \qquad \mathbf{x}^j \subseteq \mathbf{x}^M \,, \qquad \mathbf{x}^j \in \mathbb{R}^N \,.$$

Hierarchical Multiple Factor Analysis. A more sophisticated compression relies on HMFA which can be seen as a reweighted Principal Component Analysis (PCA) on predefined groups of variables defined into a hierarchy. This compression will be detailed in a longer version.

2.4. Model

With a compact representation, Problem (1) turns to a classical linear model

$$y_i = \sum_{g \in \mathcal{G}^k} \sum_{m \in \mathcal{M}^\ell} \underbrace{(\tilde{x}_i^g \cdot \tilde{x}_i^m)}_{(\varphi_{gm}^{k\ell})_i} \theta_{gm}^{k\ell} + \epsilon_i \,, \tag{2}$$

with $\theta_{gm}^{k\ell} \in \mathbb{R}$, which can be reformulated in a matrix form.

We denote by $\boldsymbol{\phi}_{gm}^{k\ell} \in \mathbb{R}^N$ the vector whose ith row is given in Equation (2) by $(\varphi_{gm}^{k\ell})_i$. We also define the matrix $\boldsymbol{\Phi}^{k\ell} \in \mathbb{R}^{N \times N_I^{k\ell}}$ by

$$\boldsymbol{\Phi}^{k\ell} = \left(\boldsymbol{\phi}_{11}^{k\ell} \cdots \boldsymbol{\phi}_{1N_M^\ell}^{k\ell} \cdots \boldsymbol{\phi}_{gm}^{k\ell} \cdots \boldsymbol{\phi}_{N_G^k 1}^{k\ell} \cdots \boldsymbol{\phi}_{N_G^k N_M^\ell}^{k\ell} \right) \,,$$

and the vector $\boldsymbol{\theta}^{k\ell} \in \mathbb{R}^{N_I^{k\ell}}$ by

$$\boldsymbol{\theta}^{k\ell} = \left(\theta_{11}^{k\ell} \cdots \theta_{1N_M^\ell}^{k\ell} \cdots \theta_{gm}^{k\ell} \cdots \theta_{N_G^k 1}^{k\ell} \cdots \theta_{N_G^k N_M^\ell}^{k\ell} \right)^T \,.$$

We may finally write Problem (2) for heights k and ℓ as a linear regression problem which reads

$$\mathbf{y} = \boldsymbol{\Phi}^{k\ell} \boldsymbol{\theta}^{k\ell} + \boldsymbol{\epsilon} \,, \tag{3}$$

where $\mathbf{y} \in \mathbb{R}^N$ and $\boldsymbol{\epsilon} \in \mathbb{R}^N$ are respectively the responses $\mathbf{y} = (y_1 \cdots y_i \cdots y_N)^T$ and the residual errors $\boldsymbol{\epsilon} = (\epsilon_1 \cdots \epsilon_i \cdots \epsilon_N)^T$.

2.5. Partially Ordered Models

It worth noting that we resort to compressions that define a structure of nested models shaped according to the hierarchies. If we denote $\mathcal{H}_{k\ell}$, the model obtained at heights k and ℓ for $k, \ell \geq 0$, then $\mathcal{H}_{k'\ell'}$ is defined as a richer decomposition of $\mathcal{H}_{k\ell}$ for $k' > k$ and $\ell' > \ell$.

For instance, with averaged groups of variables, we have

$$\mathcal{H}_{k\ell}: \quad y_i = \sum_{g \in \mathcal{G}^k} \tilde{x}_i^g \sum_{m \in \mathcal{M}^\ell} \tilde{x}_i^m \, \theta_{gm}^{k\ell} + \epsilon_i \,,$$

$$\mathcal{H}_{k\ell'}: \quad y_i = \sum_{g \in \mathcal{G}^k} \tilde{x}_i^g \sum_{m \in \mathcal{M}^\ell} \underbrace{\sum_{m' \in \mathcal{M}^{\ell'} \subset \mathcal{M}_m^\ell} \tilde{x}_i^{m'}}_{\tilde{x}_i^m} \theta_{gm'}^{k\ell'} + \epsilon_i \,.$$

We introduce the definition of nested models as follows.

Definition 1. *Let $\mathcal{H}_{k\ell}$ and $\mathcal{H}_{k'\ell'}$ being two models. $\mathcal{H}_{k\ell} \subseteq \mathcal{H}_{k'\ell'}$ if $k \leq k'$ and $\ell \leq \ell'$.*

Remark 1. *The complexity of a model $\mathcal{H}_{k\ell}$ increases with k and ℓ according to their respective number interactions. For $k' > k$ and $\ell' > \ell$,*

- *$\mathcal{H}_{k'\ell'}$ will always be more complex than $\mathcal{H}_{k\ell}$;*
- *$\mathcal{H}_{k\ell'}$ will be more complex than $\mathcal{H}_{k'\ell}$ iff $N_I^{k\ell'} > N_I^{k'\ell}$.*

Hence, the possible combinations of models form a partial order set that can be organized into a Directed Acyclic Graph (DAG) as pictured in Figures 1(a) and 1(b). This nested collection of models might be convenient for warm starting strategies or for model selection approaches for wich a nested structure is mandatory, such as likelihood ratio tests for instance. Regardless of these specific aspects, we will show in the next section how to build on this DAG structure to explore of the hierarchies efficiently.

3. Greed is Great

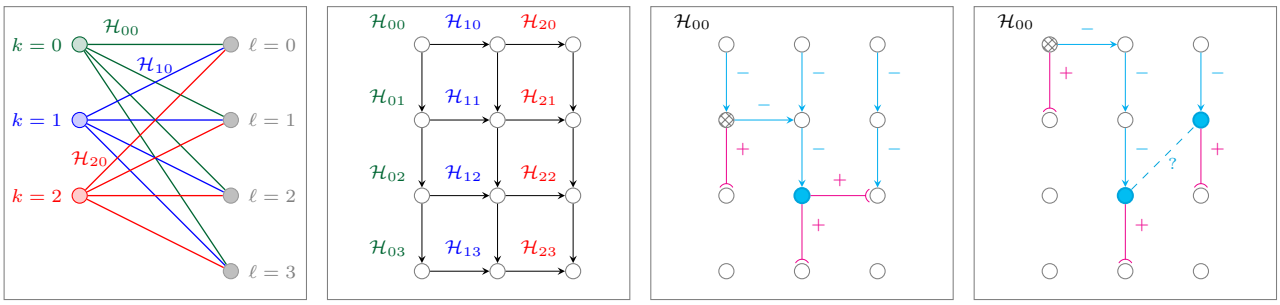
Algorithm 1 describes the greedy approach used to select the appropriate heights k and ℓ of Problem (3) efficiently. The two criterions $C_{\mathcal{H}_\bullet}$ and $C_{\mathcal{A}}$ involved to perform model selection at different stages are explained below.

3.1. Exploring the hierarchies

To explore all the possible combinations induced by the tree structures related to G and M, Problem (3) should be solved $H_G \cdot H_M$ times with $N_I^{k\ell}$ coefficients to estimate for each association of heights k and ℓ.

The resolution becomes more and more costly as heights k and ℓ increase. To avoid a complete exploration of the hierarchies, we build on the DAG structure presented in Figure 1(b) to derive a strategy based on criterion of convex trend for models of growing complexities (see Remark 1), such as AIC or BIC.

Our approach consists to explore the DAG along a first dimension chosen according to the depth of the hierarchies: k will be related to the inner loop in Algorithm 1



(a) Combination of models (b) Nested structure (c) DAG structured exploration (d) DAG structured exploration

Figure 1. (a) The set of possible models induced by both hierarchies with $H_G = 2$ and $H_M = 3$. (b) A DAG that describes the partial order on the models. If a path exists between $\mathcal{H}_{k\ell}$ and $\mathcal{H}_{k'\ell'}$, that is $\mathcal{H}_{k\ell} \cdots \rightarrow \cdots \mathcal{H}_{k'\ell'}$, then $\mathcal{H}_{k\ell} \subseteq \mathcal{H}_{k'\ell'}$. (c) and (d) Two examples of exploration reached with the proposed greedy strategy. A unique optimal model is obtained in Example (c) but a set of candidate models may be found as in Example (d). A rule to select the optimal model may be defined as described in Section 3.1.

Algorithm 1 Structured Hierarchical Active Set Algorithm

Inputs

$\mathcal{S}$	the training set
$\{\mathcal{G}^k\}_{k=0}^{H_G}$	the structure on view G
$\{\mathcal{M}^\ell\}_{\ell=0}^{H_M}$	the structure on view M
$P(\cdot)$	P builds matrices $\mathbf{\Phi}^{k\ell}$
$AS(\cdot)$	AS computes coefficients $\boldsymbol{\theta}^{k\ell}$
$R(\cdot)$	R chooses between $\mathcal{H}_{k\ell}$ and $\mathcal{H}_{k'\ell'}$

Outputs

$\hat{\mathcal{H}}$	the optimal model

// Global initialization
$\hat{\mathcal{H}} = \mathcal{H}_{00}, \quad \hat{k} = 0, \quad \hat{\ell} = 0$

for $k = 0$ **to** H_G **do**
  for $\ell = 0$ **to** H_M **do**
    // Initializations at heights k and ℓ
    $\mathbf{\Phi}^{k\ell} = P(\mathcal{S}, \mathcal{G}^k, \mathcal{M}^\ell) \in \mathbb{R}^{N \times N_I^{k\ell}}$
    $\bar{\mathcal{A}}^{k\ell} = \mathcal{G}^k \times \mathcal{M}^\ell, \qquad \mathcal{A}^{k\ell} = \{\emptyset\}$
    // Active Set procedure
    while $C_{\mathcal{A}}(\mathbf{\Phi}^{k\ell}, \boldsymbol{\theta}^{k\ell}, \mathbf{y})$ is not met **do**
      $[\mathcal{H}_{k\ell,}, \mathcal{A}^{k\ell}, \boldsymbol{\theta}^{k\ell}_\dagger] = AS(\mathbf{\Phi}^{k\ell}, \boldsymbol{\theta}^{k\ell}, \mathbf{y})$
      $\boldsymbol{\theta}^{k\ell} = \boldsymbol{\theta}^{k\ell}_\dagger$
    end while
    // Early stopping criterion
    if $C_{\mathcal{H}_1}(\mathcal{H}_{k\ell})$ or $C_{\mathcal{H}_2}(\mathcal{H}_{k\ell})$ is met **then**
      $[\hat{\mathcal{H}}, \hat{k}, \hat{\ell}] = R(\mathcal{H}_{k\ell}, \hat{\mathcal{H}})$
      break on the outer loop
    end if
  end for
end for

if $\max(H_G, H_M) = H_G$ and ℓ will be related to the inner loop otherwise.

Let k being the first dimension, ℓ being the second one [2] and $\hat{k}$ and $\hat{\ell}$ being the heights of the current best model according to the chosen criterion $C_{\mathcal{H}}$ which should be of convex shape for models of growing complexities. We stop going deeper along the first dimension ℓ if

- The criterion increases along ℓ

$$C_{\mathcal{H}_1}: \quad C_{\mathcal{H}}(\mathcal{H}_{k\ell}) > C_{\mathcal{H}}(\mathcal{H}_{k\ell-1}), \quad \text{or if}$$

- The criterion increases along k, for $k \geq \hat{k}$

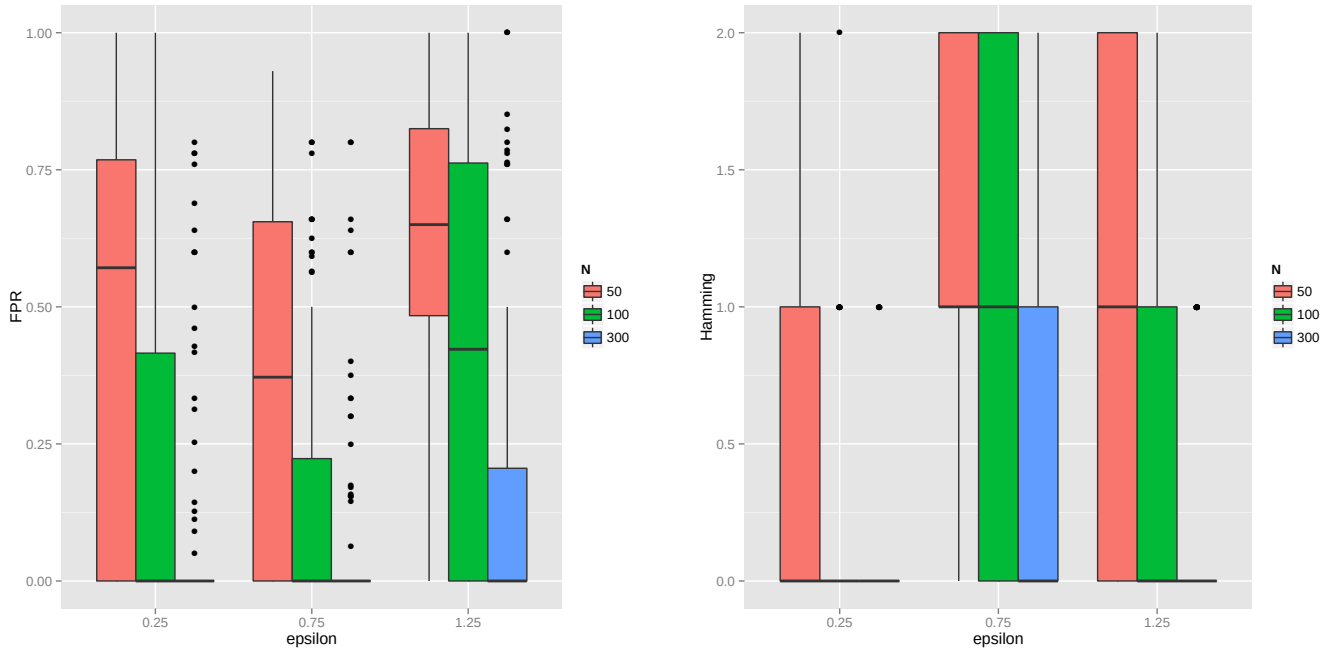
$$C_{\mathcal{H}_2}: \quad C_{\mathcal{H}}(\mathcal{H}_{k\ell}) > C_{\mathcal{H}}(\mathcal{H}_{\hat{k}\ell}).$$

This strategy is illustrated on two examples in Figures 1(c) and 1(d). One can notice on Figure 1(d) that two candidate models can emerge along the main dimension for non nested models. In this case, the rule $R(\mathcal{H}_{k\ell}, \mathcal{H}_{k'\ell'})$ for selecting the current best model may be defined according to the targeted application. For instance, we may choose the model that minimizes the MSE for prediction, the sparser model in terms of interactions involved at the level of variables for interpretation issues or the model that minimizes $C_{\mathcal{H}}$ for a compromise between these two aspects.

3.2. Active set procedure

The procedure $AS(\cdot)$ in Algorithm 1 can be based either on homotopy (Osborne et al., 2000; Efron et al., 2004) or on pursuit (Mallat and Zhang, 1993) strategies. When there are enough examples N compared to the number of interactions $N_I^{k\ell}$, a classical stepwise regression procedure may also be used.

[2] For the sake of clarity, we use this convention to describe the strategy and Algorithm 1 in relation with the examples given in Figures 1(c) and 1(d).



(a) FDP

(b) Hamming distance

Figure 2. Illustration of the behavior of Algorithm 1 in terms of variable selection considering FDP (a) and Hamming distance (b) for different numbers of examples and different levels of noise.

The purpose of $C_{\mathcal{A}}$ is to stop adding variables involved into a model $\mathcal{H}^{k\ell}$ according to the chosen strategy. It can be based on optimality conditions, on the value of the mean squared error or simply on the number of variables the user wants to add at given levels k and ℓ of the hierarchies.

4. Numerical simulations

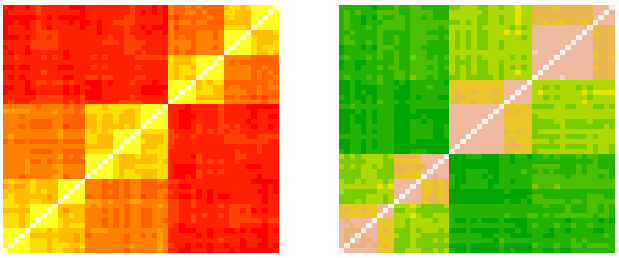


Figure 3. Examples of hierarchical structures on the matrices of correlation of $\mathbf{X}^G$ (red) and $\mathbf{X}^M$ (green).

To illustrate our approach, we have simulated two Gaussian hierarchies, one for each view, with 50 variables and 5 levels as illustrated on Figure 3. For fixed values of k and ℓ, we have generated compressed data, using averaged groups of variables, according to Model (3) so that 10% of the coefficients of $\boldsymbol{\theta}^{k\ell}$ have a non zero value. The process is repeated 100 times for each couple of parameters in $N = \{50, 100, 300\} \times \text{mean}(\boldsymbol{\epsilon}) = \{0.25, 0.75, 1.25\}$.

For each run, we evaluate the quality of the variable selection through an Hamming distance and a False Discovery Proportion (FDP). For $\hat{\mathcal{H}}_{k\ell}$, the model given by Algorithm 1, the Hamming distance counts the inconsistent matchings between $\hat{k}$ and $\hat{\ell}$ and the heights of the true model so that its value can be 0 for a perfect matching, 1 when only one height is recovered or 2 when there is no correspondance. The FDP consists to test the relevance of the estimated interactions at heights $k = H_G$ and $\ell = H_M$. More precisely, we compare the true model $\mathcal{H}_{k\ell}$ with $\hat{\mathcal{H}}_{k\ell}$ and extend the corresponding $\boldsymbol{\theta}^{k\ell}$ and $\hat{\boldsymbol{\theta}}^{k\ell}$ at the variable levels. For all possible interactions $\{gm\}$, we then determine the following confusion matrix and hence the FDP:

	$\hat{\theta}_{gm}^{H_G H_M} = 0$	$\hat{\theta}_{gm}^{H_G H_M} \neq 0$
$\theta_{gm}^{H_G H_M} = 0$	True Negative	False Positive
$\theta_{gm}^{H_G H_M} \neq 0$	False Negative	True Positive

The results given in Figure 2 show that Algorithm 1 behaves rather well when there are enough examples ($N \geq 100$) especially with a reasonable level of noise ($\text{mean}(\boldsymbol{\epsilon}) \leq 0.75$).

References

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Escofier, B. and Pagès, J. (2008). *Analyses factorielles simples et multiples: objectifs, méthodes et interprétation*. Dunod.

Lê, S., Josse, J., Husson, F., et al. (2008). Factominer: an R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Mallat, S. G. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403.