Nonlinear network-based traits prediction from cytokines/chemokines secretion profiles

Emeline Perthame [1] & Vincent Guillemot [1] & Emilie Devijver [2] & Mélina Gallopin [3] & Violetta Zujovic [4] & Jennifer Fransson [4] & Charles Sanson [4] Hervé Abdi [5]

[1] *Institut Pasteur - Hub Bioinformatique et Biostatistique - C3BI, USR 3756 IP CNRS - Paris, France*

[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, 38000 Grenoble, France*

[3] *Institute for Integrative Biology of the Cell, CEA, CNRS, Université Paris-Sud, Université Paris Saclay, Gif-sur-Yvette, France*

[4] *Sorbonne-Universités-UPMC 06, INSERM, CNRS, UMR ICM-75-1127-7225, 47 boulevard de l'Hôpital, 75013 Paris, France*

[5] *School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA*

Résumé. Dans le cadre d'une étude sur l'hétérogénéité de patients atteints de sclérose en plaques, nous présentons l'analyse d'un jeu de données composé de cas et de témoins, pour lesquels un ensemble de cytokines a été mesuré, ainsi que plusieurs marqueurs de remyélinisation. Pour cela, nous proposons d'appliquer un modèle de mélanges de régressions intitulé BLLiM permettant d'extraire des clusters d'individus présentant des profils similaires, ainsi que des modules de cytokines corrélées entre elles. Ce modèle de prédiction a l'avantage d'offrir une grande interprétabilité. Dans cet exposé, nous présenterons dans un cadre statistique ce modèle paramétrique ainsi que les résultats biologiques déduits de son application sur les données. De plus, ce modèle sera comparé à des approches similaires, tant en termes de classification que de prédiction.

Keywords. Biostatistique, interprétabilité, modélisation, mélange de régressions, réseaux de régulations, heuristique de pente

Abstract. We are interested in the heterogeneity of patients affected by multiple sclerosis. The dataset is composed of patients and healthy subjects, for which a set of cytokines were measured, as well as remyelination markers. We propose to apply a model based on a mixture of regressions called BLLiM which allows to extract clusters of subjects exhibiting similar profiles as well as modules of correlated cytokines. This prediction model is fully parametric and highly interpretable. In this talk, the statistical characteristics of the model will be presented and the clustering and prediction accuracy of this model will be compared to similar approaches. The biological results deduced from the application of BLLiM to the described dataset will be presented and discussed.

Mots-clés. Biostatistics, interpretability, model-based approach, mixture of regressions, regulatory networks, slope heuristics

[1]Institute of Engineering Univ. Grenoble Alpes

1 Biological and statistical questions

Multiple sclerosis is a chronic disease of the central nervous system characterized by a strong degradation of myelin sheaths all around the patient's nervous system. In healthy individuals, myelin sheaths are also degraded but a remyelination process exists that is able to repair the damage. In this article, we focus on the analysis of a dataset aiming at exploring the link between cytokines (the cell messengers) and the process handling the remyelination of damaged axons.

For each one of the 33 subjects of the study (26 patients and 7 healthy donors), the concentrations of 67 cytokines are measured, in addition to 3 remyelination markers: M2Arg1, M1iNOS and CNPase. These three remyelination markers are characteristic of different processes. Biologists and clinicians have separated the subjects into 3 groups according to their capacity to remyelinate: normal (healthy donors), high and low remyelination capacity. Although the partition into 3 groups was relevant and led to interesting biological results [4], it does not reflect the whole complexity of the measurements. We propose to apply a model-based approach, named Block diagonal Locally Linear Mapping (BLLiM) [3], that identifies clusters of patients exhibiting similar profiles of relations between cytokines and remyelination markers. This parametric model allows to investigate the distribution of the 3 biological markers and the correlation pattern of the selected cytokines within each identified cluster of patients.

In the first part of the analysis, we take into account the initial group partition (in three groups: healthy donors, high and low remyelinators) by initializing the cluster affectations in the algorithm used to estimate the parameters of the BLLiM model. This initialization leads to interesting results from both biological and statistical viewpoints. Indeed, the model exhibits a satisfying adjustment on the data, which suggests that it can be reliably interpreted. Moreover, the results are concomitant with some of the conclusions of a previous study of this dataset detailed in [4] regarding both the clusters of individuals and the modules of cytokines. The specificity of this initialization is that it is based solely on the CNPase profiles. BLLiM's model being based on a mixture of regressions models, it is therefore designed to cluster data according to the joint profiles of all the 3 responses and covariates. Interestingly, BLLiM's clustering of the individuals is very similar to the initialization even for the two other responses (except for 2 patients).

In the second part of the analysis, we use a standard procedure to initialize the algorithm, meaning that the clusters are randomly attributed and the number of clusters is supposed unknown. This more flexible approach has the advantage that the number of clusters is not anymore fixed to 3 and needs to be optimized. This analysis leads to new biological results that need to be interpreted and may give a new insight on the data. According to an adjustment quality criterion based on the likelihood, we notice that random initialization leads to a better fit of the model.

During the talk, the two models will be deeper interpreted and compared, both graphically and according to their prediction accuracy using cross-validation.

2 Filtering out unwanted variables with GSPPCA

First, the cytokine concentrations are log2 transformed in order to get closer to a Gaussian distribution, which is an usual transformation in such a field of application.

Besides, the resulting modules that BLLiM computes are quite sensitive to the ratio n/p: the larger the number of variables, the more likely the modules are to contain only one variable. To increase the interpretability of the results, we filter out the variables that have the lowest variance with GSPPCA [5]. In a nutshell, this Bayesian method maximizes a Bayesian PCA likelihood (based on a symmetric multivariate Bessel distribution) via a variational expectation-maximization algorithm and, in the process, manages to identify a group of variables as "noiseless", and a group of noise variables. GSPPCA identifies 25 cytokines as noise variables which are discarded from the analysis.

3 Predict markers from cytokine concentration levels using BLLiM

Let $\mathbf{Y}_i \in \mathbb{R}^L$ denote, for a given subject i, a response vector ($L = 3$ biological markers in this application) and $\mathbf{x}_i$ denotes its corresponding profile of cytokine levels. We are interested in exploring and modeling the non linear relation between $\mathbf{Y}_i$ and $\mathbf{x}_i$. In this context of non linear modeling, several authors (cite [2], [6] et [3]) proposed to approximate the regression function of interest by K linear affine functions, viz: mixture of regressions model:

$$\mathbf{Y}_i = \sum_{k=1}^{K} \mathbb{I}_{Z_i=k}(\mathbf{A}_k^*\mathbf{x}_i + \mathbf{b}_k^* + \boldsymbol{\varepsilon}_{ki}) \text{ with } \boldsymbol{\varepsilon}_{ki} \sim \mathcal{N}_D(0, \boldsymbol{\Sigma}_k^*) \tag{1}$$

where the latent variable $Z_i = k$ indicates that individual $i \in \{1, \ldots, n\}$ comes from cluster k, $k \in \{1, \ldots, K\}$. Within each cluster k, an affine relation occurs between cytokines $\mathbf{x}_i$ and the response $\mathbf{Y}_i$. The non linear regression function is approximated by a weighted combination of these K affine functions, with weights depending only on the cytokine profiles $\mathbf{x}_i$. This model has therefore the advantage to propose a fully parametric approach to approximate a non linear regression function.

One can notice that this model can be written as a mixture of regressions model:

$$\begin{aligned} &\mathbb{P}(\mathbf{X}_i = \mathbf{x}|Z_i = k) = \varphi_D(\mathbf{x}; \mathbf{c}_k^*, \boldsymbol{\Gamma}_k^*); \\ &\mathbb{P}(\mathbf{Y}_i = \mathbf{y}|\mathbf{X}_i = \mathbf{x}, Z_i = k) = \varphi_L(\mathbf{y}; \mathbf{A}_k^*\mathbf{x} + \mathbf{b}_k^*, \boldsymbol{\Sigma}_k^*) \end{aligned} \tag{2}$$

where φ stands for the probability distribution function of a Gaussian distribution. We also assume that the covariates $\mathbf{x}$ are Gaussian with expectation $\mathbf{c}_k^*$ and variance $\boldsymbol{\Gamma}_k^*$. If a new profile $\mathbf{x}_{n+1}$ of cytokines is observed, this model can be used to perform p rediction

by taking the expectation in model (1) and integrating on the latent variable Z:

$$\hat{\mathbf{Y}}_{n+1} = \mathbb{E}(\mathbf{Y}_{n+1}|\mathbf{X}_{n+1} = \mathbf{x}_{n+1}).$$

To estimate such a model, several authors (initially [2], then [6] and [3]) proposed an estimation approach based on the inverse regression model deduced from the forward regression model presented at Equation (2). Indeed, model (2) is totally equivalent to the following inverse mixture of regressions:

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}_i = \mathbf{y}|Z_i = k) = \varphi_L(\mathbf{y}; \mathbf{c}_k, \mathbf{\Gamma}_k) \\
&\mathbb{P}(\mathbf{X}_i = \mathbf{x}|\mathbf{Y}_i = \mathbf{y}, Z_i = k) = \varphi_D(\mathbf{x}; \mathbf{A}_k\mathbf{y} + \mathbf{b}_k, \mathbf{\Sigma}_k).
\end{aligned} \tag{3}$$

Notice that closed form expressions allow to go from model (2) to model (3) equivalently:

$$\begin{aligned}
\mathbf{c}_k^* &= \mathbf{A}_k\mathbf{c}_k \\
\mathbf{\Gamma}_k^* &= \mathbf{\Sigma}_k + \mathbf{A}_k\mathbf{\Gamma}_k\mathbf{A}_k^T \\
\mathbf{A}_k^* &= \mathbf{\Sigma}_k^*\mathbf{A}_k^T\mathbf{\Sigma}_k^{-1} \\
\mathbf{b}_k^* &= \mathbf{\Sigma}_k^*(\mathbf{\Gamma}_k^{-1}\mathbf{c}_k - \mathbf{A}_k^T\mathbf{\Sigma}_k^{-1}\mathbf{b}_k) \\
\mathbf{\Sigma}_k^* &= (\mathbf{\Gamma}_k^{-1} + \mathbf{A}_k^T\mathbf{\Sigma}_k^{-1}\mathbf{A}_k)^{-1}.
\end{aligned}$$

Without any assumption on the parameters, estimating the forward or the inverse model is equivalent. Nevertheless, under relevant assumptions on the inverse model, this one becomes easier to estimate than forward model, while being less stringent and more realistic. The authors of [2] assume that the residual matrix $\mathbf{\Sigma}_k$ is diagonal, with the possibility to add latent factors and assume a low rank decomposition for $\mathbf{\Sigma}_k$.

The same assumption is applied in [6], except that the authors considered a Generalized multivariate Student distribution in order to get a robust counterpart of [2]. In [3], the authors assume a block-diagonal structure for $\mathbf{\Sigma}_k$. To find the right shape for $\mathbf{\Sigma}_k$, the authors build a model collection of block structures $\mathbf{B}$ ordered by complexity and driven by the data. Among this collection of structures, the one which reaches a relevant compromise between number of parameters and goodness of fit is chosen using slope heuristic proposed by [1]:

$$\hat{\mathbf{B}}_K = \underset{\mathbf{B}}{\operatorname{argmax}} \left\{\ell(\mathbf{B}) - 2a \times \Delta_{\mathbf{B}}\right\} \tag{4}$$

where $\ell(\mathbf{B})$ is the log-likelihood of each model of the collection, $\Delta_{\mathbf{B}}$ is the number of parameters and a is the penalization parameter deduced by the slope heuristic.

More details about the method can be found in [3].

4 Preliminary results

This section presents preliminary results and the output of the model initialized with the partition of patients into 3 groups (healthy, high and low remyelination capacity).

Figure 1a presents the distribution of the 3 responses by cluster. It shows that the algorithm globally retrieves the affectations except for 2 subjects. Next, Figure 1b shows the inferred components of cytokines in the matrix $(\mathbf{\Sigma}_k)_{k=1}^K$. For the cluster of healthy donors, the method tends to build small modules of 2 or 3 connected cytokines while for the 2 clusters of patients, the method tends to create highly interconnected components of cytokines. We also notice that all cytokines are positively correlated so each edge corresponds to a positive correlation. Finally, Figure 1c presents the regression coefficients of matrix $(\mathbf{A}_k^*)_{k=1}^K$ displaying the effect of each cytokines on each response, within each cluster. This figure allows to detect a positive or negative effect of the cytokine's level on the response.

In the future, we will work on including sparsity constraints in the model. This is motivated by the fact that a lot of coefficients found by BLLiM are close to 0, the interpretability of the model will therefore be increased by a frank removal of these coefficients from the model.

References

[1] Birgé, L. and Massart, P. (2006). Minimal penalties for gaussian model selection. Probability Theory and Related Fields, 138(1-2):3373.

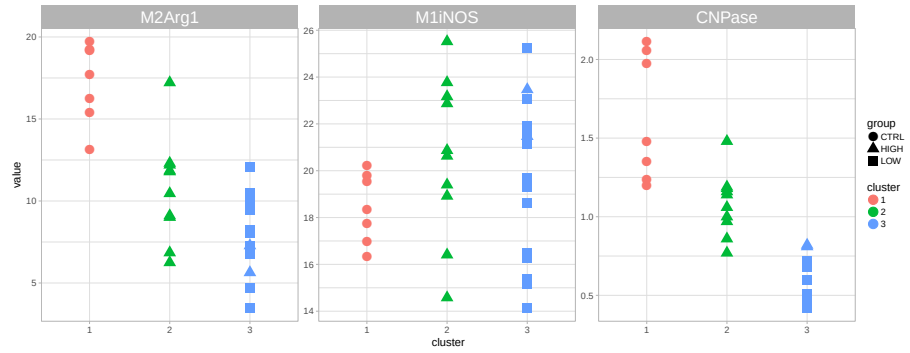
[2] Deleforge, A. and Forbes, F. and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. Statistics and Computing, 25(5):893–911.

[3] Devijver, E. and Gallopin, M. and Perthame, E. (2017), Nonlinear network-based quantitative trait prediction from transcriptomic data, *arXiv:1701.07899*

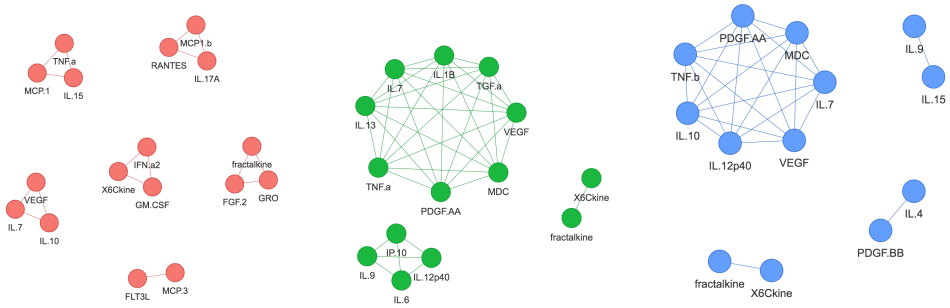
[4] El Behi, M. and Sanson, C. and Bachelin, C. and Guillot-Noël, L. and Fransson, J. and Stankoff, B. and Maillart, E. and Sarrazin, N. and Guillemot, V. and Abdi, H. and Cournu-Rebeix, I. and Fontaine, B. and Zujovic, V. (2017), Adaptive human immunity drives remyelination in a mouse model of demyelination. Brain, 4(170):967–980.

[5] Mattei, P.-A. and Bouveyron C. and Latouche, P. (2016). Globally Sparse Probabilistic PCA, Proc. AISTATS 2016, pp. 976-984

[6] Perthame E. and Forbes, F. and Deleforge, A. (2018). Inverse regression approach to robust nonlinear high-to-low dimensional mapping. Journal of Multivariate Analysis, 163:1–14.



(a)



(b)

(c)

Figure 1: (a) **Distribution of the response within clusters.** The three graphs represent the distribution for each one of the three biomarkers (M2Arg1, M1iNOS and CNPase). For each graph, the x-axis represents the assigned clusters of individuals (cluster 1, 2 or 3). The y-axis represents the value of the biomarkers. The color of the points represents the assigned clusters and their shape the original clusters. (b) **Inferred components of cytokines.** The 3 colored graphs represent the correlation between cytokines for each cluster of individuals. The color corresponds to the assigned clusters. (c) **Regression coefficients for each clusters for each biomakers.** The three graphs correspond to each biomarkers. The x-axis corresponds to the cluster of individuals. The y-axis corresponds the 42 cytokines used to predict the biomarkers. The color represents the value of the regression coefficients in the BLLiM model.