Multiple Linear Regression: Bayesian Inference for Distributed and Big Data in the Medical Informatics Platform of the Human Brain Project

Lester Melie-Garcia[1]✉, Bogdan Draganski[1,2], John Ashburner[3], Ferath Kherif[1]

[1] LREN, Department of Clinical Neurosciences, Lausanne University Hospital (CHUV), Switzerland
[2] Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[3] Wellcome Trust Centre for Imaging Neuroscience, University College London (UCL), United Kingdom (UK)

✉ Correspondence: Mont Paisible 16, 1011 Lausanne, Switzerland. Phone: +41 21 314 9593. Fax: +41 21 314 1256;
E-mail address: lester.melie@gmail.com

ABSTRACT: We propose a Multiple Linear Regression (MLR) methodology for the analysis of distributed and Big Data in the framework of the Medical Informatics Platform (MIP) of the Human Brain Project (HBP). MLR is a very versatile model, and is considered one of the workhorses for estimating dependences between clinical, neuropsychological and neurophysiological variables in the field of neuroimaging. One of the main concepts behind MIP is to federate data, which is stored locally in geographically distributed sites (hospitals, customized databases, etc.) around the world. We restrain from using a unique federation node for two main reasons: first the maintenance of data privacy, and second the efficiency in management of big volumes of data in terms of latency and storage resources needed in the federation node. Considering these conditions and the distributed nature of data, MLR cannot be estimated in the classical way, which raises the necessity of modifications of the standard algorithms. We use the Bayesian formalism that provides the armamentarium necessary to implement the MLR methodology for distributed Big Data. It allows us to account for the heterogeneity of the possible mechanisms that explain data sets across sites expressed through different models of explanatory variables. This approach enables the integration of highly heterogeneous data coming from different subjects and hospitals across the globe. Additionally, it offers general and sophisticated ways, which are extendable to other statistical models, to suit high-dimensional and distributed multimodal data. This work forms part of a series of papers related to the methodological developments embedded in the MIP.

Keywords: multiple linear regression, general linear model, linear model, linear regression, Bayesian linear model, Bayesian linear regression, MLR, parallel computation, distributed computation, Human Brain Project, Bayesian modeling, model averaging, variational Bayes.

INTRODUCTION

The amount of neuroimaging data (i.e. MRI, fMRI, PET, EEG, etc.) continues to expand exponentially with the production of thousands of studies worldwide as part of clinical and research activities. As an example, after just a few decades of the Magnetic Resonance Imaging technique (MRI), hundreds of thousands of individuals of all ages and conditions have been scanned and the rate of MRI data collection is set to grow considerably over the coming years. Unfortunately, the majority of this data is locally warehoused in laboratories and hospitals and is therefore poorly capitalized into new knowledge, scientific publications and the development of novel medical therapies. This raises the urgent need for a more efficient exploitation of such limitless richness of information in order to shed light on the principles of brain anatomy and function in both healthy and pathological states. In recent years, in light of this necessity, the neuroscientific and software developer's communities have joined efforts to embrace new paradigms related to data sharing, virtualization and federation. These paradigms provide a framework to integrate large amounts of data that will undoubtedly help in generating more realistic models of brain function. Data integration is an effective approach devoted to appropriately combining neuropsychological, genetic, clinical and neuroimaging multimodal data (at the micro, meso and macro scales), in a very heterogeneous population coming from different parts of the globe under dissimilar conditions (race, education level, dietary conditions, etc.).

Precisely, the Medical Informatics Platform (MIP) as part of the Human Brain Project (Subproject 8 (SP8)) (https://www.humanbrainproject.eu/en/medicine/medical-informatics-platform/), is conceived to provide an informatics infrastructure to the neuroscientists and clinicians studying brain anatomy and function using clinical along with multimodal big data. This platform faces the challenges of adapting, designing and implementing novel technologies in terms of hardware (clusters etc.), software (i.e. NoSQL, Apache Spark) and data analytics (i.e. decentralized machine learning) to deal with the large and geographically distributed nature of the data. In this paper, we propose a theoretical framework based on the 'Bayesian formalism' that contributes to the development of mathematical tools to deal with big and distributed data as part of the MIP environment. Rather than introducing a fully new mathematical development, we gather together and adapt general advances previously described in the Bayesian modeling literature to ultimately propose our methodological framework. We extended these developments to the versatile Multiple Linear Regression (MLR), which is considered one of the workhorses of statistics and data modeling in the neuroimaging field. The Bayesian formalism is ideal for the core of our theoretical machinery since it deals, in a natural way, with the problem of having big, heterogeneous and distributed data.

The paper in general is organized as follows. The main section, 'Theory', is divided into two main subsections. The first is devoted to providing the grounds of the general Bayesian framework for distributed data in two 'regimes': Parallel and Streaming. A general schema is illustrated that defines the local (hospitals) and federation (global) nodes; the latter federates and delivers the information to the end user. Based on previous works, a model selection/averaging step is proposed, constituting the third level of Bayesian inference. The second subsection describes a special case of the general principles outlined in the former subsection, which are applied to the multiple linear regression model. In addition, a summary of the general equations to be used in practice is shown.

THEORY

1. Problem Statement

One of the main goals of data analysis is to identify the models that best explain the data, along with their intrinsic parameter distributions, in order to discover basic organizational and functional principles of natural systems. Conventional modeling setups and algorithms presuppose that data is stored in one single location as a unique data source (i.e. database). Therefore, new techniques are required to model data when it is large and geographically distributed.

2. General Bayesian computational framework: Parallel approach.

Figure 1 shows the general scheme of our computational framework for distributed data. Data is dispersed across different remote sites, which in our case are hospitals/databases that may be located in different countries. The upper part of the figure represents the Federation node (aka master node, centralized node) that federates the aggregated information (or aggregates) coming from local nodes or hospitals (aka decentralized nodes, worker nodes) represented in the lower part of the figure. Within the framework of this paper, these aggregates are expectations of various model parameters. In this case, the federated estimation of parameter distributions is obtained using local aggregates provided in parallel by all local nodes. This setting is denominated as 'parallel' (Broderick et al., 2013).

In this paradigm, the information can be interchanged in both directions. Within an iterative algorithm, the federation node provides intermediate parameter estimates $\hat{\boldsymbol{\Theta}}$ downstream and receives aggregates ($\boldsymbol{\Theta}^{(i)}$) from each local node. The set of aggregates from Hospital '*i*' is represented by $\boldsymbol{\Theta}^{(i)} = \left[\theta_1^{(i)}, \ldots, \theta_N^{(i)}\right]$ and y_i is the data for the hospital '*i*', where *H* is the number of hospitals so that $\mathbf{y} = \left[y_1 \ldots, y_H\right]$.

We proposed the use of the Bayesian formalism as a theoretical framework to define our computational engine, which we named '*Bayesian computational framework*' (BCF). Therefore our BCF is based on the fundamental Bayes theorem so that the estimation of the parameters $\hat{\boldsymbol{\Theta}}$ at the federation node is expressed by the following equation:

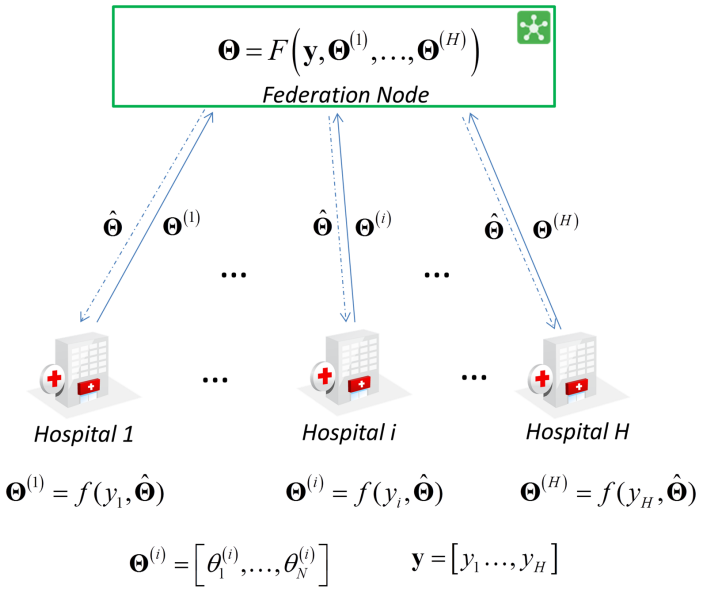

Figure 1. General scheme of the parallel paradigm in the Medical Informatics Platform (Human Brain Project).

$$p\left(\boldsymbol{\Theta}|\mathbf{y}, M_k\right) \propto p\left(\mathbf{y}|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right) \tag{1}$$

where $p\left(\boldsymbol{\Theta}|\mathbf{y}, M_k\right)$ is the posterior distribution of parameters $\boldsymbol{\Theta}$ for the specific model M_k and $\hat{\boldsymbol{\Theta}}$ the set of parameter values that maximize $p\left(\boldsymbol{\Theta}|\mathbf{y}, M_k\right)$; $p\left(\mathbf{y}|\boldsymbol{\Theta}, M_k\right)$ is the likelihood (taking into account data from all Hospitals) and $p\left(\boldsymbol{\Theta}\right)$ the prior probability of the parameters of interest.

Assuming that chunks of data belonging to each hospital are independent, the likelihood can be expressed as:

$$p\left(\mathbf{y}|\boldsymbol{\Theta}, M_k\right) = p\left(\mathbf{y}_1|\boldsymbol{\Theta}, M_k\right) p\left(\mathbf{y}_2|\boldsymbol{\Theta}, M_k\right) \cdots p\left(\mathbf{y}_H|\boldsymbol{\Theta}, M_k\right) \tag{2}$$

Substituting equation (2) into (1), we obtain:

$$p\left(\boldsymbol{\Theta}|\mathbf{y}, M_k\right) \propto p\left(\mathbf{y}_1|\boldsymbol{\Theta}, M_k\right) p\left(\mathbf{y}_2|\boldsymbol{\Theta}, M_k\right) \cdots p\left(\mathbf{y}_H|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right) \tag{3}$$

Multiplying and dividing *M-1* times by the prior probability $p\left(\boldsymbol{\Theta}\right)$ in equation (3) leads to:

$$p\left(\boldsymbol{\Theta}|\mathbf{y}, M_k\right) \propto \frac{\overbrace{p\left(\mathbf{y}_1|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right)}^{p(\boldsymbol{\Theta}|\mathbf{y}_1, M_k)} \overbrace{p\left(\mathbf{y}_2|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right)}^{p(\boldsymbol{\Theta}|\mathbf{y}_2, M_k)} \cdots \overbrace{p\left(\mathbf{y}_H|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right)}^{p(\boldsymbol{\Theta}|\mathbf{y}_H, M_k)}}{\left(p\left(\boldsymbol{\Theta}\right)\right)^{H-1}} \tag{4}$$

$$p\left(\boldsymbol{\Theta}|\mathbf{y}, M_k\right) \propto \frac{p\left(\boldsymbol{\Theta}|\mathbf{y}_1, M_k\right) p\left(\boldsymbol{\Theta}|\mathbf{y}_2, M_k\right) \cdots p\left(\boldsymbol{\Theta}|\mathbf{y}_H, M_k\right)}{\left(p\left(\boldsymbol{\Theta}\right)\right)^{H-1}} = \frac{1}{\left(p\left(\boldsymbol{\Theta}\right)\right)^{H-1}} \prod_{j=1}^{H} p\left(\boldsymbol{\Theta}|\mathbf{y}_j, M_k\right) \tag{5}$$

Then the posterior probability of the parameters $\boldsymbol{\Theta}$ for a specific model M_k at the federation node is expressed as a product of the posterior probabilities coming from the '*H*' Hospitals. The maximum of the posterior distribution for $\boldsymbol{\Theta}$ is achieved at $\hat{\boldsymbol{\Theta}}$ values.

2.1 General Bayesian computational framework: Streaming approach.

In many situations, providing a federated estimate of the parameters of interest when all local estimates are ready to use is very inefficient. Such situations become more critical when there are long delays between local parameter estimations. Hence it has been proposed, when it is tractable, the use of a subset of local estimates to obtain intermediate federated results.

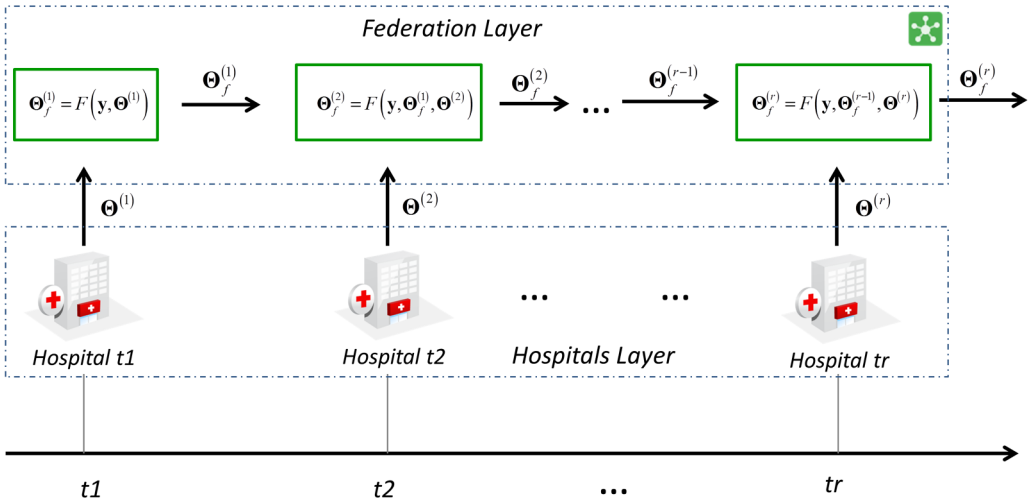

Figure 2. General scheme of the streaming paradigm in the Medical Informatics Platform (Human Brain Project). At the federation layer the new local estimates coming from the hospitals layer are combined with previous ones to deliver federated results.

Therefore as a new local estimate arrives to federation node an update of the federated estimation is performed, without recalculating parameters over again. This general scheme (Figure 2) is known as 'Streaming' (Broderick et al., 2013), and under the Bayesian formalism can be expressed through the following equations:

$$
\begin{aligned}
&- \quad p\left(\boldsymbol{\Theta}|\mathbf{y}_1, M_k\right) \propto p\left(\mathbf{y}_1|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right) \\
&- \quad p\left(\boldsymbol{\Theta}|\mathbf{y}_1, \mathbf{y}_2, M_k\right) \propto p\left(\mathbf{y}_1, \mathbf{y}_2|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right) = p\left(\mathbf{y}_2|\boldsymbol{\Theta}, M_k\right) \overbrace{p\left(\mathbf{y}_1|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right)}^{p\left(\boldsymbol{\Theta}|\mathbf{y}_1, M_k\right)} \\
&\qquad\qquad = p\left(\mathbf{y}_2|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}|\mathbf{y}_1, M_k\right) \\
&\qquad\qquad \vdots \\
&- \quad p\left(\boldsymbol{\Theta}|\mathbf{y}_1, \ldots, \mathbf{y}_r, M_k\right) \propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_r|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right) = \\
&= p\left(\mathbf{y}_r|\boldsymbol{\Theta}, M_k\right) \overbrace{p\left(\mathbf{y}_{K-1}|\boldsymbol{\Theta}, M_k\right) \cdots \overbrace{p\left(\mathbf{y}_1|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}\right)}^{p\left(\boldsymbol{\Theta}|\mathbf{y}_1\right)}}^{p\left(\boldsymbol{\Theta}|\mathbf{y}_1, \ldots, \mathbf{y}_{r-1}\right)}} = \\
&= p\left(\mathbf{y}_r|\boldsymbol{\Theta}, M_k\right) p\left(\boldsymbol{\Theta}|\mathbf{y}_1, \ldots, \mathbf{y}_{r-1}, M_k\right)
\end{aligned}
\tag{6}
$$

From the recurrence equations in (6) we have that a new posterior at the federation node, when a specific '*r*' data is ready from a local site, will be the multiplication of the previous posterior distribution by the likelihood function of the new data. Since the data cannot be distributed out of local sites, the posterior

$p\left(\boldsymbol{\Theta}\middle|\mathbf{y}_1,\ldots,\mathbf{y}_{K-1}\right)$ should be sent to the local site '*r*' for updating the federated posterior that is finally, in a new version, pull back to the federation node.

On the other hand, equation (6) can also be rewritten as:

$$p\left(\boldsymbol{\Theta}\middle|\mathbf{y},M_k\right)\propto\frac{\overbrace{p\left(\mathbf{y}_r\middle|\boldsymbol{\Theta},M_k\right)p\left(\boldsymbol{\Theta}\right)}^{p\left(\boldsymbol{\Theta}\middle|\mathbf{y}_r\right)}}{p\left(\boldsymbol{\Theta}\right)}p\left(\boldsymbol{\Theta}\middle|\mathbf{y}_1,\ldots,\mathbf{y}_r,M_k\right)=\frac{p\left(\boldsymbol{\Theta}\middle|\mathbf{y}_r,M_k\right)p\left(\boldsymbol{\Theta}\middle|\mathbf{y}_1,\ldots,\mathbf{y}_r,M_k\right)}{p\left(\boldsymbol{\Theta}\right)}\tag{7}$$

Therefore the new posterior can in addition be computed by equation (7) passing the posterior of parameters $\boldsymbol{\Theta}$ from the '*r*'th hospital to the federation node to be accordingly combined with the previous posterior. That way is more convenient since the communication between federation and local nodes is preserved in only one direction, which turns out in practice to be simpler to implement.

2.2 Model selection and model averaging: general definitions

As we stated at the beginning of the THEORY section, the selection of one or a set of models that explain our data is key in data analysis. In this subsection we treated the equations for the third level of inference in the Bayesian formalism known as model selection and model averaging (Hoeting et al., 1999; MacKay, 1992; Penny et al., 2006) applied to our particular case.

Given a set of '*K*' candidate models $M_1,\ldots,M_K$ for explaining our data, we need to calculate the posterior probability of the models to ultimately either select the best one or apply a model averaging step. In our case we assume the model selection/model averaging to be performed at the federation node, though, this can be implemented at each local site to finally be combined at the federation node. If we assume the data is generated by the same model M_k in all local sites we are in the presence of a Fixed Effects (FFX) analysis. Otherwise if a more relaxed and general assumption is adopted, allowing for the possibility that different data sites use different models, a Random Effect (RFX) analysis is carried out. In the following subsections we separately analyzed both types of analysis.

2.2.1 Model selection: Fixed Effects (FFX) Analysis

The probability distribution for specific model '*r*' is expressed by the well-known equation:

$$p\left(M_k\middle|\mathbf{y}\right)=\frac{p\left(\mathbf{y}\middle|M_k\right)p\left(M_k\right)}{\sum_{i=1}^{K}p\left(\mathbf{y}\middle|M_i\right)p\left(M_i\right)}\tag{8}$$

where $p\left(\mathbf{y}\middle|M_k\right)$ is the evidence of the model or the marginal likelihood that accounts for all data chunks associated with local sites, $p\left(M_k\right)$ the prior probability of M_k that expresses our belief or prior knowledge about the occurrence of the model M_k. The evidence of the model M_k is expressed by:

$$p\left(\mathbf{y}\middle|M_k\right)=\int p\left(\mathbf{y},\boldsymbol{\Theta},M_k\right)d\boldsymbol{\Theta}=\int p\left(\mathbf{y}\middle|\boldsymbol{\Theta},M_k\right)p\left(\boldsymbol{\Theta}\middle|M_k\right)d\boldsymbol{\Theta}\tag{9}$$

Because of the condition of independence across data sites expressed in equation (2) we have that $p\left(M_k\middle|\mathbf{y}\right)$ is:

$$p\left(M_k|\mathbf{y}\right)=\frac{p\left(\mathbf{y}|M_k\right)p\left(M_k\right)}{\sum_{j=1}^{K}p\left(\mathbf{y}|M_j\right)p\left(M_j\right)}=\frac{p\left(\mathbf{y}_1|M_k\right)\cdots p\left(\mathbf{y}_H|M_k\right)p\left(M_k\right)}{\sum_{j=1}^{K}p\left(\mathbf{y}_1|M_j\right)\cdots p\left(\mathbf{y}_H|M_j\right)p\left(M_j\right)}$$

$$p\left(M_k|\mathbf{y}\right)=\frac{\overbrace{p\left(\mathbf{y}_1|M_k\right)p\left(M_k\right)}^{p\left(M_k|\mathbf{y}_1\right)}\cdots\overbrace{p\left(\mathbf{y}_i|M_k\right)p\left(M_k\right)}^{p\left(M_k|\mathbf{y}_i\right)}\cdots\overbrace{p\left(\mathbf{y}_H|M_k\right)p\left(M_k\right)}^{p\left(M_k|\mathbf{y}_H\right)}}{\sum_{j=1}^{K}p\left(\mathbf{y}_1|M_j\right)\cdots p\left(\mathbf{y}_H|M_j\right)p\left(M_j\right)}\left(\frac{1}{p\left(M_k\right)}\right)^{H-1} \tag{10}$$

From equation (10) the posterior probability distribution of a specific model M_k, at the federation node is proportional to the product of the probabilities of this model from all local data sites.

In order to select the model at federation node, that best suits the data across all local sites, we made use of the Group Bayes Factor function as has been proposed by other authors (Stephan et al., 2007):

$$BF_{ij}=\frac{p\left(M_i|\mathbf{y}\right)}{p\left(M_j|\mathbf{y}\right)} \tag{11}$$

Assuming uniform model priors $p\left(M_1\right)=p\left(M_2\right)=\ldots=p\left(M_K\right)=cte$ equation (11) turns:

$$BF_{ij}=\frac{p\left(M_i|\mathbf{y}\right)}{p\left(M_j|\mathbf{y}\right)}=\prod_{h=1}^{H}\frac{p\left(M_i|\mathbf{y}_h\right)}{p\left(M_j|\mathbf{y}_h\right)}=\prod_{h=1}^{H}\frac{p\left(\mathbf{y}_h|M_i\right)}{p\left(\mathbf{y}_h|M_j\right)}=\prod_{h=1}^{H}BF_{ij}^{(h)} \tag{12}$$

where $BF_{ij}^{(h)}$ is the Bayes factor for the local data site '*h*':

$$BF_{ij}^{(h)}=\frac{p\left(\mathbf{y}_h|M_i\right)}{p\left(\mathbf{y}_h|M_j\right)} \tag{13}$$

Group Bayes factor encodes the relative probability that the data were generated by one model relative to another assuming that all local data were produced by the same model. It's assumed in practice that for $BF_{ij}>20$ provides strong evidence in favor of model M_i over M_j.

2.2.2 Model selection: Random Effects (RFX) Analysis

In this subsection we assume the local sites use different models to fit the data that defines a Random Effect (RFX) analysis. RFX accounts for the presence of different underlying mechanisms that explain the data. This approach has been proven (Penny et al., 2010; Stephan et al., 2009) to outperform the classical FFX since its markedly more robustness in the presence of outlier data.

We adopted the methodology developed in Stephan et al. (Stephan et al., 2009) and later extended for model averaging (for family of models) by Penny et. al. (Penny et al., 2010). The proposed RFX approach determines the probability density from which the models that generate the data are sampled at different local sites.

Given the '*K*' candidate models, we define a labeling process binary variable or model switch that helps us to calculate the probability that a specific hospital data was generated by the model M_k. Hence, matrix S is defined as:

$$\mathbf{S}=\begin{bmatrix}s_{11} & \cdots & s_{1K}\\ \vdots & \ddots & \vdots\\ s_{H1} & \cdots & s_{HK}\end{bmatrix} \tag{14}$$

where '*K*' is the number of models and '*H*' number of data sites.

$$p(\mathbf{S}) = \prod_{j=1}^{H} p(\mathbf{s}_j) \tag{15}$$

The row vector $\mathbf{s}_j = [s_{j1}, \ldots, s_{jK}]$ expresses the indicator variable to pick one of the '*K*' models for the hospital '*j*'. We assume a priori that the distribution of this variable is independent across local sites, and $p(\mathbf{s}_j)$ is in general a Multinomial distribution so that:

$$p(\mathbf{s}_j) = Mult(\pi_{kj}) \propto \prod_{k=1}^{K} \pi_{jk}^{s_{jk}} \tag{16}$$

In our case $\sum_{k=1}^{K} s_{jk} = 1$ that guarantees that only one model will be selected reducing the prior to a Categorical distribution. The element π_{jk} is the prior probability the data in hospital '*j*' was generated by the model '*k*'. Therefore the prior probability of **S** is finally expressed as:

$$p(\mathbf{S}) = \prod_{j=1}^{H} p(\mathbf{s}_j) = \prod_{j=1}^{H} Mult(\pi_{kj}) \propto \prod_{j=1}^{H} \prod_{k=1}^{K} \pi_{jk}^{s_{jk}} \tag{17}$$

The model frequency variable represented by $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ expresses the frequency of each of the '*K*' models, where $\sum_{k=1}^{K} \pi_k = 1$. We assume that $p(\boldsymbol{\pi})$ is a Dirichlet distribution so that:

$$p(\boldsymbol{\pi}) = Dir(\boldsymbol{\lambda}) \tag{18}$$

with $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]$. If a symmetric Dirichlet distribution is assumed, then $\lambda_1 = \lambda_2 = \lambda_3 \ldots = \lambda_K = \lambda_0$, and:

$$p(\boldsymbol{\pi}) = \frac{\Gamma(K\lambda_0)}{\Gamma(\lambda_0)^K} \prod_{k=1}^{K} \pi_k^{\lambda_0 - 1} \tag{19}$$

The posterior probability $p(\boldsymbol{\pi}|\mathbf{y}, \lambda)$ provides the information required to perform model selection or model averaging at the federation node, accounting for model variability across local sites.

We used the Variational Bayes approach (see later in next subsections) to approximate $p(\boldsymbol{\pi}|\mathbf{y}, \lambda)$, which is based on the model evidence shown in equation (9). Further details about the derivation of these equations, and an algorithm implementation for estimating the parameters of interest, can be found in (Stephan et al., 2009). Hereafter, only the final equations will be provided.

For model indicator variable **s**:

$$q(\mathbf{s}_l) = Mult(\tilde{\gamma}_m^l) \propto \prod_{m=1}^{K} (\tilde{\gamma}_m^l)^{s_{lm}} \tag{20}$$

$\tilde{\gamma}_m^l$ is the mean of the multinomial distribution. This is the expected number of times the model '*m*' is responsible of generating data '*l*'. The expression for $\tilde{\gamma}_m^l$ is the following:

$$\tilde{\gamma}_m^l = \tilde{\pi}_m \cdot \exp\left[\log\left(p(y_l | s_{lm})\right)\right] \tag{21}$$

where $\log\tilde{\pi}_m = \Psi(\lambda_m) - \Psi\left(\sum_{k=1}^{K}\lambda_k\right)$, $\Psi(\cdot)$ is the digamma function. Finally, $\tilde{\gamma}_m^l$ is normalized to obtain γ_m^l that is used in the rest of parameter estimators:

$$\gamma_m^l = \frac{\tilde{\gamma}_m^l}{\sum_{k=1}^{K}\tilde{\gamma}_k^l} \tag{22}$$

For model frequency $\boldsymbol{\pi}$:

The posterior $p(\boldsymbol{\pi}|\mathbf{y},\lambda)$ is approximated by $q(\boldsymbol{\pi}|\mathbf{y})$, which takes the form of a Dirichlet distribution:

$$q(\boldsymbol{\pi}) = Dir(\hat{\boldsymbol{\lambda}}) \tag{23}$$

where $\hat{\boldsymbol{\lambda}} = \left[\hat{\lambda}_1,\ldots,\hat{\lambda}_K\right]$, with $\hat{\lambda}_k$ defined as:

$$\hat{\lambda}_k = \sum_{j=1}^{H}\gamma_k^j + \lambda_0 \tag{24}$$

Then $\hat{\boldsymbol{\pi}} = \left[\hat{\pi}_1,\ldots,\hat{\pi}_K\right]$, with:

$$\hat{\pi}_k = E(\pi_k) = \int q(\pi_k)\,\pi_k d\pi_k = \frac{\hat{\lambda}_k}{\sum_{i=1}^{K}\hat{\lambda}_i} \tag{25}$$

The model that best explains the overall data is defined as that with highest probability $\hat{\boldsymbol{\pi}}$, according to equation (25). The probability of model '*k*', in a RFX analysis at the federation node, is expressed by:

$$p(M_k|\mathbf{y}) = \hat{\pi}_k \tag{26}$$

The most probable model is obtained by:

$$M_i = \max_k\left[p(M_1|\mathbf{y}),\ldots,p(M_k|\mathbf{y}),\ldots,p(M_K|\mathbf{y})\right] \tag{27}$$

Having estimates of the model probabilities accounting for all model evidences and occurrences (at the federation node) allows us to compare two models, rank them or compare families of models. The last use-case is especially important when comparing model families of different natures.

2.2.3 Model averaging

For models of a similar nature, the marginal distribution of the parameters $\boldsymbol{\Theta}$ that accounts for the uncertainty of selecting a specific model at the federation node is given by:

$$p(\boldsymbol{\Theta}|\mathbf{y}) = \sum_{k=1}^{K} p(\boldsymbol{\Theta}|\mathbf{y},M_k)\,p(M_k|\mathbf{y}) \tag{28}$$

The posterior probability of model *k* at the federation node $p(M_k|\mathbf{y})$ may be calculated using a FFX - equation (10) - or RFX - equation (26) - analysis. Posterior probabilities of the parameters $p(\boldsymbol{\Theta}|\mathbf{y},M_k)$ for a specific model would be provided either by the equation (5) or (7).

3. Multiple Linear regression (MLR) for distributed data: probabilistic generative model.

In this subsection we apply the previous general approach to the special case of the multiple linear regression model (MLR). It allows comparing different MLRs as considering subsets of explanatory variables that best explain the dependent data. This approach accounts for distributed data where only aggregates can be passed to the federation node to ultimately estimate the distribution of the parameters that best explained the data across local data sites.

Multiple linear regression allows to study the relationship between a dependent variable **y** and the explanatory variables (or independent variables) denoted by **A**. The general problem is to solve the standard linear equation posed as following:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{29}$$

y: vector of dependent variable $p \times 1$; $\mathbf{A} = [A_1, \ldots, A_N]$: matrix of regressors or design matrix so that $A_i = [a_{1i}, \ldots, a_{Hi}]^T$, $p \times N$; $\boldsymbol{\beta}$: vector of regression coefficients $N \times 1$; $\boldsymbol{\varepsilon}$: vector of experimental noise/error, $p \times 1$.

We have the special situation in the MIP that data **y** and **A** are naturally split and distributed over *H* different sites (hospitals). Additionally the data cannot be fetched to a federation site to solve the equation (1). Thus we have to adapt classical methods for estimating $\boldsymbol{\beta}$, which requires combining aggregates from hospitals data.

We propose to use the Variational Bayes approach (VB) in order to solve MLR equation (29) at each local site. VB maximizes the negative 'Free energy' (NFE) as a surrogate measure of the log evidence or marginal log likelihood. Using either parallel or streaming paradigms, the MLR model parameters distributions and model evidences are properly combined in the federation node to provide global model parameters distributions, which account for the heterogeneity of data across local data sites.

3.1 MLR equations for local sites (Hospitals): Hierarchical Bayesian Model

The MLR equation (29) is solved at each local data site. Therefore we omitted the sub index referring to a specific data site. The equations derived in this subsection are applied in the same way for each local site. In the following a VB formalism will be developed for MLR, partially treated by previous authors (Tzikas et al., 2008). The variables' dependences of our MLR formalism is represented via the Directed Acyclic graph (DAG) shown in Figure 3.

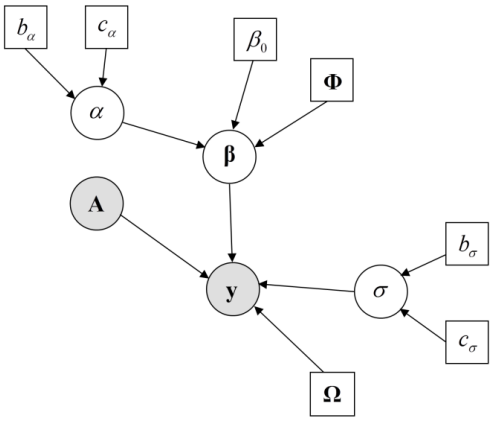

Figure 3. Directed Acyclic graph (DAG) of the MLR. The variables shown in squares are defined a priori. Shaded circles represent observed data whereas not shaded circles represent the model parameters (random variables) to be estimated.

The joint distribution of the data and model parameters for a specific model M_k is given by:

$$p\left(\mathbf{y},\boldsymbol{\beta},\sigma,\alpha\middle|M_k\right)=p\left(\mathbf{y}\middle|\boldsymbol{\beta},\sigma,\boldsymbol{\Omega},M_k\right)\,p\left(\boldsymbol{\beta}\middle|\alpha,\boldsymbol{\Phi},M_k\right)\,p(\alpha)\,p(\sigma) \tag{30}$$

Where the vector of model parameters is $\boldsymbol{\Theta}=[\boldsymbol{\beta},\sigma,\alpha]$; so that σ and α are the data and linear coefficients $\boldsymbol{\beta}$ precisions respectively.

Table 1. Example of the MLR definition for three explanatory variables case. The ones in the table indicate that specific variables are included in the model. In this case we have *N*=3 variables with *K*=7 possible models.

	β_1	β_2	β_3	*MLR Equation*
M_1	0	0	1	$y=A_3\ \beta_3$
M_2	0	1	0	$y=A_2\ \beta_2$
M_3	0	1	1	$y=\begin{bmatrix}A_2 & A_3\end{bmatrix}\begin{bmatrix}\beta_2\\ \beta_3\end{bmatrix}$
M_4	1	0	0	$y=A_1\ \beta_1$
M_5	1	0	1	$y=\begin{bmatrix}A_1 & A_3\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_3\end{bmatrix}$
M_6	1	1	0	$y=\begin{bmatrix}A_1 & A_2\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$
M_7	1	1	1	$y=\begin{bmatrix}A_1 & A_2 & A_3\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}$

A specific model M_k is defined as a subset of explanatory variables. Among models the full set of variables is included whereas the absence of any explanatory variable (null model) is excluded. Thus the maximum number of models is $K=2^N-1$. Hence, matrix A changes the number of columns for each model. Table 1 shows an example of *N*=3 explanatory variables and the seven possible models.

3.2 Likelihood

We assumed **y** is normally distributed as follows:

$$p\left(\mathbf{y}\middle|\boldsymbol{\beta},\sigma,\boldsymbol{\Omega},M_k\right)=N\left(\mathbf{A}\,\boldsymbol{\beta},\sigma^{-1}\boldsymbol{\Omega}\right)=\frac{\sigma^{p/2}}{(2\pi)^{p/2}\left|\boldsymbol{\Omega}\right|^{1/2}}e^{-\frac{\sigma}{2}(\mathbf{y}-\mathbf{A}\,\boldsymbol{\beta})^T\,\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{A}\,\boldsymbol{\beta})} \tag{31}$$

The experimental error (noise) is considered to have a normal distribution so that:

$$p(\boldsymbol{\varepsilon})=N\left(\mathbf{0},\sigma^{-1}\boldsymbol{\Omega}\right) \tag{32}$$

where σ is the error precision; $\boldsymbol{\Omega}$ is the covariance matrix modeling the correlation structure present in our data.

3.3 Prior distributions

This subsection provides the mathematical form of the prior probability distributions assumed for the parameters of interest.

Regression coefficients $\boldsymbol{\beta}$:

$$p\left(\boldsymbol{\beta}\middle|\alpha,\boldsymbol{\Phi}\right)=N\left(\boldsymbol{\beta}_0,\alpha^{-1}\boldsymbol{\Phi}\right)=\frac{\alpha^{N/2}}{\left(2\pi\right)^{N/2}\left|\boldsymbol{\Phi}\right|^{1/2}}e^{-\frac{\alpha}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)^T\boldsymbol{\Phi}^{-1}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_0\right)} \tag{33}$$

Precision of the regression coefficients α *:*

$$p\left(\alpha\right)=Ga\left(b_\alpha,c_\alpha\right)=\frac{c_\alpha^{b_\alpha}}{\Gamma\left(b_\alpha\right)}\alpha^{b_\alpha-1}e^{-c_\alpha\alpha} \tag{34}$$

Precision of the likelihood function σ *:*

$$p\left(\sigma\right)=Ga\left(b_\sigma,c_\sigma\right)=\frac{c_\sigma^{b_\sigma}}{\Gamma\left(b_\sigma\right)}\sigma^{b_\sigma-1}e^{-c_\sigma\sigma} \tag{35}$$

3.4 Posteriors

Our aim is to estimate the posterior distributions of the unknown parameters $\boldsymbol{\Theta}=\left[\boldsymbol{\beta},\sigma,\alpha\right]$ and the evidence of the different models $M_1,\ldots,M_K$. In doing so, we use the *Variational Bayes* formalism (VB) (Beal, 2003; Lappalainen and Miskin, 2000). This is an efficient method already used in previous publications (Penny et al., 2005; Trujillo-Barreto et al., 2008). It is based on the Variational Free Energy method of Feynman and Bogoliubov, which provides an approximate factorized posterior distribution $q\left(\boldsymbol{\Theta}\middle|\mathbf{Y}\right)$, with minimal Kullback–Leibler (KL) divergence from the true posterior $p\left(\boldsymbol{\Theta}\middle|\mathbf{y}\right)=p\left(\boldsymbol{\beta},\sigma,\alpha\middle|\mathbf{y}\right)$.

We consider the following factorization of the approximate posterior distribution:

$$q\left(\boldsymbol{\beta},\sigma,\alpha\middle|\mathbf{y}\right)=q\left(\boldsymbol{\beta}\middle|\mathbf{y}\right)\ q\left(\sigma\middle|\mathbf{y}\right)q\left(\alpha\middle|\mathbf{y}\right) \tag{36}$$

The following subsections provide details about the derivation of the approximate posteriors for each parameter. In general, the approximate posterior $q\left(\boldsymbol{\Theta}\middle|\mathbf{Y}\right)$ for the *i*-th subset of parameters is expressed by $q\left(\theta_i\middle|\mathbf{Y}\right)$ as:

$$\begin{aligned}q\left(\theta_i\middle|\mathbf{Y}\right)&=\frac{\exp\left(\int q\left(\boldsymbol{\Theta}_{\setminus i}\middle|\mathbf{Y}\right)\log p\left(\mathbf{Y},\boldsymbol{\Theta}\right)d\boldsymbol{\Theta}_{\setminus i}\right)}{\int\exp\left(\int q\left(\boldsymbol{\Theta}_{\setminus i}\middle|\mathbf{Y}\right)\log p\left(\mathbf{Y},\boldsymbol{\Theta}\right)d\boldsymbol{\Theta}_{\setminus i}\right)d\theta_i}\\&=\frac{\exp\left(I\left(\theta_i\right)\right)}{\int\exp\left(I\left(\theta_i\right)\right)d\theta_i}\end{aligned} \tag{37}$$

where $I\left(\theta_i\right)=\int q\left(\boldsymbol{\Theta}_{\setminus i}\middle|\mathbf{Y}\right)\log p\left(\mathbf{Y},\boldsymbol{\Theta}\right)d\boldsymbol{\Theta}_{\setminus i}=\left\langle\log p\left(\mathbf{Y},\boldsymbol{\Theta}\right)\right\rangle_{q\left(\boldsymbol{\Theta}_{\setminus i}\right)}$, and $\boldsymbol{\Theta}_{\setminus i}$ indicates parameters not present in the *i*-th group of parameters.

3.4.1 Posterior for regression coefficients $\boldsymbol{\beta}$

The approximated posterior distribution $q(\boldsymbol{\beta}|\mathbf{y})$ of the regression coefficients $\boldsymbol{\beta}$ is expressed by:

$$q(\boldsymbol{\beta}|\mathbf{y}) \propto e^{I(\boldsymbol{\beta})} \tag{38}$$

where

$$I(\boldsymbol{\beta}) = \int q(\sigma|\mathbf{y}) q(\alpha|\mathbf{y}) \log\left[p(\mathbf{y}|\boldsymbol{\beta},\sigma)\, p(\boldsymbol{\beta}|\alpha)\, p(\sigma)\, p(\alpha) \right] d\sigma d\alpha \tag{39}$$

Finally $q(\boldsymbol{\beta}|\mathbf{y})$ is a normal distribution so that,

$$q(\boldsymbol{\beta}|\mathbf{y}) = N\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}\right) = \frac{1}{(2\pi)^{N/2} \left|\hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}\right|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})} \tag{40}$$

The mean and covariance matrix of this multivariate normal distribution is given by:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}^{-1} \left(\hat{\sigma} \mathbf{A}^T \boldsymbol{\Omega}^{-1} \mathbf{y} + \hat{\alpha} \boldsymbol{\Phi}^{-1} \boldsymbol{\beta}_0 \right) \tag{41}$$

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}} = \hat{\sigma} \mathbf{A}^T \boldsymbol{\Omega}^{-1} \mathbf{A} + \hat{\alpha} \boldsymbol{\Phi} \tag{42}$$

3.4.2 Posterior for the precision of the regression coefficients α

In this case the approximated posterior distribution for α is obtained by:

$$q(\alpha|\mathbf{y}) \propto e^{I(\alpha)} \tag{43}$$

where

$$I(\alpha) = \int q(\boldsymbol{\beta}|\mathbf{y}) q(\sigma|\mathbf{y}) \log\left[p(\mathbf{y}|\boldsymbol{\beta},\sigma)\, p(\boldsymbol{\beta}|\alpha)\, p(\sigma)\, p(\alpha) \right] d\boldsymbol{\beta} d\sigma \tag{44}$$

After some algebra solving equation (44) we see that the posterior distribution of α is approximated by a gamma distribution so that:

$$q(\alpha|\mathbf{y}) = Ga\left(\hat{b}_\alpha, \hat{c}_\alpha\right) \tag{45}$$

where

$$\begin{aligned} \hat{b}_\alpha &= b_\alpha + \frac{N}{2} - 1 \\ \hat{c}_\alpha &= c_\alpha + \frac{1}{2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^T \boldsymbol{\Phi}^{-1} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + Tr\left(\boldsymbol{\Phi}^{-1} \hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}\right) \end{aligned} \tag{46}$$

Finally, the expected value of α is:

$$\hat{\alpha} = \frac{\hat{b}_\alpha}{\hat{c}_\alpha} \tag{47}$$

3.4.3 Posterior for the precision σ

The approximate posterior distribution for σ is obtained by:

$$q(\sigma|\mathbf{y}) \propto e^{I(\sigma)} \tag{48}$$

where

$$I(\sigma) = \int q(\boldsymbol{\beta}|\mathbf{y}) q(\alpha|\mathbf{y}) \log\left[p(\mathbf{y}|\boldsymbol{\beta},\sigma)\, p(\boldsymbol{\beta}|\alpha)\, p(\sigma)\, p(\alpha) \right] d\boldsymbol{\beta} d\alpha \tag{49}$$

After some algebra, we obtain the following gamma distribution:

$$q(\sigma|\mathbf{y}) = Ga\left(\hat{b}_\sigma, \hat{c}_\sigma\right) \tag{50}$$

where

$$\begin{aligned} \hat{b}_\sigma &= b_\sigma + \frac{p}{2} - 1 \\ \hat{c}_\sigma &= c_\sigma + \frac{1}{2}\left(\mathbf{y} - \mathbf{A}\,\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{\Omega}^{-1}\left(\mathbf{y} - \mathbf{A}\,\hat{\boldsymbol{\beta}}\right) + Tr\left(\mathbf{A}^T \boldsymbol{\Omega}^{-1} \mathbf{A} \hat{\boldsymbol{\Phi}}_\beta\right) \end{aligned} \tag{51}$$

The expected value of σ is given by:

$$\hat{\sigma} = \frac{\hat{b}_\sigma}{\hat{c}_\sigma} \tag{52}$$

3.5 Free Energy and KL divergences

The free energy function is expressed as:

$$F = \int q(\boldsymbol{\Theta}|\mathbf{y}) \log p(\mathbf{y}|\boldsymbol{\Theta})\, d\boldsymbol{\Theta} - \int q(\boldsymbol{\Theta}|\mathbf{y}) \log \frac{q(\boldsymbol{\Theta}|\mathbf{y})}{p(\boldsymbol{\Theta})} d\boldsymbol{\Theta} \tag{53}$$

where the first term $\left\langle \log p(\mathbf{y}|\boldsymbol{\Theta}) \right\rangle_{q(\boldsymbol{\Theta}|\mathbf{y})} = \int q(\boldsymbol{\Theta}|\mathbf{y}) \log p(\mathbf{y}|\boldsymbol{\Theta})\, d\boldsymbol{\Theta}$ is the mean likelihood over $q(\boldsymbol{\Theta}|\mathbf{y})$ and $KL\left(q(\boldsymbol{\Theta}|\mathbf{y}) \| p(\boldsymbol{\Theta})\right) = \int q(\boldsymbol{\Theta}|\mathbf{y}) \log \frac{q(\boldsymbol{\Theta}|\mathbf{y})}{p(\boldsymbol{\Theta})} d\boldsymbol{\Theta}$ the Kullback Leibler divergence between the approximated posterior and prior distributions.

After some algebraic work, the expectation of the Likelihood is given by:

$$\begin{aligned} \left\langle \log p(\mathbf{y}|\boldsymbol{\Theta}) \right\rangle_{q(\boldsymbol{\Theta}|\mathbf{y})} = {} & \frac{N}{2}\left(\Psi\left(\hat{b}_\sigma\right) - \ln \hat{c}_\sigma\right) - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Omega}| - \\ & - \frac{\hat{\sigma}}{2}\left(\mathbf{y} - \mathbf{A}\,\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{\Omega}^{-1}\left(\mathbf{y} - \mathbf{A}\,\hat{\boldsymbol{\beta}}\right) - \frac{\hat{\sigma}}{2} Tr\left(\mathbf{A}^T \boldsymbol{\Omega}^{-1} \mathbf{A} \hat{\boldsymbol{\Phi}}_\beta\right) \end{aligned} \tag{54}$$

where $Tr(\cdot)$ is the trace operator.

Similarly, the Kullback Leibler divergence between the approximated posterior and prior distributions is expressed as:

$$KL\left(q\left(\boldsymbol{\Theta}|\mathbf{y}\right)\middle\|p\left(\boldsymbol{\Theta}\right)\right)=\int q\left(\boldsymbol{\beta}|\mathbf{y}\right)q\left(\sigma|\mathbf{y}\right)q\left(\alpha|\mathbf{y}\right)\log\frac{q\left(\boldsymbol{\beta}|\mathbf{y}\right)q\left(\sigma|\mathbf{y}\right)q\left(\alpha|\mathbf{y}\right)}{p\left(\boldsymbol{\beta}|\alpha\right)p\left(\sigma\right)p\left(\alpha\right)}$$
$$=KL\left(q\left(\boldsymbol{\beta}|\mathbf{y}\right)\middle\|p\left(\boldsymbol{\beta}|\alpha\right)\right)+KL\left(q\left(\sigma|\mathbf{y}\right)\middle\|p\left(\sigma\right)\right)+KL\left(q\left(\alpha|\mathbf{y}\right)\middle\|p\left(\alpha\right)\right) \tag{55}$$

After some algebraic operations we obtain:

$$KL\left(q\left(\boldsymbol{\beta}|\mathbf{y}\right)\middle\|p\left(\boldsymbol{\beta}|\alpha\right)\right)=-\frac{1}{2}\log\frac{\left|\hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}\right|}{\left|\boldsymbol{\Phi}\right|}-\frac{N}{2}-\frac{N}{2}\left(\Psi\left(\hat{b}_{\alpha}\right)-\log\hat{c}_{\alpha}\right)+$$
$$+\frac{\hat{\alpha}}{2}Tr\left(\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}\right)+\frac{\hat{\alpha}}{2}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right)^{T}\boldsymbol{\Phi}^{-1}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right) \tag{56}$$

$$KL\left(q\left(\sigma|\mathbf{y}\right)\middle\|p\left(\sigma\right)\right)=\left(\hat{b}_{\sigma}-1\right)\Psi\left(\hat{b}_{\sigma}\right)+\log\hat{c}_{\sigma}-\hat{b}_{\sigma}-\log\Gamma\left(\hat{b}_{\sigma}\right)+\log\Gamma\left(b_{\sigma}\right)-$$
$$-b_{\sigma}\log\left(c_{\sigma}\right)-\left(b_{\sigma}-1\right)\left(\Psi\left(\hat{b}_{\sigma}\right)-\log\hat{c}_{\sigma}\right)+\frac{\hat{b}_{\sigma}c_{\sigma}}{\hat{c}_{\sigma}} \tag{57}$$

$$KL\left(q\left(\alpha|\mathbf{y}\right)\middle\|p\left(\alpha\right)\right)=\left(\hat{b}_{\alpha}-1\right)\Psi\left(\hat{b}_{\alpha}\right)+\log\hat{c}_{\alpha}-\hat{b}_{\alpha}-\log\Gamma\left(\hat{b}_{\alpha}\right)+\log\Gamma\left(b_{\alpha}\right)-$$
$$-b_{\alpha}\log\left(c_{\alpha}\right)-\left(b_{\alpha}-1\right)\left(\Psi\left(\hat{b}_{\alpha}\right)-\log\hat{c}_{\alpha}\right)+\frac{\hat{b}_{\alpha}c_{\alpha}}{\hat{c}_{\alpha}} \tag{58}$$

Equations (54), (56), (57) and (58) provides the necessary quantities to evaluate the Free energy function. The dependence among the posteriors of the model parameters $\boldsymbol{\Theta}=\left[\boldsymbol{\beta},\sigma,\alpha\right]$ in equations (41), (46) and (51) leads to an iterative algorithm. The Free Energy is evaluated at each iteration, and is expected to continue to increase until the convergence criterion is met. This is defined as $\frac{F_i-F_{i-1}}{F_{i-1}}<Tol$, where F_i and F_{i-1} are the free energy at iteration '*i*' and '*i*-1' respectively, *Tol* is the tolerance with values defined in $0\le Tol\le 1$.

Since the Free Energy is a surrogate measure of the log model evidence, the model probabilities - equation (8) - at local nodes, are given by the expression:

$$p\left(M_k|\mathbf{y}\right)=\frac{e^{F\left(\mathbf{y}|M_k\right)}p\left(M_k\right)}{\sum_{i=1}^{K}e^{F\left(\mathbf{y}|M_i\right)}p\left(M_i\right)} \tag{59}$$

where the model evidence is expressed by:

$$p\left(\mathbf{y}|M_k\right)\approx e^{F\left(\mathbf{y}|M_k\right)} \tag{60}$$

As a final step, the model evidences and the distributions of the parameters $\boldsymbol{\Theta}=\left[\boldsymbol{\beta},\sigma,\alpha\right]$ from each local node are submitted to the federation node to be properly federated.

The general iterative algorithm that runs at each local node is the following:

Local Node –General Algorithm -

- Local data **y** and **A** assessment.
- Definition of the stopping criteria tolerance '*Tol*'
- Definition of $\boldsymbol{\Omega}$, $\boldsymbol{\Phi}$ as the covariance structure for **y** and $\boldsymbol{\beta}$ respectively.
- Initialize hyperparameters: $b_\sigma, c_\sigma, b_\alpha, c_\alpha$
- For *k*=1 to *K* (Number of Models)
  - Iteration '*i*'
    - Compute $\hat{\alpha}$ using equation (47)
    - Compute $\hat{\sigma}$ using equation (52)
    - Compute $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\beta}}$ using equations (41) and (42) respectively
    - Compute Free Energy F_i for iteration '*i*' combining equations (54), (56), (57) and (58)

      Repeat until $\frac{F_i - F_{i-1}}{F_{i-1}} < Tol$
  - Evaluate evidence of model M_k using equation (60)

  End (Number of Models)
- Send to the Federation node: evidences of the '*K*' models and distributions of the parameters $\hat{\boldsymbol{\Theta}} = \left[\hat{\boldsymbol{\beta}}, \hat{\sigma}, \hat{\alpha}\right]$

3.6 MLR equations for the federation node

At the federation node, the *MLR* is estimated accounting for parameter distributions and model evidences coming from local nodes. In order to obtain expressions (see next subsection) for the federated *MLR* model parameters we made use of the general equations developed in all subsections 2.x.

3.6.1 Linear model coefficients $\boldsymbol{\beta}$:

At each local node, this parameter has a normal distribution given by equations (40), (41) and (42). The distribution of $\boldsymbol{\beta}$ at the federation node using equation (5) is given by:

$$p\left(\boldsymbol{\beta}\middle|\mathbf{y}, M_k\right) = \frac{1}{\left[p\left(\boldsymbol{\beta}\middle|\alpha\right)\right]^{H-1}} \prod_{j=1}^{H} q\left(\boldsymbol{\beta}\middle|\mathbf{y}_j, M_k\right) \tag{61}$$

$$p\left(\boldsymbol{\beta}\middle|\mathbf{y}, M_r\right) = N\left(\hat{\boldsymbol{\beta}}_f, \hat{\boldsymbol{\Phi}}_f\right) = \frac{1}{\left(2\pi\right)^{N/2}\left|\hat{\boldsymbol{\Phi}}_f\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_f\right)^T \hat{\boldsymbol{\Phi}}_f^{-1}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_f\right)\right] \tag{62}$$

where,

$$\hat{\boldsymbol{\beta}}_f = \hat{\boldsymbol{\Phi}}_f^{-1}\left(\sum_{i=1}^{H} \boldsymbol{\Phi}_{\boldsymbol{\beta}_i}^{-1}\hat{\boldsymbol{\beta}}_i - \hat{\alpha}_f\left(H-1\right)\boldsymbol{\Phi}^{-1}\boldsymbol{\beta}_0\right) \tag{63}$$

$$\hat{\boldsymbol{\Phi}}_f = \sum_{i=1}^{M} \boldsymbol{\Phi}_{\boldsymbol{\beta}_i}^{-1} - \hat{\alpha}_f\left(M-1\right)\boldsymbol{\Phi}^{-1} \tag{64}$$

3.6.2 Precision of the regression coefficients α

At each local node, α follows a Gamma distribution given by equations (45), (46) and (47). After some algebraic operations on equation (5) we see that α also distributes as gamma at the federation node so that:

$$p\left(\alpha\middle|\mathbf{y},M_k\right)=\frac{1}{\left[p(\alpha)\right]^{H-1}}\prod_{j=1}^{H}q\left(\alpha\middle|\mathbf{y}_j,M_k\right) \tag{65}$$

$$p\left(\alpha\middle|\mathbf{y},M_k\right)=Ga\left(b_{\alpha_f},c_{\alpha_f}\right)=\frac{c_{\alpha_f}^{b_{\alpha_f}}}{\Gamma\left(b_{\alpha_f}\right)}\alpha_f^{b_{\alpha_f}-1}e^{-c_{\alpha_f}\alpha_f} \tag{66}$$

$$\begin{aligned} b_{\alpha_f}&=\sum_{j=1}^{H}\hat{b}_{\alpha_j}-(H-1)b_{\alpha}\\ c_{\alpha_f}&=\sum_{j=1}^{H}\hat{c}_{\alpha_j}-(H-1)c_{\alpha}\end{aligned} \tag{67}$$

The expected value for α is expressed by:

$$\hat{\alpha}_f=\frac{b_{\alpha_f}}{c_{\alpha_f}} \tag{68}$$

3.6.3 Precision of the likelihood function σ

At each local node, σ has a Gamma distribution given by equations (50), (51) and (52). After combining the local estimates through equation (5) this model parameter distributes as gamma at the federation node. This probability distribution is expressed as:

$$p\left(\sigma\middle|\mathbf{y},M_k\right)=\frac{1}{\left[p(\sigma)\right]^{H-1}}\prod_{j=1}^{H}q\left(\sigma\middle|\mathbf{y}_j,M_k\right) \tag{69}$$

$$p\left(\sigma\middle|\mathbf{y},M_k\right)=Ga\left(b_{\sigma_f},c_{\sigma_f}\right)=\frac{c_{\sigma_f}^{b_{\sigma_f}}}{\Gamma\left(b_{\sigma_f}\right)}\sigma^{b_{\sigma_f}-1}e^{-c_{\sigma_f}\sigma} \tag{70}$$

$$\begin{aligned} b_{\sigma_f}&=\sum_{j=1}^{H}\hat{b}_{\sigma_j}-(H-1)b_{\sigma}\\ c_{\sigma_f}&=\sum_{j=1}^{H}\hat{c}_{\sigma_j}-(H-1)c_{\sigma}\end{aligned} \tag{71}$$

with the expected value given by:

$$\hat{\sigma}_f=\frac{b_{\sigma_f}}{c_{\sigma_f}} \tag{72}$$

3.6.4 Model selection and averaging for MLR at the federation node

As we mentioned before, a specific model M_k in the MLR is defined as a subset of explanatory variables (see Table 1). This means we are looking for a subset of β_i in $\boldsymbol{\beta}$ that best contributes explaining the data across hospital sites. The number of possible models 'K' is $K=2^N-1$, taking into account that each

explanatory variable can be present or not. The model selection at the federation node is conducted by using either a FFX or RFX approach given by equation (10) or (26) respectively. The model averaging step, via expression (28), provides the estimation of the parameter distributions irrespective of the model selected. For the regression coefficients this equation takes the mathematical form:

$$p(\boldsymbol{\beta}|\mathbf{y}) = \sum_{k=1}^{K} p(\boldsymbol{\beta}|\mathbf{y}, M_k)\, p(M_k|\mathbf{y}) \tag{73}$$

The general algorithm that runs at the federation node is the following:

Federation Node –General Algorithm -

- Fetch from the '*H*' local nodes the local parameter expectations: $\hat{\boldsymbol{\beta}}_j$, $\boldsymbol{\Phi}_{\boldsymbol{\beta}_j}, \hat{b}_{\alpha_j}, \hat{c}_{\alpha_j}, \hat{b}_{\sigma_j}, \hat{c}_{\sigma_j}$ for all models $M_k, k = 1,\ldots,K$.
- Estimation of the federated expected value $\hat{\alpha}_f$ of the precision of the linear model coefficients using equation (68)
- Estimation of the federated expected value $\hat{\sigma}_f$ of the precision of the data distribution using equation (72).
- Estimation of the federated linear coefficients $\hat{\boldsymbol{\beta}}_f$ and the covariance matrix $\hat{\boldsymbol{\Phi}}_f$ using equations (63) and (64) respectively.
- Execution of the Model selection using a '*Random Effects (RFX) Analysis'* (see the corresponding section in the text).
- Model averaging using equation (73).
- Submit to the user the distributions and its expectations $\hat{\boldsymbol{\beta}}_f$, $\hat{\boldsymbol{\Phi}}_f$, $\hat{\sigma}_f, \hat{\alpha}_f$ for the model parameters and the model that best explains the data (set of explanatory variables that best explains the data).

CONCLUDING REMARKS

In this paper we presented a Bayesian formalism to deal with distributed, highly heterogeneous and big data, based on previous theoretical developments. This formalism is part of the Bayesian computational core embedded in the Medical Informatics Platform (MIP) of the European Human Brain Project (www.humanbrainproject.eu). As an application example we applied the general approach outlined in this work to the multiple linear regression (MLR), which is considered one of the workhorses of statistics in the neuroimaging field. Though the adaptation of the MLR to a distributed environment might be considered straightforward, in this work we extended it with a model selection and averaging approach. This is crucial for selecting appropriate models to explain highly heterogeneous data.

The federation node is defined as the site in which the local aggregates quantities are combined to deliver a final result accounting for the diversity of the different hospital data (local nodes). We provided approximated probability distributions of the MLR parameters using the Variational Bayes formalism (VB). VB approximates the parameters probability distributions based on the minimization of the negative free energy (NFE) as a surrogate measure of the log evidence. NFE enables the estimation of model probabilities to ultimately performing the model selection and averaging step at the federation node. In this respect, we brought into our formalism the methodology proposed by Stephan et al. (Stephan et al., 2009) (for '*Dynamic Causal Modeling*'-DCM-) to account for the heterogeneity of the possible mechanisms that explain data at different hospitals. This is an important aspect in order to integrate highly heterogeneous data coming from different subjects and hospitals across the globe. For the case of a large number of explanatory variables ($N \gg 1$) an exhaustive evaluation of all model evidences is computationally prohibitive. In such cases a Markov Chain Monte Carlo (MCMC) approach ought to be adopted. MCMC would sample the mass of the models probability distribution allowing, as before, model selection and model averaging steps to be conducted. In addition, a factorization over $\boldsymbol{\beta}$ coefficients in the VB scheme -equation (36)- should be assumed at local nodes in order to get tractable estimations.

The implementation of our theoretical formalism and its application, in both simulated and real data, will be addressed in a separate work. Its performance under different conditions and the limits of its application will be evaluated through simulated data. Several situations should be considered: a) unbalanced data across local nodes; b) influence of the number of local nodes on model selection and parameters estimation; c) robustness to the presence of outliers; d) robustness of the estimations when the number of explanatory variables increases, etc.

The present work is the first of a series of papers adapting, through Bayesian formalism, the existing statistical models to distributed and big data in order to ultimately discover underlying mechanisms of brain anatomy and function (using the Medical Informatics Platform). This approach is not limited to the Neuroscience field; it can be applied in other branches of science and technology.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 604102 and the European Union's Horizon 2020 research and innovation programme under grant agreement No 720270 (HBP SGA1). BD is supported by the Swiss National Science Foundation (NCCR Synapsy, project grant Nr 32003B_159780 and SPUM 33CM30_140332/1), Foundation Parkinson Switzerland and Foundation Synapsis.

REFERENCES

Beal, M.J., 2003. Variational algorithms for approximate Bayesian inference. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, London.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A.C., Jordan, M.I., 2013. Streaming Variational Bayes, in: Advances in Neural Information Processing Systems. pp. 1727–1735.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 382–401.

Lappalainen, H., Miskin, J., 2000. Ensemble learning, Advances in Independent Component Analysis.

MacKay, D.J.C., 1992. Bayesian Interpolation. Neural Comput. 4, 415–447. doi:10.1162/neco.1992.4.3.415

Penny, W., Mattout, J., Trujillo-Barreto, N., 2006. Bayesian model selection and averaging. Stat. Parametr. Mapp. Anal. Funct. brain images. London Elsevier.

Penny, W.D., Stephan, K.E., Daunizeau, J., Rosa, M.J., Friston, K.J., Schofield, T.M., Leff, A.P., 2010. Comparing Families of Dynamic Causal Models. PLoS Comput. Biol. 6, e1000709. doi:10.1371/journal.pcbi.1000709

Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. Neuroimage 24, 350–362. doi:10.1016/j.neuroimage.2004.08.034

Stephan, K.E., Marshall, J.C., Penny, W.D., Friston, K.J., Fink, G.R., 2007. Interhemispheric Integration of Visual Processing during Task-Driven Lateralization. J. Neurosci. 27.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., Friston, K.J., 2009. Bayesian model selection for group studies. Neuroimage 46, 1004–1017. doi:10.1016/j.neuroimage.2009.03.025

Trujillo-Barreto, N.J., Aubert-Vázquez, E., Penny, W.D., 2008. Bayesian M/EEG source reconstruction with spatio-temporal priors. Neuroimage 39, 318–335. doi:10.1016/j.neuroimage.2007.07.062

Tzikas, D., Likas, A., Galatsanos, N., 2008. The variational approximation for Bayesian inference. IEEE Signal Process. Mag. 25, 131–146. doi:10.1109/MSP.2008.929620