applied sciences

MDPI

Article

A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data

Albert Belenguer-Llorens [1], Carlos Sevilla-Salcedo [1], Manuel Desco [2,3,4,5], Maria Luisa Soto-Montenegro [3,4,*] and Vanessa Gómez-Verdejo [1,*]

[1] Department of Signal Processing and Communications, University Carlos III of Madrid Leganés, 28911 Leganés , Spain; abelenguer@tsc.uc3m.es (A.B.-L.); casevill@pa.uc3m.es (C.S.-S.)
[2] Department of Bioengineering and Aerospace Engineering, University Carlos III of Madrid Leganés, 28911 Leganés, Spain; desco@hggm.es
[3] CIBER of Mental Health (CIBERSAM), 28029 Madrid, Spain
[4] Instituto de Investigación Sanitaria Gregorio Marañón, 28007 Madrid, Spain
[5] Centro Nacional de Investigaciones Cardiovasculares (CNIC), 28029 Madrid, Spain
* Correspondence: marisa@hggm.es (M.L.S.-M.); vanessag@ing.uc3m.es (V.G.-V.)

Abstract: In this paper, we propose a novel Machine Learning Model based on Bayesian Linear Regression intended to deal with the low sample-to-variable ratio typically found in neuroimaging studies and focusing on mental disorders. The proposed model combines feature selection capabilities with a formulation in the dual space which, in turn, enables efficient work with neuroimaging data. Thus, we have tested the proposed algorithm with real MRI data from an animal model of schizophrenia. The results show that our proposal efficiently predicts the diagnosis and, at the same time, detects regions which clearly match brain areas well-known to be related to schizophrenia.

Keywords: Bayesian learning; neuroimaging; feature selection; kernel formulation; mental disorders; schizophrenia; MRI

check for updates

Citation: Belenguer-Llorens, A.; Sevilla-Salcedo, C.; Desco, M.; Soto-Montenegro, M.L.; Gomez-Verdejo, V. A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data. *Appl. Sci.* **2022**, *12*, 2571. https://doi.org/10.3390/app12052571

Academic Editors: Alexander E. Hramov, Kyungtae Kang, Hyo-Joong Suh and Junggab Son

Received: 7 December 2021
Accepted: 21 February 2022
Published: 1 March 2022

1. Introduction

Neuroimaging has undergone a major breakthrough in recent years and has helped in the diagnosis, prognosis, and treatment monitoring of psychiatric disorders. The clinical diagnosis of these disorders is troublesome due to the lack of specific biomarkers [1] and to the fact that many of them share clinical features, thus hindering an accurate diagnosis. Specifically, schizophrenia is one of the most complex pathologies to diagnose [2] since it is commonly confused with other psychotic disorders in up to 20% of cases [3]. As consequence, new tools for the diagnosis of mental disorders are emerging [4,5].

Machine Learning (ML) techniques have emerged as a promising tool for the analysis of neuroimaging data. These algorithms are capable of analyzing any data source, either images (structural or functional), genetic information [6] or behavioral information [7], to carry out an automatic diagnosis of the pathology. Recent approaches based on Support Vector Machine algorithm (SVM) have been applied in Magnetic Resonance Imaging (MRI), showing great results in this field and detecting relevant brain areas involved in the pathology, as well as inferring new useful biomarkers for their diagnosis [8–10]. However, although these models have provided accurate results for automatic classification, the lack of interpretability in their results prevents the characterization of the pathology. In particular, in contexts where only a few features are relevant for the problem, it is advisable to detect the informative variables and eliminate the useless ones. For this reason, many authors combine ML models with Feature Selection (FS) approaches, such as the Recursive Feature Elimination (RFE) [11], consisting of the direct elimination of the less representative features, methods based on decision tree formulations, such as Random Forest Importance (RFI) [12,13], or embedded approaches which include L1 or

L1–L2 regularizations to promote sparsity, such as Lasso and elastic-net algorithms [14,15]. Nevertheless, in neuroimaging, we have to deal with large datasets, where the number of cases is significantly smaller than the number of variables, and many of these approaches fail in this scenario, tending to over-fit. To avoid this problem, some authors propose Bayesian approaches but work over a reduced set of features [16–18], whereas others point to the use of more refined techniques better adapted to the problem needs [19–21].

To overcome these limitations, we present a novel formulation for the Bayesian Linear Regression model. Our proposal, called the Dual Bayesian Linear regression model with Feature Selection (DBL-FS), is formulated to work efficiently with a reduced number of samples characterized in high-dimensional spaces, e.g., neuroimaging data. For this purpose, the model is formulated in the dual space and simultaneously includes an Automatic Relevance Determination (ARD) prior over the primal weights to provide the model with FS capabilities so that it can remove irrelevant input features. Here, we have tested our formulation on rodent data in an animal model of schizophrenia that show similar brain anatomical deficits than patients with schizophrenia [22–24]. One advantage of using rodent data is a more solid knowledge of the ground truth due to the controlled experimental induction of the pathology.

2. Materials

Rodent MRI data were obtained from the Biomedical Imaging and Instrumentation Group (Biig) of the Gregorio Marañón Hospital. The dataset consisted of 53 rat brain MRI images divided into two groups: healthy rats (N = 24) and pathological rats (N = 29). Pathology was induced by the administration of the viral mimic polyriboinosinic-polyribocytidilic acid (poly I:C) in gestational day 15 to pregnant Wistar rats, since maternal immune stimulation (MIS) is associated with increased risk of onset of schizophrenia in the offspring, with behavioral abnormalities as well as neurophysiological and morphological traits. Model details can be found elsewhere [25–27].

All images were preprocessed following the standard preprocessing pipeline in neuroimaging research, using the processing toolbox of the Statistical Parametric mapping software (SPM12) [28], as shown in Figure 1. Output consisted of: White Matter (WM), Gray Matter (GM), and CerebroSpinal Fluid (CSF) regions, with 464,487, 582,467, and 30,702 voxels, respectively.

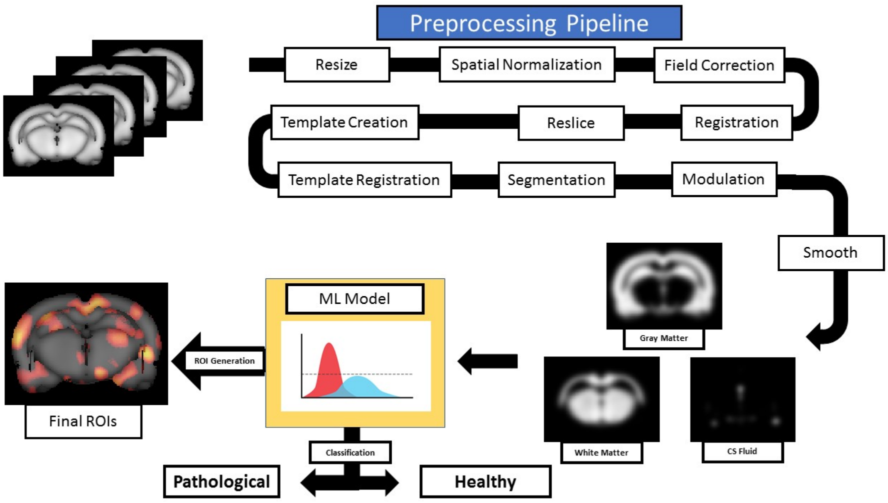

Figure 1. MRI pipeline for data processing [28]. First, images were corrected for field homogeneity, resized by a factor of 10 and spatially normalized to create a custom template based on a Wistar rat brain template [29]. All images were resliced to this custom template and were segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). Later, all images were modulated using the Jacobian determinants and smoothed with a 10-mm FWHM Gaussian kernel. Finally, the segmented tissues were processed by the ML model to classify them into healthy and pathological subjects and to identify the brain areas relevant to this decision.

3. Methods

This section introduces the formulation of the proposed Dual Bayesian Linear regression model with Feature Selection (DBL-FS). Later, we also introduce some reference methods that we will use as baselines to show the advantages of the proposed approach together to the experimental setup.

3.1. A Dual Bayesian Linear Regression Model with Feature Selection

3.1.1. Model Definition

The proposed model borrows some ideas from the Bayesian Principal Component Analysis (BPCA) [30] and Bayesian Canonical Correlation Analysis (BCCA) [31] algorithms to endow a Bayesian Linear Regression (BLR) [32] with a dual formulation able to carry out automatic feature selection over the primal variables. This relies on including an ARD prior over the weight matrices to automatically infer the feature relevance in the input feature space by assigning higher/lower relevance values when there are more/less relevant features. Meanwhile, the model works with a formulation in the dual space. In turn, this allows the model to efficiently deal with large data problems by working in the data space while it applies a feature selection over the variable space. In addition, we can exploit the DBL-FS Bayesian formulation to facilitate including prior expert knowledge to guide the FS process. This way, we can guide the learning process and compensate the limited number of samples to train the model.

To define the model, let us consider $\mathbf{X}$ as the observation matrix with the MRI information of N subjects; this way, each row, $\mathbf{x}_{n,:}$ for $n = 1, \ldots, N$, is a D-dimensional vector containing the brain image of the n-th subject, and each column, $\mathbf{x}_{:,d}$ for $d = 1, \ldots, D$, contains the information of the d-th voxel over the N subjects. On the other hand, the column vector $\mathbf{y}$ represents the diagnosis labels (control or schizophrenic) for the N subjects under study. Although each label, y_n, belongs to the set $\{0, 1\}$ (indicating the subject is control or not), for the model formulation, we consider $y_n \in \Re$, and thus, we will generalize the model for regression problems. Later, we will apply a threshold over the model output to classify each subject into one of two categories.

3.1.2. Generative Model

As the graphical model of Figure 2 shows, the generative model of DBL-FS considers that each datum, $\mathbf{x}_{n,:}$, is combined with a weight vector $\mathbf{w}$ plus some Gaussian noise to generate the output variable:

$$y_n = \mathbf{x}_{n,:}\mathbf{w} + \eta, \tag{1}$$

where η is a Gaussian noise with zero mean and precision τ. In turn, the noise precision is modeled with a gamma distribution with parameters a_0^τ, b_0^τ:

$$\tau \sim \Gamma(a_0^\tau, b_0^\tau) \tag{2}$$

In addition, DBL-FS considers that the weight associated to the d-th input feature follows a normal distribution:

$$w_d \sim \mathcal{N}\left(0,\ \alpha_d^{-1}\right) \quad d = 1, \ldots, D \tag{3}$$

where its precision, α_d, is modeled with a gamma distribution as:

$$\alpha_d \sim \Gamma(a_0^\alpha, b_0^\alpha) \quad d = 1, \ldots, D \tag{4}$$

This ARD prior over w_d allows us to obtain the relevance over the elements of $\mathbf{w}$, and therefore, DBL-FS is capable of automatically setting to zero the features that are irrelevant for the problem.

As the model will have to work with MRI data, composed by few samples (less than 100) and tens or hundreds of thousands of voxels, it is clear that working in the primal space is not the most efficient way to proceed. So, we propose to reformulate the model making use of the Representer Theorem [33] (RT). That is, as the RT states that the primal weights of any regression model resulting from minimizing an empirical error (risk) can be expressed as a linear combination of the input data and its equivalent dual coefficients, we can express $\mathbf{w}$ as:

$$\mathbf{w} = \mathbf{X}^T\mathbf{a} \tag{5}$$

where $\mathbf{a}$ is a vector of length N containing the dual variables. As we will see later (see Equation (19)), the lower bound that maximizes our variational inference is equivalent to minimizing an empirical cost. This way, the model can be formulated to work in the dual space as:

$$y_n = \mathbf{k}_{n,:}\mathbf{a} + \eta, \tag{6}$$

where $\mathbf{k}_{n,:}$ denotes to the n-th row of the linear kernel matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^T$. This way, with the dual formulation, the target variables y_n, for $n = 1, \ldots, N$, are modeled as:

$$y_n \sim \mathcal{N}\left(\mathbf{k}_{n,:}\mathbf{a}, \tau^{-1}\right) \quad n = 1, \ldots, N. \tag{7}$$

With this new formulation, the model will be able to work in the space of $\mathbf{a}$, where only N parameters have to be inferred. Thus, model complexity and overfitting risks are drastically reduced, as long as we are able to maintain the feature relevance determination over $\mathbf{w}$, providing the model with feature selection capabilities.

Finally, it is important to note that the model formulation does not need to specifically include the distribution of $\mathbf{a}$ since the relation between $\mathbf{w}$ and $\mathbf{a}$ is deterministic, and therefore, the statistical characterization of $\mathbf{w}$ is also characterizing $\mathbf{a}$.

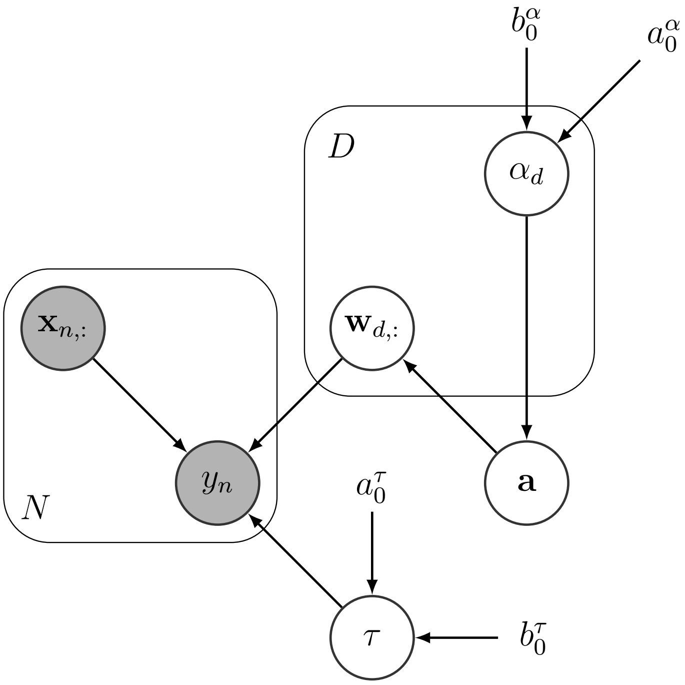

Figure 2. Plate diagram for the DBL-FS graphic model. Grey circles denote observed variables, white circles unobserved variables. Model hyperparameters do not have a circle.

3.1.3. Variational Inference

Once the generative model is defined, we should evaluate the posterior distribution of the variables to estimate their optimum values. Although, in this case, the posterior distribution is not tractable, we can use variational inference together with the mean-field technique [34] to find an approximation to this posterior $q(\boldsymbol{\Theta}) \approx p(\boldsymbol{\Theta}|\boldsymbol{y}, \mathbf{X})$, where $\boldsymbol{\Theta}$ contains all model variables. Then, we can define a Lower Bound (LB) using the Kullback–

Leibler divergence between the posterior and its approximation; so, maximizing this LB, we can obtain the optimum model parameters. Therefore, using the mean-field approximation to factorize over the posterior, we obtain:

$$p(\boldsymbol{\Theta}|\boldsymbol{y}, \mathbf{X}) \approx q(\boldsymbol{\Theta}) = q(\mathbf{w})q(\mathbf{a})q(\boldsymbol{\alpha})q(\tau), \tag{8}$$

and we can determine each approximated distribution by calculating:

$$\ln(q_j^*) = \mathbb{E}_{-q_j}[\ln(p(\mathbf{X}, \boldsymbol{y}, \boldsymbol{\Theta}))] + const, \tag{9}$$

where $\mathbb{E}_{-q_j}$ implies that we calculate the expectation over all random variables except the j-th variable, and $p(\mathbf{X}, \boldsymbol{y}, \boldsymbol{\Theta})$ is the joint probability.

Therefore, we can apply (9) to the joint probability for each random variable to obtain the model update rules. Firstly, the distribution of the dual weights **a** is:

$$q(\mathbf{a}) = \mathcal{N}(\mathbf{a}|\langle\mathbf{a}\rangle, \boldsymbol{\Sigma}_a), \tag{10}$$

with mean and variance determined by:

$$\langle\mathbf{a}\rangle = \langle\tau\rangle\boldsymbol{\Sigma}_a\mathbf{K}^T y \tag{11}$$

$$\boldsymbol{\Sigma}_a^{-1} = \mathbf{X}\mathrm{diag}(\langle\boldsymbol{\alpha}\rangle)\mathbf{X}^T + \langle\tau\rangle\mathbf{K}^T\mathbf{K}, \tag{12}$$

where $\mathrm{diag}(\langle\boldsymbol{\alpha}\rangle)$ represents an identity matrix with vector α as the diagonal. The distribution of variable $\boldsymbol{\alpha}$ is:

$$q(\boldsymbol{\alpha}) = \Gamma(\boldsymbol{\alpha}|a_\alpha, \mathbf{b}_\alpha), \tag{13}$$

with parameters

$$a_\alpha = a_0^\alpha + \frac{D}{2} \tag{14}$$

$$\mathbf{b}_\alpha = b_0^\alpha + \frac{1}{2}\mathrm{diag}(\mathbf{X}^T\langle\mathbf{a}\mathbf{a}^T\rangle\mathbf{X}), \tag{15}$$

where α_0^α and β_0^α are hyperparameters, and the operator diagonal returns a column vector formed by the main diagonal of the matrix. Moreover, the distribution of the noise precision τ is given by:

$$q(\tau) = \Gamma(\tau|a_\tau, b_\tau), \tag{16}$$

with parameters

$$a_\tau = a_0^\tau + \frac{N}{2} \tag{17}$$

$$b_\tau = b_0^\tau + \frac{1}{2}(\sum_{n=1}^{N} y_n^2 - 2Tr\{\boldsymbol{y}^T\mathbf{K}\langle\mathbf{a}\rangle\} + \mathrm{Tr}\{\mathbf{K}^T\mathbf{K}\langle\mathbf{a}\mathbf{a}^T\rangle\}), \tag{18}$$

where α_0^τ and β_0^τ are hyperparameters, and $Tr\{\dot{}\}$ is the trace operator. See Appendix A for the full development of these mean field distribution approximations.

Once we have defined the different distributions, the model updates the different random variables in an iterative coordinate-ascent-like optimization where the distribution of each factor is obtained using (10) to (22). This optimization process is guided by the *LB* cost function defined as:

$$LB = const + \sum_{n=1}^{N}\left(\frac{D}{2} + a_0^\alpha + 1\right)\ln(b_\alpha) - \left(\frac{D}{2} + a_0^\tau + 1\right)\ln(b_\tau) - \frac{D}{2}\ln(|\boldsymbol{\Sigma}_a|), \tag{19}$$

where we analyze its convergence to stop the distribution parameters update. See Appendix B for the full development of the LB.

For an efficient optimization of the model, in practice, we will work in the dual space updating the Equations (10) to (18). However, when the model convergence is reached, we can obtain the approximate posterior distribution of $\mathbf{w}$ as:

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\langle\mathbf{w}\rangle, \boldsymbol{\Sigma}_w), \tag{20}$$

with parameters

$$\langle\mathbf{w}\rangle = \langle\mathbf{a}\rangle\mathbf{X} \tag{21}$$

$$\boldsymbol{\Sigma}_w = \mathbf{X}^T\boldsymbol{\Sigma}_a\mathbf{X}. \tag{22}$$

Once the model is trained, we can analyze the distribution of $\mathbf{w}$ and check which feature components are zero and, therefore, are eliminated, having an automatic selection of the relevant input voxels. This is due to the fact that, despite working in the dual space, the precision of $\mathbf{w}$ components, $\boldsymbol{\alpha}$, is considered in the distribution of $\mathbf{a}$ (see Equation (12)).

Moreover, the inclusion of a prior over $\boldsymbol{\alpha}$ (see Equation (13)) in the generative model has an additional advantage since we can use it to adapt the prior distribution of $\mathbf{w}$ and include expert knowledge in the model. Thus, in case we want to add more relevance to a particular region (for instance, a neurobiologically meaningful Region of Interest (ROI)), we can initialize the parameters b_α associated to the voxels of this region with higher values than the rest to promote that the distribution of $\mathbf{w}$ has also higher values for these voxels. Otherwise, if we do not want to include this expert knowledge, this variable will be uniformly initialized over all voxels.

3.2. Baselines

Here, we present the baseline methods used during the experimental section, whose performances will be compared with those of the proposed DBL-FS model. In particular, we considered three approaches, one baseline aimed to solve regression problems (as DBL-FS) and the other two methods specifically designed to solve classification tasks:

- As the first baseline, we included a regression Gaussian Process (GP) [35], using the implementation provided by the GPy library (available at github accessed on 9 December 2021). We have selected this model since it allows us to define lineal kernels with ARD, so that we can work in the dual space and learn the relevance of the different input features.
- Next, we included an SVM [36] with a linear kernel using the scikit-learn library [37] to also optimize the model in the dual space.
- The last selected baseline is the recently proposed adaptation of Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA) [38] to work in the dual space, the Kernelized SSHIBA (KSSHIBA) [39] is available at github accessed on 29 March 2021. This algorithm can simultaneously combine different data sources or views (in our case, different tissues) in a common latent space providing a low-dimensional representation of the data. In addition, this model can also include an additional output view to categorically model the target variable (patient or control sample), as well as a linear kernel with ARD coefficients over the input features (equivalent to the GP configuration).

Both GP and KSSHIBA use an ARD to determine the relevance of the input features, but they do not have a prior distribution or constraint to force their input weights to be sparse and, therefore, obtain a real FS. Meanwhile, DBL-FS imposes sparsity with the Gamma prior to actually promote zero values in the model weights which, in turn, automatically eliminates the least relevant features.

Furthermore, it is important to mention that deep learning models are not included in this study, as these methods are severely limited by the sample size required to learn the model parameters. Therefore, although models such as convolutional neural networks have promising results in image analysis, they also pose serious challenges when working with datasets of small sample size. Furthermore, we have explored other baselines such as random forests but have not included the results due to their poor results. Nevertheless, all the methods under study will be evaluated with different configurations to be able to analyze different properties and, hence, to carry out an extensive study and analyze their advantages and disadvantages in comparison to our model.

3.3. Experimental Setup

MRI data were standardized to zero mean and unitary standard deviation. As we have a reduced number of subjects (only 53 samples), we have used a Leave-One-Out (LOO) framework to evaluate the model performance. This way, we have trained as many models as available samples, using in each training partition all the subjects except one, which was used afterwards for testing. Then, to evaluate the model performance, we present the results in terms of average accuracy, that is, the percentage of correctly classified test subjects computed over all LOO iterations. Furthermore, since the performance of some methods depends on their initialization we repeated the LOO process 10 times (with different initializations) and depicted the average accuracy over them in order to obtain more statically significant results.

To complete the performance analysis, the result table includes the final number of voxels selected by each model (and their percentage with respect to the total), computed as the average number of voxels used by each model for each LOO iteration and each run.

Regarding the different models under study, we considered several configurations to carry out a more comprehensive analysis and more adequate evaluation of the different methods.

For GPs, we have considered two versions: (1) the standard GP with a linear kernel, denoted as GP, and (2) the previous GP but including ARD capabilities and an FS stage. That is, we first trained a GP with ARD and analyzed the ARD coefficients to select the most relevant features, and then trained a standard GP with the chosen features. Thus, this two-step approach provided a GP with FS capabilities, denoted as GP+FS. For this pruning, we selected the 25% most relevant features in order to compare the performance of this method with DBL-FS. In addition, as both DBL-FS and GPs were formulated for regression problems and our predictive task is a binary classification (0 or 1), we set the threshold to 0.5.

We have implemented two different approaches for SVMs: (1) a standard SVM with linear kernel and (2) an SVM with a Multi-Kernel Learning (MKL) strategy, denoted as MKL-SVM. In the latter case, we independently considered the different tissues (GM, CSF, and WM) and a different linear kernels for each of them, and subsequently, the model learned the combination of these three kernels, including two parameters for their combination. These parameters were defined as scalars multiplying each kernel term, and a subsequent inner LOO was used to find their optimal values. Thus, the defined combinations coefficients gave more or less relevance to each kernel (therefore, to each tissue), providing additional flexibility to the model.

For KSSHIBA, we have included two versions, similarly to what was done in GP: (1) the standard KSSHIBA model and (2) a two-stage version of KSSHIBA (denoted as KSSHIBA+FS), in which KSSHIBA was first trained with ARD functionality, and subsequently, we selected the most relevant features to train the model using this subset of features. For these experiments, we initially had 1000 latent factors, from which the model will automatically prune the irrelevant ones. For FS, we kept the highest 25% of voxels equivalently to the number of selected features from the DBL-FS model.

Finally, we have also defined two approaches for the DBL-FS model, with and without expert knowledge. In the latter one, we have equally initialized the ARD prior for all voxels, setting the parameters a and b of random variable α to 2 and 1, respectively. In the expert knowledge case (denoted as DBL-FS+EK), we have initialized the parameters a and b in such a way that the areas of the prefrontal cortex, ventral hippocampus, and lateral ventricles (which are known to be more intensely affected [23]) had more relevance than the rest. In particular, parameter a was set to 50 and parameter b was fixed to either 1 or 0.001, depending on whether the voxel belonged to the indicated ROIs or not.

4. Experimental Results

Table 1 shows the LOO accuracy results for the classification problem together with the number of selected voxels (the approaches without FS used 100%). Despite using different initializations in the evaluated models, the results were stable across them with a negligible standard deviation, showing that the initialization hardly influences the results. For this reason, we did not include the standard deviation in Table 1. The results show that GPs, KSSHIBA+FS, and MKL-SVM obtained the worst classification accuracies, while SVM and KSSHIBA achieved the best performances among the baselines. However, DBL-FS and DBL-FS+EK still obtained an improvement of 5.7% in terms of accuracy over the best baseline while learning the most restrictive selection mask. From these results, we need to highlight that (1) KSSHIBA obtained a predictive performance similar to that of SVM while summarizing the information of the original data (distributed in more than 10^6 voxels) in only nine latent variables, and (2) MKL-SVM showed worse results than the standard SVM, probably due to the higher number of hyperparameters it needed to learn in order to perform the MKL, which may be causing overfitting.

Table 1. Performance of the different methods under study showing the model accuracy and the number of selected voxels (with their percentages with respect to the total). In addition, models with the best accuracy have been highlighted in bold and placed at the bottom of the table, which corresponds to the proposed DBL-FS approaches.

Experiment	Accuracy	# Selected Voxels
GP	67.9%	1,077,656 (100%)
GP+FS	67.9%	269,414 (25%)
SVM	71.6%	1,077,656 (100%)
MKL-SVM	67.9%	1,077,656 (100%)
KSSHIBA	69.8%	1,077,656 (100%)
KSSHIBA+FS	64.1%	269,414 (25%)
DBL-FS	**77.3%**	**287,996 (26.72%)**
DBL-FS+EK	**77.3%**	**242,754 (22.52%)**

Figure 3 shows the brain areas selected by the methods with FS capabilities. As each voxel has an associated weight, the image masks represent the absolute value of these weight magnitudes for the selected of voxels as an indicator of the voxel relevance. In addition, since we have a model for each LOO iteration, Figure 3 displays the average values of these relevances (over all LOO iterations) and includes a normalization of their scales to the range $(0,1)$ to simplify their analysis. As a result, we can observe that the GP-FS selected meaningless voxels in neurobiological terms while KSSHIBA detected well-defined areas corresponding to only WM tissues. Finally,the DBL-FS and DBL-FS+EK approaches obtained well-defined regions in the GM and WM tissues and the CSF, which are interpretable in neurobiological terms. Although both methods provided similar selections, DBL-FS+EK selected a reduced set of voxels.

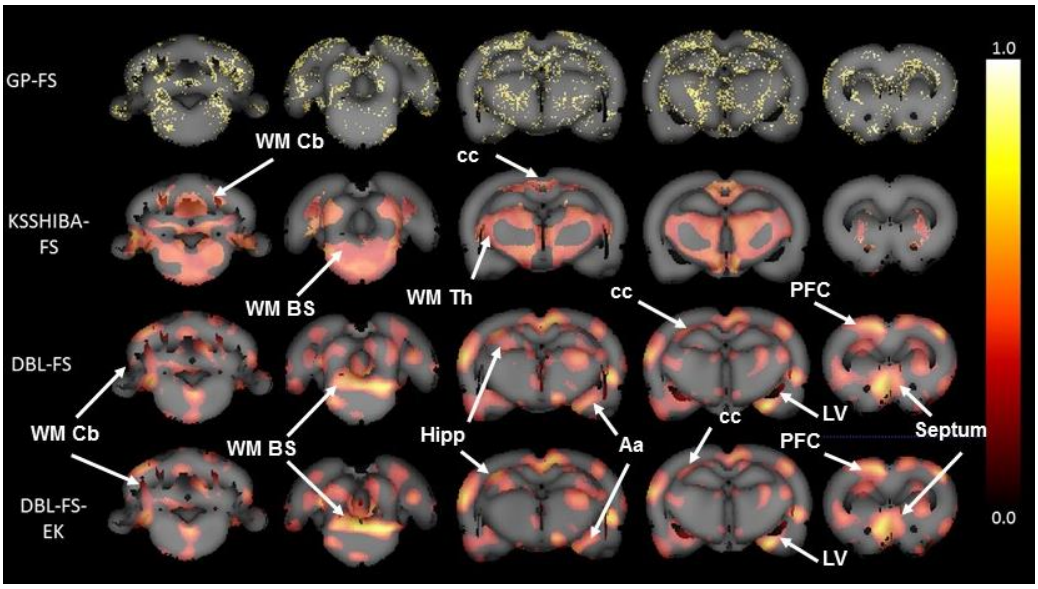

Figure 3. Brain masks obtained by the FS of each model. Colors are defined in a linear scale and associated to the relevance of the voxel (white: more relevant; dark red: less relevant). GP-FS model yields meaningless results in neurobiological terms, where no anatomical regions can be identified. KSSHIBA-FS model only identifies brain areas related to WM deficits in schizophrenia. Both DBL-FS and DBL-FS+EK learn similar relevance in WM, GM, and CSF brain areas of interest in schizophrenia, such as the hippocampus (hipp), prefrontal cortex (PFC), amygdala (Aa), septum, lateral ventricles (LV), corpus callosum (cc), WM cerebellar (WM Cb), and WM brainstem (WM BS) fibers.

5. Discussion

This study shows, for the first time, the great advantage of using DBL-FS for the detection and characterization of the morphometric brain changes in a rodent model of schizophrenia. This Bayesian model was adapted for neuroimaging data, characterized by a low sample-to-variable ratio (53 samples vs. 1,077,656 voxels in our case) relying on a dual formulation of the Bayesian Linear Regression model. Furthermore, as the main novelty of this proposal, we combine this dual formulation with a prior over the primal weights to learn the feature relevance over the input features, forcing an automatic FS. Finally, we can exploit the Bayesian nature of the model to include specific prior knowledge to guide the learning process and counterweight the limitations caused by the low sample size of the problem.

Thus, the comparison in terms of performance with the baselines provides clear evidence of the promising results of the proposed DBL-FS model in the characterization of neuroimaging data in mental disorders. Note that DBL-FS is able to largely outperform the baselines in prediction accuracy, showing an advantage of 5.7% in terms of accuracy over the best baseline. In addition, DBL-FS is the only method capable of detecting regions within the three brain tissues that are known to be relevant in the biology of schizophrenia. In this sense, the relevance learned by the GPs is inconsistent between the different LOO iterations, generating a scattered voxel selection and, therefore, a non-localized, unreliable, and uninterpretable mask. On the other hand, KSSHIBA provides a consistent result but only finds relevant regions within WM tissue and, therefore, ignores relevant regions and includes some irrelevant areas.

Analyzing in detail the regions selected by the DBL-FS and DBL-FS+Ek models, we can verify that these areas belong to brain regions whose morphometric changes have been related to schizophrenia, based on the literature. First, as for CSF, the areas with the greatest weight were the most frontal areas of the lateral ventricles and the third ventricle. One of the morphometric hallmarks in schizophrenia is the enlargement of the ventricles [23,40], which is consistent with the learned selection. Second, regarding GM, our model clearly defined anatomical areas, such as the prefrontal cortex (PFC), hippocampus, amygdala, and septum, some of them in both hemispheres. Numerous studies have demonstrated the relevance of the morphological changes of these areas in mental disorders [41,42] together with the disconnection and lack of symmetry between both cerebral hemispheres [43,44]. Similar volumetric abnormalities have also been reported for the animal model used in this study [24,45]. In addition, the method also detected the medial septum, which plays a significant role in dopamine-related disorders such as schizophrenia [46,47] and addictions [48–50], which highlights the relevance of this structure in mental disorders. Regarding WM, the method found three well-defined brain areas, the frontal part of corpus callosum and WM tracts of the brainstem and the cerebellum [51,52].

Regarding the inclusion of expert knowledge by means of the $\boldsymbol{\alpha}$ prior, it reveals two interesting behaviors. First, it demonstrates the robustness and potential of the standard DBL-FS since it is able to obtain similar accuracy and roughly similar brain masks to its DBL-FS+EK extension without the need for expert information. Second, the possibility of including expert knowledge makes the model converge faster, and it also refines the brain region selection. It is important to note that, although the expert knowledge guides the inference process, the model is also learning from the data, allowing it to redefine the initial expert knowledge into a specific set of voxels. For instance, looking at Figure 3, we can see that using expert knowledge, we obtain a higher relevance associated with the core of the WM brainstem and hippocampal areas.

6. Conclusions

This article shows a novel Bayesian approach using linear regression to characterize neuroimaging data, tested in an animal model of schizophrenia. The proposed DBL-FS+EK model allowed us to efficiently work with neuroimaging data, characterized by a low sample-to-variable ratio. This is achieved by taking advantage of its Bayesian formulation to work in the dual space while learning a voxel importance for feature selection. Furthermore, the use of a specific prior to force sparsity can be combined with expert knowledge to guide the model. The proposed model was analyzed using MRI data from a rodent model of a schizophrenia database and compared to different baselines. The results provided an outstanding classification performance of DBL-FS+EK, improving the accuracy of the second best classifier, SVM, in $\sim$6%. Furthermore, looking at the selected voxels and their associated relevance, we can confirm that the proposed model is able to detect biologically relevant areas for the characterization of this disease, as it clearly agrees with known literature.

Author Contributions: Conceptualization, V.G.-V. and M.L.S.-M.; methodology, V.G.-V. and C.S.-S.; software, A.B.-L. and C.S.-S.; validation, M.L.S.-M. and A.B.-L.; formal analysis, A.B.-L., V.G.-V. and C.S.-S.; investigation, V.G.-V., M.L.S.-M. and C.S.-S.; resources, M.L.S.-M.; data curation, V.G.-V. and M.L.S.-M.; writing—review and editing, all; visualization, all; supervision, V.G.-V., M.L.S.-M. and C.S.-S.; project administration, V.G.-V. and M.L.S.-M.; funding acquisition, V.G.-V., M.L.S.-M. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is part of the project PID2020-115363RB-I00 funded by MCIN/AEI/10.13039/501100011033. A.B.-L.'s work is funded by the Community of Madrid through the "Excellence of University Teaching Staff" line of the Multi-year Agreement with UC3M (EPUC3M27), within the framework of the V PRICIT. M.L.S.-M.'s was supported by Ministerio de Ciencia, Innovación y Universidades, Instituto de Salud Carlos III (project number PI17/01766, and grant number BA21/00030), co-financed by European Regional Development Fund (ERDF), "A way to make Europe", CIBER de Salud Mental (project number CB07/09/0031), Delegación del Gobierno para el Plan Nacional sobre Drogas (project number 2017/085); Fundación Mapfre and Fundación Alicia Koplowitz. M.D.'s work was supported by Ministerio de Ciencia e Innovación (MCIN) and Instituto de Salud Carlos III (ISCIII) (PT20/00044). The CNIC is supported by the ISCIII, the MCIN and the Pro CNIC Foundation, and is a Severo Ochoa Center of Excellence (SEV-2015-0505).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. DBL-FS Variational Inference

This section explains in detail the development of the variational inference of the proposed DBL-FS indicated in the Methods section. In particular, here we present the calculation of the mean field approximation of the model parameters:

$$q(\boldsymbol{\Theta}) = q(\mathbf{w})q(\mathbf{a})q(\boldsymbol{\alpha})q(\tau), \tag{A1}$$

where each term is calculated applying Equation (9) to the joint probability for each random variable to obtain the updated model rules.

Appendix A.1. Mean Field Approximation of ***a***

Using the mean field approximation over variable **a**, we find that the logarithm of its approximate posterior is:

$$\ln(q(\mathbf{a})) = \mathbb{E}[\ln(p(\boldsymbol{y}|\mathbf{X}, \mathbf{a}, \tau))] + \mathbb{E}[\ln(p(\mathbf{w}|\boldsymbol{\alpha}, \mathbf{a}))] + const. \tag{A2}$$

If we develop the first term in the equation, we have:

$$\begin{aligned}\ln(p(\boldsymbol{y}|\mathbf{X}, \mathbf{a}, \tau)) &= \sum_{n=1}^{N} \ln p(y_n|\mathbf{x}_{n,:}, \mathbf{a}, \tau) = \sum_{n=1}^{N} \ln\mathcal{N}\left(\mathbf{x}_{n,:}\mathbf{X}^T\mathbf{a}, \tau^{-1}\right) \\ &= \sum_{n=1}^{N} (\frac{1}{2}\ln(\tau) - \frac{\tau}{2}(y_n - \mathbf{a}^T\mathbf{X}\mathbf{x}_{n,:}^T)(y_n - \mathbf{x}_{n,:}\mathbf{X}^T\mathbf{a})) + const \\ &= \frac{N}{2}\ln(\tau) - \frac{\tau}{2}\sum_{n=1}^{N}\left(y_n^2 - 2y_n\mathbf{x}_{n,:}\mathbf{X}^T\mathbf{a} + \mathbf{a}^T\mathbf{X}\mathbf{x}_{n,:}^T\mathbf{x}_{n,:}\mathbf{X}^T\mathbf{a}\right) + const,\end{aligned} \tag{A3}$$

and, calculating the expectation of this expression, we obtain:

$$\mathbb{E}_\tau[\ln(p(\boldsymbol{y}|\mathbf{X}, \mathbf{a}, \tau))] = \frac{N}{2}\ln(\langle\tau\rangle) + \langle\tau\rangle\boldsymbol{y}\mathbf{K}\mathbf{a} - \frac{\langle\tau\rangle}{2}\mathbf{a}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{K}\mathbf{a} + const. \tag{A4}$$

Equivalently, the second term can be calculated as:

$$\begin{aligned}\ln(p(\mathbf{w}|\boldsymbol{\alpha}, \mathbf{a})) &= \sum_{d=1}^{D} \ln p(w_d|\alpha_d, \mathbf{a}) = \sum_{d=1}^{D} \ln\mathcal{N}\left(0, \alpha_d^{-1}\right) \\ &= \sum_{d=1}^{D}\left(\frac{1}{2}\ln(\alpha_d) - \frac{1}{2}\mathbf{a}^\mathbf{T}\mathbf{x}_{:,\mathbf{d}}\alpha_d\mathbf{x}_{:,\mathbf{d}}^\mathbf{T}\mathbf{a}\right) + const \\ &= \frac{1}{2}\sum_{d=1}^{D}\ln(\alpha_d) - \frac{1}{2}\sum_{d=1}^{D}\left(\mathbf{a}^\mathbf{T}\mathbf{x}_{:,\mathbf{d}}\alpha_d\mathbf{x}_{:,\mathbf{d}}^\mathbf{T}\mathbf{a}\right) + const,\end{aligned} \tag{A5}$$

and, we if we use the expectation, we have:

$$\mathbb{E}_\alpha[\ln(p(\mathbf{w}|\boldsymbol{\alpha}, \mathbf{a}))] = \frac{1}{2}\sum_{d=1}^{D} ln(\langle\alpha_d\rangle) - \frac{1}{2}\mathbf{a}^\mathbf{T}\mathbf{X}\mathrm{diag}(\langle\alpha\rangle)\mathbf{X}^\mathbf{T}\mathbf{a} + const. \tag{A6}$$

Now, joining Equations (A4) and (A6), we obtain:

$$\ln(q(\mathbf{a})) = \langle\tau\rangle\boldsymbol{y}^\mathbf{T}\mathbf{K}\mathbf{a} - \frac{\langle\tau\rangle}{2}\mathbf{a}^\mathbf{T}\mathbf{K}^\mathbf{T}\mathbf{K}\mathbf{a} - \frac{1}{2}\mathbf{a}^\mathbf{T}\mathbf{X}\mathrm{diag}(\langle\alpha\rangle)\mathbf{X}^\mathbf{T}\mathbf{a} + const. \tag{A7}$$

Therefore, we can identify the parameters of the q distribution on this equation, having:

$$q(\mathbf{a}) = \mathcal{N}(\mathbf{a}|\langle\mathbf{a}\rangle, \boldsymbol{\Sigma}_a) \tag{A8}$$

where the mean is:

$$\langle\mathbf{a}\rangle = \tau\boldsymbol{\Sigma}_a\mathbf{K}^{\mathbf{T}}\boldsymbol{y} \tag{A9}$$

and the variance is:

$$\boldsymbol{\Sigma}_a^{-1} = \mathbf{X}\mathrm{diag}(\langle\boldsymbol{\alpha}\rangle)\mathbf{X}^{\mathbf{T}} + \langle\tau\rangle\mathbf{K}^{\mathbf{T}}\mathbf{K} \tag{A10}$$

Appendix A.2. Mean Field Approximation of $\boldsymbol{\alpha}$

Now, using the mean field approximation over variable $\boldsymbol{\alpha}$, we find that the logarithm of its approximate posterior is:

$$\ln(q(\boldsymbol{\alpha})) = \mathbb{E}[\ln(p(\mathbf{w}|\boldsymbol{\alpha},\mathbf{a}))] + \mathbb{E}[\ln(p(\boldsymbol{\alpha}))] + const \tag{A11}$$

Developing the first term, we obtain:

$$\ln(p(\mathbf{w}|\boldsymbol{\alpha},\mathbf{a})) = \frac{1}{2}\sum_{d=1}^{D}\ln(\alpha_d) - \frac{1}{2}\sum_{d=1}^{D}\mathrm{Tr}\{\mathbf{a}^{\mathbf{T}}\mathbf{x}_{:,\mathbf{d}}\alpha_d\mathbf{x}^{\mathbf{T}}_{:,\mathbf{d}}\mathbf{a}\} + const, \tag{A12}$$

and we can apply the expectation to obtain:

$$\mathbb{E}_{a,\tau}[p(\mathbf{w}|\boldsymbol{\alpha},\mathbf{a})] = \frac{1}{2}\sum_{d=1}^{D}\ln(\alpha_d) - \frac{1}{2}\sum_{d=1}^{D}\alpha_d\mathrm{Tr}\{\mathbf{x}^{\mathbf{T}}_{:,\mathbf{d}}\langle\mathbf{a}\mathbf{a}^{\mathbf{T}}\rangle\mathbf{x}_{:,\mathbf{d}}\} \tag{A13}$$

If we look at the second term, we have

$$\ln(p(\boldsymbol{\alpha})) = \sum_{d=1}^{D}(-b_0^{\alpha}\alpha_d + (a_0^{\alpha} - 1)\ln(\alpha_d)) + const, \tag{A14}$$

where we can apply the expectation of the function as:

$$\mathbb{E}[\ln(p(\boldsymbol{\alpha}))] = \sum_{d=1}^{D}\ln(p(\alpha_d)) = \sum_{d=1}^{D}(-b_0^{\alpha}\alpha_d + (a_0^{\alpha} - 1)\ln(\alpha_d)) + const. \tag{A15}$$

Now, if we sum Equations (A13) and (A15), we obtain:

$$\ln(q(\boldsymbol{\alpha})) = \sum_{d=1}^{D}\left(\left(\frac{1}{2} + a_0^{\alpha} - 1\right)\ln(\alpha_d) - \frac{\alpha_d}{2}(\mathrm{Tr}\{\mathbf{x}^{\mathbf{T}}_{:,\mathbf{d}}\langle\mathbf{a}\mathbf{a}^{\mathbf{T}}\rangle\mathbf{x}_{:,\mathbf{d}}\} + 2b_0^{\alpha})\right) + const. \tag{A16}$$

Thus, if we identify terms on the variable distribution, we have:

$$q(\alpha_d) = \prod_{d=1}^{D}\Gamma(\alpha_d|a_{\alpha_d}, b_{\alpha_d}), \tag{A17}$$

where the first parameter is:

$$a_{\alpha} = \frac{1}{2} + a_0^{\alpha}, \tag{A18}$$

and the second parameter can be expressed as:

$$b_{\alpha} = b_0^{\alpha} + \frac{1}{2}\mathrm{diag}(\mathbf{X}^{\mathbf{T}}\langle\mathbf{a}\mathbf{a}^{\mathbf{T}}\rangle\mathbf{X}). \tag{A19}$$

Appendix A.3. Mean Field Approximation of τ

Following the same steps as in the two previous approaches, we use the mean field approximation over variable τ to obtain the logarithm of the approximate posterior:

$$\ln(q(\tau)) = E[\ln(p(\boldsymbol{y}|\mathbf{X},\mathbf{a},\tau))] + E[\ln(p(\tau))] + const. \tag{A20}$$

Therefore, the first term on this equation is:

$$\ln(p(\boldsymbol{y}|\mathbf{X},\mathbf{a},\tau)) = \frac{N}{2}\ln(\tau) - \frac{\tau}{2}\left(\sum_{n=1}^{N} y_n^2 - 2\mathrm{Tr}\left\{\boldsymbol{y}^{\mathbf{T}}\mathbf{K}\mathbf{a}\right\} + \mathrm{Tr}\left\{\mathbf{K}^{\mathbf{T}}\mathbf{K}\mathbf{a}\mathbf{a}^{\mathbf{T}}\right\}\right) + const, \tag{A21}$$

and, applying the expectation we obtain:

$$\begin{aligned}\mathbb{E}_a[\ln(p(\boldsymbol{y}|\mathbf{X},\mathbf{a},\tau))] = {} & \frac{N}{2}\ln(\tau) \\ & -\frac{\tau}{2}\left(\sum_{n=1}^{N} y_n^2 - 2\mathrm{Tr}\left\{\boldsymbol{y}^{T}\mathbf{K}\langle\mathbf{a}\rangle\right\} + Tr\left\{\mathbf{K}^{\mathbf{T}}\mathbf{K}\langle\mathbf{a}\mathbf{a}^{\mathbf{T}}\rangle\right\}\right) + const.\end{aligned} \tag{A22}$$

The second term is defined as:

$$\mathbb{E}[\ln(p(\tau))] = \ln(p(\tau)) = -b_0^\tau\tau + (a_0^\tau - 1)\ln(\tau) + const. \tag{A23}$$

Now, if we sum Equations (A22) and (A23), we obtain:

$$\begin{aligned}\ln(q(\alpha)) = {} & \left(\frac{N}{2} + a_0^\tau - 1\right)\ln(\tau) \\ & -\frac{\tau}{2}\left(\sum_{n=1}^{N} y_n^2 - 2\mathrm{Tr}\left\{\boldsymbol{y}^{\mathbf{T}}\mathbf{K}\langle\mathbf{a}\rangle\right\} + \mathrm{Tr}\left\{\mathbf{K}^{\mathbf{T}}\mathbf{K}\langle\mathbf{a}\mathbf{a}^{\mathbf{T}}\rangle\right\} + 2b_0^\tau\right) + const.\end{aligned} \tag{A24}$$

Therefore, following the same procedure as with the previous variables, we identify terms from the distribution and obtain:

$$q(\tau) = \Gamma(\tau|a_\tau, b_\tau), \tag{A25}$$

where the first parameter is:

$$a_\tau = \frac{N}{2} + a_0^\tau, \tag{A26}$$

and the second one is:

$$b_\tau = \frac{1}{2}(\sum_{n=1}^{N} y_n^2 - 2Tr\left\{\boldsymbol{y}^{\mathbf{T}}\mathbf{K}\langle\mathbf{a}\rangle\right\} + Tr\left\{\mathbf{K}^{\mathbf{T}}\mathbf{K}\langle\mathbf{a}\mathbf{a}^{\mathbf{T}}\rangle\right\}) + b_0^\tau. \tag{A27}$$

Appendix B. Lower Bound Inference

As mentioned in the Methods section, we use the Kullback–Leibler divergence to first determine the similarities between two distribution where, for any two probability density functions $p(x)$ and $q(x)$, we have:

$$D_{KL} = \int q(x)\ln\frac{q(x)}{p(x)}dx \tag{A28}$$

In our case, if we particularize for the true posterior and the posterior approximation, the divergence can be expressed as:

$$\begin{aligned} D_{KL} &= -\int q(\Theta)\ln\left(\frac{q(\Theta)}{p(\Theta|X)}\right)d\Theta = \int q(\Theta)\ln(q(\Theta))d\Theta - \int q(\Theta)\ln(p(\Theta|X))d\Theta \\ &= \mathbb{E}_q[\ln(q(\Theta))] - \mathbb{E}_q[\ln(p(\Theta|X))]. \end{aligned} \tag{A29}$$

Developing the conditional probability we obtain:

$$D_{KL} = \mathbb{E}_q[\ln(q(\Theta))] - \mathbb{E}_q[\ln(p(\Theta, X))] + \ln(p(X)). \tag{A30}$$

Due to the impossibility of working with this distribution because the marginal distribution $p(X)$ cannot be calculated, we use an Evidence Lower Bound (ELBO/LB) to this expression [34]. The LB is the divergence of negative KL plus $\ln(p(X))$; therefore, the greatest similarity between the two functions is achieved by maximizing this new measure. We can calculate the LB as:

$$\begin{aligned} L_q &= -\int q(\Theta)\ln\left(\frac{q(\Theta)}{p(X,\Theta)}\right)d\Theta = \int q(\Theta)\ln(p(X,\Theta))d\Theta - \int q(\Theta)\ln(q(\Theta))d\Theta \\ &= \mathbb{E}_q[\ln(p(X,\Theta))] - \mathbb{E}_q[\ln(q(\Theta))] \end{aligned} \tag{A31}$$

In order to easily calculate this lower bound, we will separately calculate the terms related to $\mathbb{E}_q[\ln(p(X,\Theta))]$ and to the entropy in the following subsections.

Appendix B.1. Terms Associated to $\mathbb{E}_q[\ln(p(\mathbf{X}, \boldsymbol{y}, \Theta))]$

This first term of the lower bound would be composed by the following terms:

$$\begin{aligned} \mathbb{E}_q[\ln(p(\mathbf{X}, \boldsymbol{y}, \Theta))] &= \mathbb{E}_q[\ln(p(\mathbf{X}))] + \mathbb{E}_q[\ln(p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{a}))] + \mathbb{E}_q[\ln(p(\boldsymbol{\alpha}))] \\ &\quad + \mathbb{E}_q[\ln(p(\boldsymbol{y} | \, \mathbf{a}, \mathbf{X}, \tau))] + \mathbb{E}_q[\ln(p(\tau))] \end{aligned} \tag{A32}$$

This way, the different elements of this equation can be calculated as:

$$\begin{aligned} \mathbb{E}_q[\ln(p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{a}))] = -\frac{D}{2}\ln(2\pi) + \frac{D}{2}\sum_{\mathrm{d}=1}^{\mathrm{D}}\left(\psi\left(a_{\mathrm{ff_d}}\right) - \ln\left(b_{\mathrm{ff_d}}\right)\right) \\ - \sum_{\mathrm{d}=1}^{\mathrm{D}}\left(a_{\mathrm{ff_d}}\right) + b_0^{\alpha}\sum_{\mathrm{d}=1}^{\mathrm{D}}\left(\frac{a_{\mathrm{ff_d}}}{b_{\mathrm{ff_d}}}\right) \end{aligned} \tag{A33}$$

$$\begin{aligned} \mathbb{E}_q[\ln(p(\boldsymbol{\alpha}))] = (a_0^{\alpha}\ln(b_0^{\alpha}) - \ln(\Gamma(a_0^{\alpha}))) \\ + \sum_{\mathrm{d}=1}^{\mathrm{D}}\left(-b_0^{\alpha}\frac{a_{\mathrm{ff_d}}}{b_{\mathrm{ff_d}}} + (a_0^{\alpha} - 1)\left(\psi\left(a_{\mathrm{ff_d}}\right) - \ln\left(b_{\mathrm{ff_d}}\right)\right)\right) \end{aligned} \tag{A34}$$

$$\begin{aligned} \mathbb{E}_q[\ln(p(\mathbf{w}, \boldsymbol{\alpha}))] &= \left(\frac{D}{2} + a_0^{\alpha} - 1\right)\sum_{\mathrm{d}=1}^{\mathrm{D}}\left(\psi\left(a_{\mathrm{ff_d}}\right) - \ln\left(b_{\mathrm{ff_d}}\right)\right) - \frac{D}{2}\ln(2\pi) \\ &\quad + (a_0^{\alpha}\ln(b_0^{\alpha}) - \ln(\Gamma(a_0^{\alpha}))) - \sum_{\mathrm{d}=1}^{\mathrm{D}}\left(a_{\mathrm{ff_d}}\right) \end{aligned} \tag{A35}$$

$$
\begin{aligned}
\mathbb{E}_q[\ln(p(\boldsymbol{y}|\,\mathbf{a},\mathbf{X},\tau))] &= -\frac{ND}{2}\ln(2\pi)+\frac{D}{2}\sum_{\mathrm{n=1}}^{\mathrm{N}}(\mathbb{E}_q[\ln(\tau)]) \\
&\quad -\frac{1}{2}\mathbb{E}_q[\tau]\left(\sum_{\mathrm{n=1}}^{\mathrm{N}} y_n^2 - 2\,\mathrm{Tr}\left\{\boldsymbol{y}^{\mathrm{T}}\,\mathbf{K}\langle\mathbf{a}\rangle\right\}+\mathrm{Tr}\left\{\langle\mathbf{a},\mathbf{a}^{\mathrm{T}}\rangle\,\mathbf{K}^{\mathrm{T}}\,\mathbf{K}\right\}\right) \\
&= -\frac{ND}{2}\ln(2\pi)+\frac{D}{2}(\psi(a_\tau)-\ln(b_\tau))-a_\tau+\frac{a_\tau}{b_\tau}b_0^\tau
\end{aligned}
\tag{A36}
$$

$$
\mathbb{E}_q[\ln(p(\tau))] = a_0^\tau\ln(b_0^\tau)-\ln(\Gamma(a_0^\tau))-b_0^\tau\frac{a_\tau}{b_\tau}+(a_0^\tau-1)(\psi(a_\tau)-\ln(b_\tau)) \tag{A37}
$$

$$
\begin{aligned}
\mathbb{E}_q[\ln(p(\boldsymbol{y},\tau|\,\mathbf{a},\mathbf{X}))] &= -\frac{ND}{2}\ln(2\pi)-a_\tau+a_0^\tau\ln(b_0^\tau)-\ln(\Gamma(a_0^\tau))+ \\
&\quad \left(\frac{D}{2}+a_0^\tau-1\right)(\psi(a_\tau)-\ln(b_\tau))
\end{aligned}
\tag{A38}
$$

Appendix B.2. Terms of Entropy

The second term in the LB expression is the entropy of the model parameters:

$$
\mathbb{E}_q[\ln(q(\Theta))] = \mathbb{E}_q[\ln(q(\mathbf{w}))]+\mathbb{E}_q[\ln(q(\boldsymbol{\alpha}))]+\mathbb{E}_q[\ln(q(\tau))], \tag{A39}
$$

where the entropy of these parameters is:

$$
\mathbb{E}_q[\ln(q(\mathbf{w}))] = \frac{D}{2}\ln(2\pi e)+\frac{D}{2}\ln|\Sigma_{\mathbf{w}}| \tag{A40}
$$

$$
\begin{aligned}
&\mathbb{E}_q\left[\ln\left(q\left(\boldsymbol{\alpha}^{(\mathrm{m})}\right)\right)\right] = \\
&\quad \sum_{\mathrm{k=1}}^{\mathrm{K_c}}\left(a_{\mathrm{ff}_{\mathrm{k}}^{(\mathrm{m})}}+\ln\left(\Gamma\left(a_{\mathrm{ff}_{\mathrm{k}}^{(\mathrm{m})}}\right)\right)-\left(1-a_{\mathrm{ff}_{\mathrm{k}}^{(\mathrm{m})}}\right)\psi\left(a_{\mathrm{ff}_{\mathrm{k}}^{(\mathrm{m})}}\right)-\ln\left(b_{\mathrm{ff}_{\mathrm{k}}^{(\mathrm{m})}}\right)\right)
\end{aligned}
\tag{A41}
$$

$$
\mathbb{E}_q\left[\ln\left(q\left(\text{ø}^{(\mathrm{m})}\right)\right)\right] = a_{\text{ø}^{(\mathrm{m})}}+\ln(\Gamma(a_{\text{ø}^{(\mathrm{m})}}))-(1-a_{\text{ø}^{(\mathrm{m})}})\psi(a_{\text{ø}^{(\mathrm{m})}})-\ln(b_{\text{ø}^{(\mathrm{m})}}) \tag{A42}
$$

Appendix B.3. Complete Lower Bound

Finally, joining Equations (A32) and (A39), the complete lower bound is calculated as:

$$
\begin{aligned}
L_q &= -\left(\frac{D}{2}+a_0^\alpha-1\right)\sum_{\mathrm{k=1}}^{\mathrm{K_c}}(\ln(b_{\alpha_k}))-\left(\frac{D}{2}+a_0^\tau-1\right)(\ln(b_\tau)) \\
&\quad -\frac{D}{2}\ln|\Sigma_{\mathbf{w}}|+\sum_{\mathrm{k=1}}^{\mathrm{K_c}}(\ln(b_{\alpha_k}))+\ln(b_\tau)+\mathrm{const}
\end{aligned}
\tag{A43}
$$

References

1. Carvalho, A.F.; Solmi, M.; Sanches, M.; Machado, M.O.; Stubbs, B.; Ajnakina, O.; Sherman, C.; Sun, Y.R.; Liu, C.S.; Brunoni, A.R.; et al. Evidence-based umbrella review of 162 peripheral biomarkers for major mental disorders. *Transl. Psychiatry* **2020**, *10*, 152. [CrossRef] [PubMed]
2. Widing, L.; Simonsen, C.; Flaaten, C.B.; Haatveit, B.; Vik, R.K.; Wold, K.F.; Åsbø, G.; Ueland, T.; Melle, I. Symptom Profiles in Psychotic Disorder Not Otherwise Specified. *Front. Psychiatry* **2020**, *11*, 580444. [CrossRef] [PubMed]
3. Correll, C.U.; Brevig, T.; Brain, C. Patient characteristics, burden and pharmacotherapy of treatment-resistant schizophrenia: Results from a survey of 204 US psychiatrists. *BMC Psychiatry* **2019**, *19*, 362. [CrossRef] [PubMed]

4. Roberts, L.W.; Chan, S.; Torous, J. New tests, new tools: Mobile and connected technologies in advancing psychiatric diagnosis. *NPJ Digit. Med.* **2018**, *1*, 20176. [CrossRef] [PubMed]
5. Li, Z.; Li, W.; Wei, Y.; Gui, G.; Zhang, R.; Liu, H.; Chen, Y.; Jiang, Y. Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls. *Comput. Med. Imaging Graph.* **2021**, *89*, 101882. [CrossRef] [PubMed]
6. Trakadis, Y.J.; Sardaar, S.; Chen, A.; Fulginiti, V.; Krishnan, A. Machine learning in schizophrenia genomics, a case-control study using 5090 exomes. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2019**, *180*, 103–112. [CrossRef]
7. Hettige, N.C.; Nguyen, T.B.; Yuan, C.; Rajakulendran, T.; Baddour, J.; Bhagwat, N.; Bani-Fatemi, A.; Voineskos, A.N.; Chakravarty, M.M.; De Luca, V. Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach. *Gen. Hosp. Psychiatry* **2017**, *47*, 20–28. [CrossRef]
8. Xiao, Y.; Yan, Z.; Zhao, Y.; Tao, B.; Sun, H.; Li, F.; Yao, L.; Zhang, W.; Chandan, S.; Liu, J.; et al. Support vector machine-based classification of first episode drug-naïve schizophrenia patients and healthy controls using structural MRI. *Schizophr. Res.* **2019**, *214*, 11–17. [CrossRef]
9. Guo, Y.; Qiu, J.; Lu, W. Support Vector Machine-Based Schizophrenia Classification Using Morphological Information from Amygdaloid and Hippocampal Subregions. *Brain Sci.* **2020**, *10*, 562. [CrossRef]
10. Jahmunah, V.; Oh, S.L.; Rajinikanth, V.; Ciaccio, E.J.; Cheong, K.H.; Arunkumar, N.; Acharya, U.R. Automated detection of schizophrenia using nonlinear signal processing methods. *Artif. Intell. Med.* **2019**, *100*, 101698. [CrossRef]
11. Brownlee, J. Recursive Feature Elimination (RFE) for Feature Selection in Python. Machine Learning Mastery. 2020. Available online: https://machinelearningmastery.com/rfe-feature-selection-in-python/ (accessed on 25 May 2020).
12. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
13. Li, X.; Chen, W.; Zhang, Q.; Wu, L. Building auto-encoder intrusion detection system based on random forest feature selection. *Comput. Secur.* **2020**, *95*, 101851. [CrossRef]
14. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
15. Amini, F.; Hu, G. A two-layer feature selection method using genetic algorithm and elastic net. *Expert Syst. Appl.* **2021**, *166*, 114072. [CrossRef]
16. Shen, L.; Qi, Y.; Kim, S.; Nho, K.; Wan, J.; Risacher, S.L.; Saykin, A.J. Sparse bayesian learning for identifying imaging biomarkers in AD prediction. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Beijing, China, 20–24 September 2010; pp. 611–618.
17. Sabuncu, M.R.; Van Leemput, K. The relevance voxel machine (RVoxM): A self-tuning Bayesian model for informative image-based prediction. *IEEE Trans. Med. Imaging* **2012**, *31*, 2290–2306. [CrossRef]
18. Sabuncu, M.R. A sparse Bayesian learning algorithm for longitudinal image data. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 411–418.
19. Parrado-Hernández, E.; Gómez-Verdejo, V.; Martínez-Ramón, M.; Shawe-Taylor, J.; Alonso, P.; Pujol, J.; Menchón, J.M.; Cardoner, N.; Soriano-Mas, C. Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction. *Med. Image Anal.* **2014**, *18*, 435–448. [CrossRef]
20. Gómez-Verdejo, V.; Parrado-Hernández, E.; Tohka, J. Sign-consistency based variable importance for machine learning in brain imaging. *Neuroinformatics* **2019**, *17*, 593–609. [CrossRef]
21. Sevilla-Salcedo, C.; Gómez-Verdejo, V.; Tohka, J. Regularized Bagged Canonical Component Analysis for Multiclass Learning in Brain Imaging. *Neuroinformatics* **2020**, *18*, 641–659. [CrossRef]
22. Grimm, O.; Gass, N.; Weber-Fahr, W.; Sartorius, A.; Schenker, E.; Spedding, M.; Risterucci, C.; Schweiger, J.I.; Böhringer, A.; Zang, Z.; et al. Acute ketamine challenge increases resting state prefrontal-hippocampal connectivity in both humans and rats. *Psychopharmacology* **2015**, *232*, 4231–4241. [CrossRef]
23. Hadar, R.; Soto-Montenegro, M.L.; Götz, T.; Wieske, F.; Sohr, R.; Desco, M.; Hamani, C.; Weiner, I.; Pascau, J.; Winter, C. Using a maternal immune stimulation model of schizophrenia to study behavioral and neurobiological alterations over the developmental course. *Schizophr. Res.* **2015**, *166*, 238–247. [CrossRef]
24. Romero-Miguel, D.; Casquero-Veiga, M.; MacDowell, K.S.; Torres-Sanchez, S.; Garcia-Partida, J.A.; Lamanna-Rama, N.; Romero-Miranda, A.; Berrocoso, E.; Leza, J.C.; Desco, M.; et al. A Characterization of the Effects of Minocycline Treatment During Adolescence on Structural, Metabolic, and Oxidative Stress Parameters in a Maternal Immune Stimulation Model of Neurodevelopmental Brain Disorders. *Int. J. Neuropsychopharmacol.* **2021**, *24*, 734–748. [CrossRef] [PubMed]
25. Ozawa, K.; Hashimoto, K.; Kishimoto, T.; Shimizu, E.; Ishikura, H.; Iyo, M. Immune activation during pregnancy in mice leads to dopaminergic hyperfunction and cognitive impairment in the offspring: A neurodevelopmental animal model of schizophrenia. *Biol. Psychiatry* **2006**, *59*, 546–554. [CrossRef] [PubMed]
26. Zhu, F.; Zheng, Y.; Liu, Y.; Zhang, X.; Zhao, J. Minocycline alleviates behavioral deficits and inhibits microglial activation in the offspring of pregnant mice after administration of polyriboinosinic–polyribocytidilic acid. *Psychiatry Res.* **2014**, *219*, 680–686. [CrossRef] [PubMed]
27. Meyer, U.; Feldon, J. To poly (I: C) or not to poly (I: C): Advancing preclinical schizophrenia research through the use of prenatal immune activation models. *Neuropharmacology* **2012**, *62*, 1308–1321. [CrossRef]
28. Casquero-Veiga, M.; Garcia-Garcia, D.; MacDowell, K.S.; Perez-Caballero, L.; Torres-Sanchez, S.; Fraguas, D.; Berrocoso, E.; Leza, J.C.; Arango, C.; Desco, M.; et al. Risperidone administered during adolescence induced metabolic, anatomical and

inflammatory/oxidative changes in adult brain: A pet and mri study in the maternal immune stimulation animal model. *Eur. Neuropsychopharmacol.* **2019**, *29*, 880–896. [CrossRef]

29. Valdes Hernandez, P.A.; Sumiyoshi, A.; Nonaka, H.; Haga, R.; Aubert Vasquez, E.; Ogawa, T.; Iturria Medina, Y.; Riera, J.J.; Kawashima, R. An in vivo MRI template set for morphometry, tissue segmentation, and fMRI localization in rats. *Front. Neuroinform.* **2011**, *5*, 26.
30. Bishop, C.M. Bayesian PCA. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 1999 ; pp. 382–388.
31. Klami, A.; Virtanen, S.; Kaski, S. Bayesian canonical correlation analysis. *J. Mach. Learn. Res.* **2013**, *14*, 965–1003.
32. Bishop, C.M. Pattern recognition. *Mach. Learn.* **2006**, *128* , 1–39 .
33. Schölkopf, B.; Herbrich, R.; Smola, A.J. A generalized representer theorem. In Proceedings of the International Conference on Computational Learning Theory, Amsterdam, The Netherlands, 16–19 July 2001; pp. 416–426.
34. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. doi: [CrossRef]
35. Rasmussen, C.E. Gaussian processes in machine learning. In *Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 63–71.
36. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: New York, NY, USA, 2008.
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
38. Sevilla-Salcedo, C.; Gómez-Verdejo, V.; Olmos, P.M. Sparse semi-supervised heterogeneous interbattery bayesian analysis. *Pattern Recognit.* **2021**, *120*, 108141. [CrossRef]
39. Sevilla-Salcedo, C.; Guerrero-López, A.; Olmos, P.M.; Gómez-Verdejo, V. Bayesian Sparse Factor Analysis with Kernelized Observations. *arXiv* **2020**, arXiv:2006.00968.
40. Styner, M.; Lieberman, J.A.; McClure, R.K.; Weinberger, D.R.; Jones, D.W.; Gerig, G. Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4872–4877. [CrossRef] [PubMed]
41. Rapado-Castro, M.; Villar-Arenzana, M.; Janssen, J.; Fraguas, D.; Bombin, I.; Castro-Fornieles, J.; Mayoral, M.; González-Pinto, A.; de la Serna, E.; Parellada, M.; et al. Fronto-Parietal Gray Matter Volume Loss Is Associated with Decreased Working Memory Performance in Adolescents with a First Episode of Psychosis. *J. Clin. Med.* **2021**, *10*, 3929. [CrossRef]
42. Wen, D.; Wang, J.; Yao, G.; Liu, S.; Li, X.; Li, J.; Li, H.; Xu, Y. Abnormality of subcortical volume and resting functional connectivity in adolescents with early-onset and prodromal schizophrenia. *J. Psychiatr. Res.* **2021**, *140*, 282–288 . [CrossRef]
43. Guo, S.; Kendrick, K.M.; Zhang, J.; Broome, M.; Yu, R.; Liu, Z.; Feng, J. Brain-wide functional inter-hemispheric disconnection is a potential biomarker for schizophrenia and distinguishes it from depression. *Neuroimage Clin.* **2013**, *2*, 818–826. [CrossRef]
44. Boklage, C.E. Schizophrenia, brain asymmetry development, and twinning: Cellular relationship with etiological and possibly prognostic implications. *Biol. Psychiatry* **1977**, *12*, 19–35.
45. Casquero-Veiga, M.; Romero-Miguel, D.; MacDowell, K.S.; Torres-Sanchez, S.; Garcia-Partida, J.A.; Lamanna-Rama, N.; Gómez-Rangel, V.; Romero-Miranda, A.; Berrocoso, E.; Leza, J.C.; et al. Omega-3 fatty acids during adolescence prevent schizophrenia-related behavioural deficits: Neurophysiological evidences from the prenatal viral infection with PolyI: C. *Eur. Neuropsychopharmacol.* **2021**, *46*, 14–27. [CrossRef]
46. Bortz, D.M.; Grace, A.A. Medial septum activation produces opposite effects on dopamine neuron activity in the ventral tegmental area and substantia nigra in MAM vs. normal rats. *NPJ Schizophr.* **2018**, *4*, 17. [CrossRef]
47. Takeuchi, Y.; Nagy, A.; Barcsai, L.; Li, Q.; Ohsawa, M.; Mizuseki, K.; Berényi, A. The medial septum as a potential target for treating brain disorders associated with oscillopathies. *Front. Neural Circuits* **2021**, *15*, 701080. [CrossRef] [PubMed]
48. McGlinchey, E.M.; Aston-Jones, G. Dorsal hippocampus drives context-induced cocaine seeking via inputs to lateral septum. *Neuropsychopharmacology* **2018**, *43*, 987–1000. [CrossRef] [PubMed]
49. Pantazis, C.B.; Aston-Jones, G. Lateral septum inhibition reduces motivation for cocaine: Reversal by diazepam. *Addict. Biol.* **2020**, *25*, e12742. [CrossRef] [PubMed]
50. Gárate-Pérez, M.F.; Méndez, A.; Bahamondes, C.; Sanhueza, C.; Guzmán, F.; Reyes-Parada, M.; Sotomayor-Zárate, R.; Renard, G.M. Vasopressin in the lateral septum decreases conditioned place preference to amphetamine and nucleus accumbens dopamine release. *Addict. Biol.* **2021**, *26*, e12851. [CrossRef] [PubMed]
51. Yang, M.; Gao, S.; Zhang, X. Cognitive deficits and white matter abnormalities in never-treated first-episode schizophrenia. *Transl. Psychiatry* **2020**, *10*, 368. [CrossRef]
52. Kim, S.E.; Jung, S.; Sung, G.; Bang, M.; Lee, S.H. Impaired cerebro-cerebellar white matter connectivity and its associations with cognitive function in patients with schizophrenia. *NPJ Schizophr.* **2021**, *7*, 38. [CrossRef]