Sevilla, C., Gómez, V. & Olmos, P. M. (2021). Sparse semi-supervised heterogeneous interbattery bayesian analysis. *Pattern Recognition, 120*, 108141.

DOI: 10.1016/j.patcog.2021.108141

Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis

Carlos Sevilla-Salcedo*, Vanessa Gómez-Verdejo, Pablo M. Olmos**

Department of Signal Processing and Communications, Universidad Carlos III de Madrid Leganés, 28911 Spain

Abstract

The Bayesian approach to feature extraction, known as factor analysis (FA), has been widely studied in machine learning to obtain a latent representation of the data. An adequate selection of the probabilities and priors of these bayesian models allows the model to better adapt to the data nature (i.e. heterogeneity, sparsity), obtaining a more representative latent space.

The objective of this article is to propose a general FA framework capable of modelling any problem. To do so, we start from the Bayesian Inter-Battery Factor Analysis (BIBFA) model, enhancing it with new functionalities to be able to work with heterogeneous data, to include feature selection, and to handle missing values as well as semi-supervised problems.

The performance of the proposed model, Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA), has been tested on different scenarios to evaluate each one of its novelties, showing not only a great versatility and an interpretability gain, but also outperforming most of the state-of-the-art algorithms.

Keywords: Bayesian model, Canonical Correlation Analysis, Principal Component Analysis, factor analysis, feature selection, semi-supervised, multi-task

Carlos Sevilla-Salcedo received the B.Sc. degree in telecommunication engineering from the Universidad de Granada in 2015. He then received the M.Sc. doing two masters in 2017 and his Ph.D. in 2021 in the Universidad

*Corresponding author. Email address: sevisal@tsc.uc3m.es

**Pablo M. Olmos is also with the Gregorio Marañón Health Research Institute.

November 23, 2021

Carlos III de Madrid. His research is focused on bayesian machine learning and latent models.
Vanessa Gómez-Verdejo received the Engineering degree in 2002 from Universidad Politécnica de Madrid. In 2007, she obtained a Ph.D. from Universidad Carlos III de Madrid, where she is currently Associated Professor. Her research is focused on machine learning and feature selection approaches. http://vanessa.webs.tsc.uc3m.es/
Pablo M. Olmos received the B.Sc./M.Sc.and Ph.D. degrees from the University of Sevilla in 2008 and 2011. He is currently an Associate Professor with the Universidad Carlos III de Madrid. His research interests range from Bayesian machine learning to information theory. A detailed CV and list of publications can be accessed at http://www.tsc.uc3m.es/~olmos.

1. Introduction

Feature Extraction (FE) is used to transform data into a new low dimensional latent space while removing correlations and noisy components [1], what has made it play an important role in the Machine Learning (ML) community. In particular, one method that has been increasingly used in this context is Canonical Correlation Analysis (CCA) [2], which constructs the latent space from the correlation between different views. Despite being commonly used for a single input and output views [3, 4], its formulation allows to combine multiple views of the data to improve the extraction of the latent features [5, 6, 7] , what is commonly known as multi-task or multi-view learning.

FE algorithms have been adapted to the Bayesian framework, introducing a probabilistic model able to correlate all involved views and latent low-dimensional variables [8, 9]. This new formulation, known as Factor Analysis (FA), has been used in multi-tasks problems such as biomarkers design and classification [10], person and digit classification [11] or modelling functional neuroimaging data [12].

Bayesian algorithms have the additional advantage of facilitating the inclusion of constraints on the model by defining particular priors over the model variables. For example, the distribution of the latent variables of a FA algorithm can be redefined to impose sparsity on the number of latent factors [13, 14]. This way, the model is capable of automatically determining which latent factors are relevant and eliminate the useless ones. Other approaches include Feature Selection (FS) so that the model is capable of learning the

feature relevance during its training [15, 16]. Furthermore, the probabilistic design allows modelling real data via continuous distributions or categorical data with discrete distributions, hence capturing the real data nature. Most methods developed for Bayesian FA centre around working with real data, whereas there are not many studies about more specific data. In particular, [17] presents an algorithm that combines FA with sparsity in the latent space, as well as working with categorical data. By treating the categorical data as whole numbers, the data distribution fits better the original data [18]. Conversely, multilabelled methods consider the correlation between labels to model them [19] improving the final results [20, 21].

Another advantage of probabilistic modelling is that we can naturally deal with missing data. In semi-supervised multi-task learning, the model learns from the available information (available views) for every data point and, with that, missing views are integrated out. In this respect, some algorithms combine this semi-supervised approaches with the sparsity in the data distribution to model words' labels [22] or with modelling multi-label and categorical data [23]. Other methods [24] propose a semi-supervised extension of a Deep Generative Model to obtain a more informative model. Or, some models, such as [25] and [26], combine a semi-supervised learning with a Bayesian Principal Component Regression to model soft sensors for industrial applications. Among the different approaches in the literature for FA and their extensions, the Bayesian Inter-Battery FA model (BIBFA) [27] has specially attracted our attention since it provides a framework for FA where one can work with multiple data views and sparsity over the latent factors to automatically select the number of latent variables. However, we miss some functionalities in the model to really have a versatile framework able to face any real problem.

In this paper, we overcome some of the limitations of the BIBFA model by introducing the following extensions: (1) We endow the model with **feature selection** capabilities. Our proposal combines the sparsity over the latent space with sparsity over the input feature space by means of a double ARD prior, providing an automatic selection of both latent factors and input features. (2) SSHIBA is able to handle **heterogeneous views** in which data views can be either real-valued vectors, binary vectors (multi-label observations), or categorical variables, widening the spectre of problems that can be faced. (3) A **semi-supervised** scheme which allows to work with unlabelled data as well as missing data. All these proposed extensions of the algorithm can be combined with each other in any way into a robust framework

named Sparse Semi-supervised Heterogeneous Inter-battery Bayesian Analysis (SSHIBA) to provide an adapted solution for any scenario according to the needs of the problem.

These proposed extensions have been analysed in terms of performance and interpretability comparing to BIBFA as well as other state-of-the-art algorithms, proving that the model is able to combine the new proposed functionalities with good performance results. An exemplary notebook, including the complete code of the proposed method, is available at https://github.com/sevisal/SSHIBA.git.

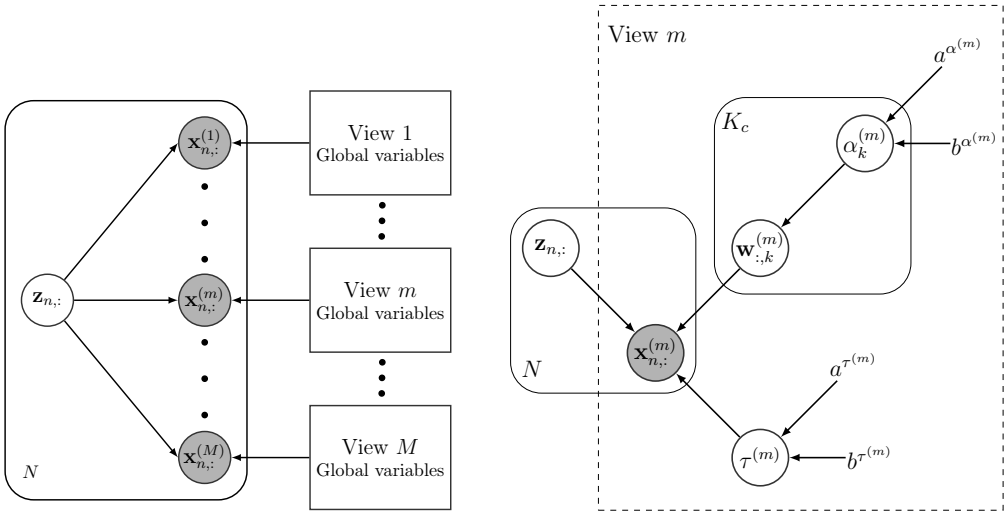
The article is organised as follows. Section 2 reviews the BIBFA algorithm presented in Klami et al. [27]. Section 3 includes a generalised formulation including the proposed extensions. This section just presents the probabilistic model and the inference learning, all mathematical development is described in the Supplementary Material. Section 4 analyses the model performance over a set of different scenarios designed to evaluate the different functionalities proposed. Finally, Section 5 gives some final remarks and conclusions.

2. Related Work: Bayesian Inter-Battery Factor Analysis

In this section we briefly review the Bayesian Inter-Battery Factor Analysis (BIBFA) model, presented in [27]. Before introducing the probabilistic formulation of this model, we first introduce the notation used. Given a matrix $\mathbf{A}$ of dimensions $I \times J$, $\mathbf{a}_{i,:}$ represents the i-th row, $\mathbf{a}_{:,j}$ represents the j-th column, and $a_{i,j}$ the i-th element of the j-th column of the matrix. In case there of multiple data views, $\mathbf{A}^{(m)}$ represents the matrix $\mathbf{A}$ of view m and $\mathbf{A}^{\{\mathcal{M}\}}$ represents all the matrices $\mathbf{A}$ of the views in the set $\mathcal{M}$.

2.1. BIBFA Generative model

The overall goal of BIBFA [27] is to jointly project different data representations, defined as "views", into a discriminative low-dimensional space. Unlike, previous FA models, BIBFA can automatically tune effective dimensionality of the projected space through automatic relevance determination (ARD) priors over the projecting matrices [28]. Assume $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{1\times D_m}$ is the m-th view of the n-th data point with $n = 1, \ldots, N$ (each view is a D_m-dimensional row vector). If $\mathcal{M} = \{1, 2, \ldots, M\}$, then $\mathbf{x}_{\mathrm{n},:}^{\{\mathcal{M}\}} = \{\mathbf{x}_{\mathrm{n},:}^{(1)}, \mathbf{x}_{\mathrm{n},:}^{(2)}, \ldots, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{M})}\}$ is the complete n-th observation. Then, the joint probability density function (pdf) of the BIBFA model can be defined as follows:



(a) Multi-view model. (b) Zoom in view m.

Figure 1: Plate diagram for the BIBFA graphical model. Gray circles denote observed variables, white circles unobserved random variables. The nodes without a circle correspond to the hyperparameters.

$$\mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(0, I_{K_c}) \tag{1}$$

$$\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})} \sim \mathcal{N}\left(0, \left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\right)^{-1} I_{K_c}\right) \tag{2}$$

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(\mathbf{z}_{\mathrm{n},:}\, \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}, \tau^{(\mathrm{m})^{-1}}\, I_{D_m}) \tag{3}$$

$$\alpha_k^{(m)} \sim \Gamma\left(a^{\alpha^{(m)}}, b^{\alpha^{(m)}}\right) \tag{4}$$

$$\tau^{(\mathrm{m})} \sim \Gamma\left(a^{\tau^{(\mathrm{m})}}, b^{\tau^{(\mathrm{m})}}\right) \tag{5}$$

where I_{K_c} is an identity matrix of dimension K_c, $\mathbf{z}_{\mathrm{n},:} \in \mathbb{R}^{1\times K_c}$ is the low-dimension latent variable for the n-th data point[1], $\Gamma(a, b)$ is a Gamma distribution with parameters a and b, $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})}$ is the k-th column of matrix $\mathbf{W}^{(\mathrm{m})}$ (of dimensions $D_m \times K_c$), and up-script (m) corresponds to the m-th view. The Gamma distribution over $\alpha_k^{(m)}$ enables the model to enforce zero values in

[1]Note we work with row-vectors.

order to maximise the model likelihood given our data. Hence, we say that (2) and (4) form an ARD prior for each of the columns of matrix $\mathbf{W}^{(\mathrm{m})}$. The BIBFA graphical model is presented in Figure 1(a). A closer look on how BIBFA models the generation of each data view is provided in Figure 1(b).

In light of the structure found in the posterior distribution of the $\mathbf{W}^{(\mathrm{m})}$ matrices, in terms of patterns of columns that are almost all zeros, one can identify common latent factors (elements of $\mathbf{z}_{\mathrm{n},:}$) across all views, specific ones only necessary to explain certain views, or irrelevant ones that are not used to explain any view. In [27], the latter are removed during inference using a threshold across all views. We adopt the same strategy, as we will later discuss.

2.2. BIBFA Variational Inference

Once the BIBFA generative model is defined, we can evaluate the posterior distribution of all the model variables given the observed data, which is unfeasible due to the intractability of computing the marginal likelihood of the data, i.e. the normalising factor in Baye's rule. In [27], the authors rely on an approximate inference approach through mean-field variational inference [29], where a lower bound to the posterior distribution is maximised, and a fully factorised variational family is chosen to approximate the posterior distribution as

$$p(\Theta|\,\mathbf{X}^{\{\mathcal{M}\}}) \;\approx\; \prod_{\mathrm{m}=1}^{\mathrm{M}} \left(q\big(\mathbf{W}^{(\mathrm{m})}\big) q\big(\tau^{(\mathrm{m})}\big) \prod_{\mathrm{k}=1}^{\mathrm{K_c}} q\Big(\alpha_{\mathrm{k}}^{(\mathrm{m})}\Big) \right) \prod_{n=1}^{N} q(\mathbf{z}_{\mathrm{n},:}) \tag{6}$$

where Θ comprises all random variables (rv) in the model.

The mean-field posterior structure along with the lower bound results in a feasible coordinate-ascent-like optimization algorithm in which the optimal maximization of each of the factors in (6) can be computed if the rest remain fixed using the following expression

$$q^*(\theta_i) \propto \mathbb{E}_{\Theta_{-i}} \left[\log p(\Theta, \mathbf{x}_{1,:}, \ldots, \mathbf{x}_{N,:}) \right], \tag{7}$$

where Θ_{-i} comprises all rv but θ_i. This new formulation is in general feasible since it does not require to completely marginalize Θ from the joint distribution.

Table 1 shows the BIBFA mean-field factor update rules derived in [27] using (7). For a compact notation, we stuck in matrix $\mathbf{Z}$, of dimension $N \times K_c$,

the latent projection of all data points and <> represents the mean value of the rv.

Variable	q^* distribution	Parameters
$\mathbf{z}_{\mathrm{n},:}$	$\mathcal{N}\left(\mathbf{z}_{\mathrm{n},:} \left\vert \mu_{\mathbf{z}_{\mathrm{n},:}}, \Sigma_{\mathbf{Z}}\right.\right)$	$\mu_{\mathbf{z}_{\mathrm{n},:}} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \langle \tau^{(\mathrm{m})} \rangle \, \mathbf{X}^{(\mathrm{m})} \langle \mathbf{W}^{(\mathrm{m})} \rangle \Sigma_{\mathbf{Z}}$ $\Sigma_{\mathbf{Z}}^{-1} = I_{K_c} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \, \mathbf{W}^{(\mathrm{m})} \rangle$
$\mathbf{W}^{(\mathrm{m})}$	$\prod_{\mathrm{d}=1}^{\mathrm{D_m}} \mathcal{N}\left(\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})} \left\vert \mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}}, \Sigma_{\mathbf{W}^{(\mathrm{m})}}\right.\right)$	$\mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}} = \langle \tau^{(\mathrm{m})} \rangle \, \mathbf{X}^{(\mathrm{m})^{\mathrm{T}}} \langle \mathbf{Z} \rangle \Sigma_{\mathbf{W}^{(\mathrm{m})}}$ $\Sigma_{\mathbf{W}^{(\mathrm{m})}}^{-1} = \mathrm{diag}(\langle \boldsymbol{\alpha}^{(\mathrm{m})} \rangle) + \langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{Z}^{\mathrm{T}} \, \mathbf{Z} \rangle$
$\alpha_{\mathrm{k}}^{(\mathrm{m})}$	$\Gamma\left(\alpha_{\mathrm{k}}^{(\mathrm{m})} \left\vert a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}, b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}\right.\right)$	$a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = \frac{D_m}{2} + a^{\boldsymbol{\alpha}^{(\mathrm{m})}}$ $b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = b^{\boldsymbol{\alpha}^{(\mathrm{m})}} + \frac{1}{2} \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \, \mathbf{W}^{(\mathrm{m})} \rangle_{k,k}$
$\tau^{(\mathrm{m})}$	$\Gamma\left(\tau^{(\mathrm{m})} \left\vert a_{\tau^{(\mathrm{m})}}, b_{\tau^{(\mathrm{m})}}\right.\right)$	$a_{\tau^{(\mathrm{m})}} = \frac{D_m N}{2} + a^{\tau^{(\mathrm{m})}}$ $b_{\tau^{(\mathrm{m})}} = b^{\tau^{(\mathrm{m})}} + \frac{1}{2} \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{d}=1}^{\mathrm{D_m}} \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})^2}$ $- \mathrm{Tr}\left\{ \langle \mathbf{W}^{(\mathrm{m})} \rangle \langle \mathbf{Z}^{\mathrm{T}} \rangle \, \mathbf{X}^{(\mathrm{m})} \right\}$ $+ \frac{1}{2} \mathrm{Tr}\left\{ \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \, \mathbf{W}^{(\mathrm{m})} \rangle \langle \mathbf{Z}^{\mathrm{T}} \, \mathbf{Z} \rangle \right\}$

Table 1: Updated q distributions for the different rv of the graphical model. These expressions have been obtained using the update rules of the mean field approximation (7). See [27] for further details.

2.3. Predictive model

The BIBFA model is also limited by the fact that it does not incorporates a semi-supervised setting, in which missing views can be properly handled. The authors rely on a training phase where the posterior distribution of the global variables of the model is computed w.r.t. complete data (i.e. no missing views), to then estimate the distribution of missing views in a test set using a predictive distribution.

Assume we use the mean field variational method to approximate the posterior distribution of the BIBFA model parameters Θ w.r.t. a fully observed training database $\mathcal{D}$, i.e. $q^*(\Theta) \approx p(\Theta|\mathcal{D})$. For a test data point $\mathbf{x}_{*,:}$

with observed views contained in the set $\mathcal{M}_{in}$ and missing views in the set $\mathcal{M}_{out}$, the BIBFA predictive model is as follows. First, the marginal posterior distribution of the latent projection $\mathbf{z}_{*,:}$ given $\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}$ is computed

$$\begin{aligned} p\big(\mathbf{z}_{*,:} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}\big) &= \int p(\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}} \,|\, \mathbf{z}_{*,:}, \Theta) p\big(\mathbf{z}_{*,:} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big) p(\Theta|\mathcal{D}) d\Theta d\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}} \\ &= \int p\big(\mathbf{z}_{*,:} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big) p(\Theta|\mathcal{D}) d\Theta, \end{aligned} \tag{8}$$

where note that the integration w.r.t. $\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}}$ is straightforward as it always integrates to one. Regarding the second term, we can either use Monte Carlo Integration by sampling from $q^*(\Theta)$ or use a point estimate for Θ (e.g. mean or mode computed from $q^*(\Theta)$). In both cases, once Θ is fixed, observe that

$$p\big(\mathbf{z}_{*,:} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big) \propto p\big(\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}} \,|\, \mathbf{z}_{*,:}, \Theta\big) p(\mathbf{z}_{*,:}), \tag{9}$$

is also Gaussian with mean $\langle \mathbf{z}_{*,:} \rangle$ and covariance matrix $\Sigma_{\mathbf{z}_{*,:}}$ given by

$$\begin{aligned} \Sigma_{\mathbf{z}_{*,:}}^{-1} &= I_{K_c} + \sum_{m \in \mathcal{M}_{in}} \Big(\tau^{(\text{m})} \, \mathbf{W}^{(\text{m})^{\text{T}}}, \mathbf{W}^{(\text{m})} \Big) \\ \langle \mathbf{z}_{*,:} \rangle &= \sum_{m \in \mathcal{M}_{in}} \big(\tau^{(\text{m})} \, \mathbf{x}_{*,:}^{(\text{m})} \, \mathbf{W}^{(\text{m})} \big) \Sigma_{\mathbf{z}_{*,:}} \end{aligned} \tag{10}$$

We can now write the expression of the distribution of the output views $\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}}$ as follows:

$$p\big(\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big) = \prod_{m \in \mathcal{M}_{out}} p\big(\mathbf{x}_{*,:}^{(\text{m})} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big), \tag{11}$$

where

$$p\big(\mathbf{x}_{*,:}^{(\text{m})} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big) = \int p\big(\mathbf{x}_{*,:}^{(\text{m})} \,|\, \mathbf{z}_{*,:}, \Theta\big) p\big(\mathbf{z}_{*,:} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\big) d\, \mathbf{z}_{*,:} \tag{12}$$

where $p\Big(\mathbf{x}_{*,:}^{(\text{m})} \,|\, \mathbf{z}_{*,:}, \Theta\Big)$ is defined in (3). Using again the properties of the Gaussian distributions we get $p\Big(\mathbf{x}_{*,:}^{(\text{m})} \,|\, \mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{in}}\}}, \Theta\Big) = \mathcal{N}\Big(\mathbf{x}_{*,:}^{(\text{m})} \,|\mu_{\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}}}, \Sigma_{\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}}}\Big)$,

where

$$\Sigma_{\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}}} = \tau^{\{\mathcal{M}_{\text{out}}\}^{-1}} I_{D_m} + \mathbf{W}^{\{\mathcal{M}_{\text{out}}\}} \Sigma_{\mathbf{z}_{*,:}} \mathbf{W}^{\{\mathcal{M}_{\text{out}}\}^{\text{T}}} \tag{13}$$

$$\mu_{\mathbf{x}_{*,:}^{\{\mathcal{M}_{\text{out}}\}}} = \mathbf{z}_{*,:} \mathbf{W}^{\{\mathcal{M}_{\text{out}}\}^{\text{T}}} \tag{14}$$

These equations complete the standard BIBFA variational model presented in [27], which works in a simple context in which the data matrices are composed of real numbers and can only work with the predictive approach. Next section presents a generalized version of this model overcoming these limitations.

3. The proposed model: SSHIBA

This section presents the Sparse Semi-supervised Heterogeneous Interbatery Bayesian Analysis (SSHIBA) method. SSHIBA generalises BIBFA in several aspects that we sequentially introduce:

1) **Feature selection**: in addition to being able to automatically select the adequate number of latent variables, by adding a double ARD prior over the matrices $\mathbf{W}^{(\text{m})}$, SSHIBA provides automatic relevant determination of both latent factors and input features for each view.

2) **Heterogeneous views**: in contrast to standard BIBFA, which only considers continuous real-valued observations, SSHIBA is able to properly incorporate binary and categorical variables. In this way, the model can handle diverse data types in the different views.

3) **Semi-supervised learning**: besides, SSHIBA provides the possibility of training the model in a semi-supervised fashion, properly handling data points with partial observations (missing views).

These proposed extensions of the method can be combined with each other in any specific way, e.g. combining a multidimensional binary view in which we want to infer some unknown values, as well as doing feature selection. Furthermore, in order to avoid hand-crafted data normalisation, the proposed generative probabilistic model also includes a bias term per view that is learned via variational inference. Namely, in the BIBFA model

above, we include the following terms:

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{z}_{\mathrm{n},:} \;\sim\; \mathcal{N}(\mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} + \boldsymbol{b}^{(\mathrm{m})}, \tau^{(\mathrm{m})^{-1}}\, I_{D_m}) \tag{15}$$

$$\boldsymbol{b}^{(\mathrm{m})} \;\sim\; \mathcal{N}(0, I_{D_m}) \tag{16}$$

In the following subsections we detail the above mentioned SSHIBA features. The mathematical derivations of the variational machinery in each case have been moved to the Supplementary Material.

3.1. Feature selection in the SSHIBA model

For this first extension of the method, we propose to redefine the priors of matrix $\mathbf{W}^{(\mathrm{m})}$ so that it is able to automatically select both the relevant latent factors and the relevant input features that are used by the model.

3.1.1. Generative model for feature selection

We propose a double ARD prior over the $\mathbf{W}^{(\mathrm{m})}$ matrices, with a different prior for each entry of $\mathbf{W}^{(\mathrm{m})}$:

$$\mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \;\sim\; \mathcal{N}\left(0, \left(\gamma_{\mathrm{d}}^{(\mathrm{m})}\,\alpha_{\mathrm{k}}^{(\mathrm{m})}\right)^{-1}\right) \tag{17}$$

$$\gamma_{\mathrm{d}}^{(\mathrm{m})} \;\sim\; \Gamma\left(a^{\gamma^{(\mathrm{m})}}, b^{\gamma^{(\mathrm{m})}}\right) \tag{18}$$

Note that the variance of $\mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})}$ is the product of two variables: a row-wise prior over $\mathbf{W}^{(\mathrm{m})}$, i.e. $\alpha_{\mathrm{k}}^{(\mathrm{m})}$, which was already present in the BIBFA model and is used to perform latent variable selection, and a column-wise prior over $\mathbf{W}^{(\mathrm{m})}$, i.e. $\gamma_{\mathrm{d}}^{(\mathrm{m})}$ which induces sparsity along the elements of such columns, allowing interpretable results by means of feature selection. With the product in (17), we provide the model with the flexibility to find the structural sparsity patterns in $\mathbf{W}^{(\mathrm{m})}$ that maximise the evidence. Figure 2 shows the graphical model of SSHIBA (assuming still real-valued observations).

3.1.2. Variational inference

When we augment the BIBFA model presented in Section 2.1 with the double ARD method summarized by equations (17) and (18), we equivalently

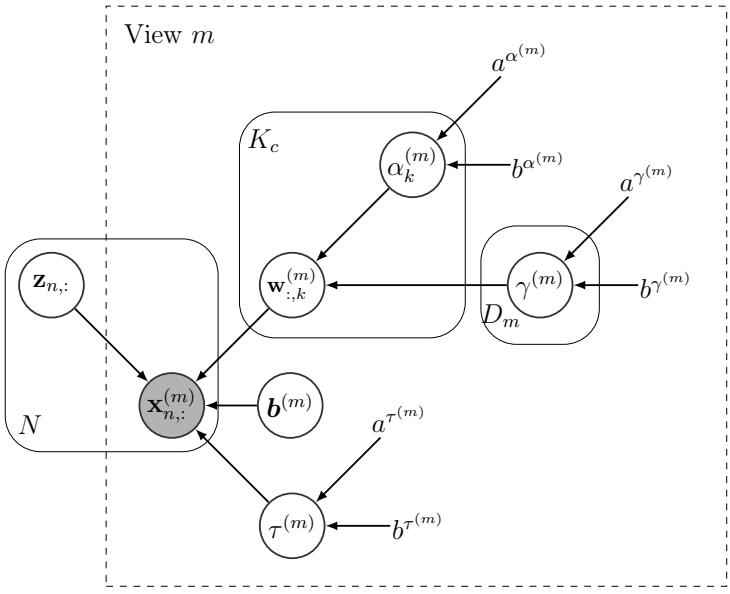


Figure 2: SSHIBA's feature selection graphical model.

need to expand accordingly the mean-field posterior distribution, namely

$$p(\Theta|\,\mathbf{X}^{\{\mathcal{M}\}}) \approx \prod_{\mathrm{m}=1}^{\mathrm{M}} \left(q\left(\mathbf{W}^{(\mathrm{m})}\right) q\left(\boldsymbol{b}^{(\mathrm{m})}\right) q\left(\tau^{(\mathrm{m})}\right) \prod_{\mathrm{k}=1}^{\mathrm{K_c}} q\left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\right) \prod_{\mathrm{d}=1}^{\mathrm{D_m}} q\left(\gamma_{\mathrm{d}}^{(\mathrm{m})}\right) \right) \prod_{n=1}^{N} q(\mathbf{z}_{\mathrm{n},:}). \quad (19)$$

Table 2 shows the update rules obtained by applying the mean-field update rule in (7) to this new model. A detailed calculation of these expressions can be found in the Supplementary Material.

Since variable $\boldsymbol{\gamma}^{(\mathrm{m})}$ provides a measure of importance for each feature (higher $\boldsymbol{\gamma}^{(\mathrm{m})}$, lower importance), the model is now capable of providing a measure of the relevance of each feature. In other words, this version allows the model to provide an online feature ranking or feature selection for any input data, improving the interpretability of the results.

Finally, note that once global rv are sampled from $q^*(\Theta)$, the predictive model remains the same w.r.t. the BIBFA predictive model in Section 2.3.

Variable	q^* distribution	Parameters
$\mathbf{z}_{\mathrm{n},:}$	$\mathcal{N}\left(\mathbf{z}_{\mathrm{n},:} \mid \mu_{\mathbf{z}_{\mathrm{n},:}}, \Sigma_{\mathbf{Z}}\right)$	$\mu_{\mathbf{z}_{\mathrm{n},:}} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\langle \tau^{(\mathrm{m})} \rangle \left(\mathbf{X}^{(\mathrm{m})} - \mathbb{1}_N \langle \boldsymbol{b}^{(\mathrm{m})} \rangle \right) \langle \mathbf{W}^{(\mathrm{m})} \rangle \right) \Sigma_{\mathbf{Z}}$ $\Sigma_{\mathbf{Z}}^{-1} = I_{K_c} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \mathbf{W}^{(\mathrm{m})} \rangle$
$\mathbf{W}^{(\mathrm{m})}$	$\prod_{\mathrm{d}=1}^{\mathrm{D_m}} \mathcal{N}\left(\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})} \mid \mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}}, \Sigma_{W_d^{(m)}}\right)$	$\mu_{\mathbf{W}^{(\mathrm{m})}} = \langle \tau^{(\mathrm{m})} \rangle \left(\mathbf{X}^{(\mathrm{m})} - \mathbb{1}_N \langle \boldsymbol{b}^{(\mathrm{m})} \rangle \right)^T \langle \mathbf{Z} \rangle \Sigma_{\mathbf{W}^{(\mathrm{m})}}$ $\Sigma_{W_d^{(m)}}^{-1} = \mathrm{diag}(\langle \boldsymbol{\alpha}^{(\mathrm{m})} \rangle) \langle \gamma_{\mathrm{d}}^{(\mathrm{m})} \rangle + \langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{Z}^{\mathrm{T}} \mathbf{Z} \rangle$
$\boldsymbol{b}^{(\mathrm{m})}$	$\mathcal{N}\left(\boldsymbol{b}^{(\mathrm{m})} \mid \mu_{\boldsymbol{b}^{(\mathrm{m})}}, \Sigma_{\boldsymbol{b}^{(\mathrm{m})}}\right)$	$\mu_{\boldsymbol{b}^{(\mathrm{m})}} = \langle \tau^{(\mathrm{m})} \rangle \sum_{\mathrm{n}=1}^{\mathrm{N}} \left(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} - \langle \mathbf{z}_{\mathrm{n},:} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \rangle \right) \Sigma_{\boldsymbol{b}^{(\mathrm{m})}}$ $\Sigma_{\boldsymbol{b}^{(\mathrm{m})}}^{-1} = \left(N \langle \tau^{(\mathrm{m})} \rangle + 1 \right) I_{D_m}$
$\boldsymbol{\alpha}^{(\mathrm{m})}$	$\prod_{\mathrm{k}=1}^{\mathrm{K_c}} \Gamma\left(\alpha_{\mathrm{k}}^{(\mathrm{m})} \mid a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}, b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}\right)$	$a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = \frac{D_m}{2} + a^{\boldsymbol{\alpha}^{(\mathrm{m})}}$ $b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = b^{\boldsymbol{\alpha}^{(\mathrm{m})}} + \frac{1}{2} \sum_{\mathrm{d}=1}^{\mathrm{D_m}} \langle \gamma_{\mathrm{d}}^{(\mathrm{m})} \rangle \langle \mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \, \mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \rangle$
$\tau^{(\mathrm{m})}$	$\Gamma\left(\tau^{(\mathrm{m})} \mid a_{\tau^{(\mathrm{m})}}, b_{\tau^{(\mathrm{m})}}\right)$	$a_{\tau^{(\mathrm{m})}} = \frac{D_m N}{2} + a^{\tau^{(\mathrm{m})}}$ $b_{\tau^{(\mathrm{m})}} = b^{\tau^{(\mathrm{m})}} + \frac{1}{2} \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{d}=1}^{\mathrm{D_m}} \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})^2}$ $- \mathrm{Tr}\left\{ \langle \mathbf{W}^{(\mathrm{m})} \rangle \langle \mathbf{Z}^{\mathrm{T}} \rangle \mathbf{X}^{(\mathrm{m})} \right\} + \frac{1}{2} \mathrm{Tr}\left\{ \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \mathbf{W}^{(\mathrm{m})} \rangle \langle \mathbf{Z}^{\mathrm{T}} \mathbf{Z} \rangle \right\}$ $- \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \langle \boldsymbol{b}^{(\mathrm{m})^{\mathrm{T}}} \rangle + \sum_{\mathrm{n}=1}^{\mathrm{N}} \langle \mathbf{z}_{\mathrm{n},:} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \rangle \langle \boldsymbol{b}^{(\mathrm{m})^{\mathrm{T}}} \rangle + \frac{N}{2} \langle \boldsymbol{b}^{(\mathrm{m})} \, \boldsymbol{b}^{(\mathrm{m})^{\mathrm{T}}} \rangle$
$\boldsymbol{\gamma}^{(\mathrm{m})}$	$\prod_{\mathrm{d}=1}^{\mathrm{D_m}} \Gamma\left(\gamma_{\mathrm{d}}^{(\mathrm{m})} \mid a_{\gamma_{\mathrm{d}}^{(\mathrm{m})}}, b_{\gamma_{\mathrm{d}}^{(\mathrm{m})}}\right)$	$a_{\gamma_{\mathrm{d}}^{(\mathrm{m})}} = \frac{K_c}{2} + a^{\boldsymbol{\gamma}^{(\mathrm{m})}}$ $b_{\gamma_{\mathrm{d}}^{(\mathrm{m})}} = b^{\boldsymbol{\gamma}^{(\mathrm{m})}} + \frac{1}{2} \sum_{\mathrm{k}=1}^{\mathrm{K_c}} \langle \alpha_{\mathrm{k}}^{(\mathrm{m})} \rangle \langle \mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \, \mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \rangle$

Table 2: Distribution q of the different rv of the graphical model for feature selection together with the different distribution parameters. Where $\mathbb{1}_N$ is a row vector of ones of dimension N.

3.2. Heterogeneous data: Multidimensional binary views

This section introduces another model extension, in this case, to model any of data views as a multidimensional binary observation. For example, this extension can be used to model the output view of a multi-label classification problem.

3.2.1. Generative model

To accommodate the model for binary views, we incorporate the Bayesian logistic regression model presented in [30], as it is summarised in the graphical model of Figure 3. $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ is now unobserved but still keeps the same conditional distribution (15); i.e. $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ is still a D_m-real valued vector following a Gaussian distribution given $\mathbf{z}_{\mathrm{n},:}$. Furthermore, we introduce a new

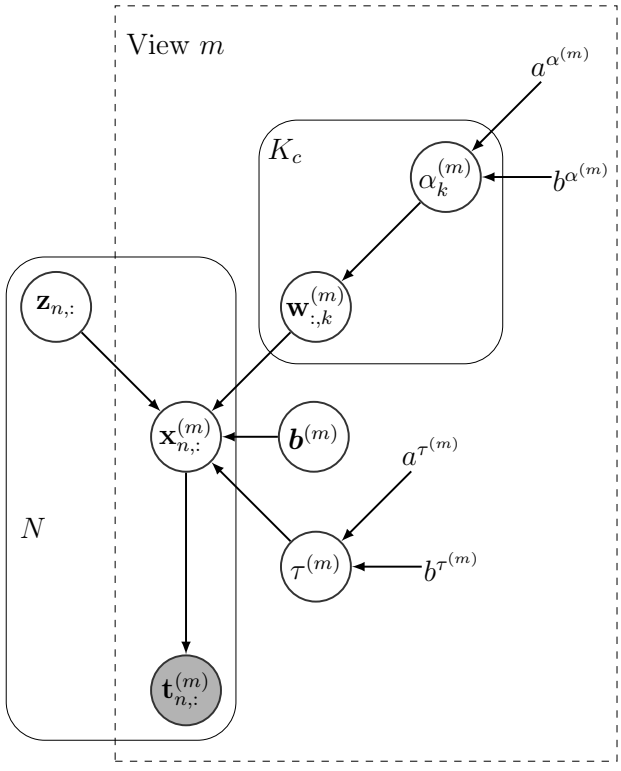


Figure 3: SSHIBA graphical model for multi-dimensional binary views.

observed variable binary vector $\mathbf{t}_{\mathrm{n},:}^{(\mathrm{m})}$, also of dimension D_m, whose conditional distribution given $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ is a product of logistic regression terms

$$p\big(\mathbf{t}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}\big) = \prod_{d=1}^{D_m} p\Big(\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})} \,|\, \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big) \tag{20}$$

$$p\Big(\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})} \,|\, \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big) = \sigma\Big(\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big)^{\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})}} \Big(1 - \sigma(\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})})\Big)^{1-\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})}} = e^{\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\,\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})}} \sigma\Big(-\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big), \tag{21}$$

where $\sigma(a) = (1 + e^{-a})^{-1}$. Following [30], to develop the variational machinery, we will use the following lower bound on the logistic regression condi-

tional probability

$$p\Big(\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})} \,|\, \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big) \;=\; e^{\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\,\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})}}\sigma\Big(-\,\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big) \geq e^{\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\,\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})}}\sigma\Big(\xi_{\mathrm{n,d}}^{(\mathrm{m})}\Big)e^{-\frac{\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}+\xi_{n,d}^{(m)}}{2}-\lambda\left(\xi_{n,d}^{(m)}\right)\left(\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})^2}-\xi_{\mathrm{n,d}}^{(\mathrm{m})^2}\right)} \tag{22}$$

where $\lambda(a) = \frac{1}{2a}\big(\sigma(a) - \frac{1}{2}\big)$, and $\xi_{\mathrm{n,d}}^{(\mathrm{m})}$ is a variational parameter optimized by maximizing the evidence lower bound as shown in Section B of the Supplementary Material. Using (22), we can lower bound $p\big(\mathbf{T}^{(\mathrm{m})} \,|\, \mathbf{X}^{(\mathrm{m})}\big)$ as follows

$$p\big(\mathbf{T}^{(\mathrm{m})} \,|\, \mathbf{X}^{(\mathrm{m})}\big) \geq h\big(\mathbf{X}^{(\mathrm{m})}, \boldsymbol{\xi}\big) = \prod_{n=1}^{N}\prod_{d=1}^{D_m}\left(\sigma\Big(\xi_{\mathrm{n,d}}^{(\mathrm{m})}\Big)e^{\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\,\mathrm{t}_{\mathrm{n,d}}^{(\mathrm{m})}-\frac{\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}+\xi_{\mathrm{n,d}}^{(\mathrm{m})}}{2}-\lambda\left(\xi_{\mathrm{n,d}}^{(\mathrm{m})}\right)\left(\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})^2}-\xi_{\mathrm{n,d}}^{(\mathrm{m})^2}\right)}\right). \tag{23}$$

3.2.2. Variational inference

Given the graphical model in Figure 3, the mean-field variational family is as follows

$$p\big(\Theta|\,\mathbf{T}^{\{\mathcal{M}_{\mathrm{t}}\}}, \mathbf{X}^{\{\mathcal{M}_r\}}\big) \;\approx\; q(\mathbf{Z})\prod_{m_t\in\mathcal{M}_t}\left(\prod_{n=1}^{N} q\big(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\big)\right)\prod_{\mathrm{m}=1}^{\mathrm{M}} q\big(\mathbf{W}^{(\mathrm{m})}\big)q\Big(\boldsymbol{b}^{(\mathrm{m})}\Big)q\big(\boldsymbol{\alpha}^{(\mathrm{m})}\big)q\big(\tau^{(\mathrm{m})}\big)q\big(\boldsymbol{\gamma}^{(\mathrm{m})}\big), \tag{24}$$

where $\mathcal{M}_t$ is the set of views in which we want to have the multidimensional binary data and $\mathcal{M}_r$ are the rest of the views. The details about the variational updates can be found in Appendix B. Note that, conditioned to a fixed $\mathbf{X}^{(\mathrm{m_t})}$, the model is equivalent to the case of real-valued observations and, hence, most of the mean-field updates remain almost the same. We only have to replace in Table 1 and 2 $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}$ (or the stacked data matrix $\mathbf{X}^{(\mathrm{m})^{\mathrm{T}}}$) by its mean $\langle\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\rangle$ ($\langle\mathbf{X}^{(\mathrm{m})^{\mathrm{T}}}\rangle$) w.r.t. $q\Big(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\Big)$ for each data point. Regarding this latter term, the variational update-rule is given in Table 3.

Unlike the predictive distribution when only real views are implemented, the predictive distribution in SSHIBA with multi-dimensional binary observations requires approximate inference (e.g. variational inference or Monte Carlo) to estimate the posterior latent distribution in (8) w.r.t. to the ob-

Variable	q distribution	Parameters
$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}$	$\mathcal{N}\Big(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})} \mid \mu_{\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}}, \Sigma_{\mathbf{X}^{(\mathrm{m_t})}}\Big)$	$\mu_{\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}} = \Big(\mathrm{t}_{\mathrm{n},:}^{(\mathrm{m_t})} - \frac{1}{2} + \langle\tau^{(\mathrm{m_t})}\rangle\langle\mathbf{z}_{\mathrm{n},:}\rangle\langle\mathbf{W}^{(\mathrm{m_t})^{\mathrm{T}}}\rangle + \langle\boldsymbol{b}^{(\mathrm{m_t})}\rangle\Big)\Sigma_{\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}}$ $\Sigma_{\mathbf{X}^{(\mathrm{m_t})}}^{-1} = \langle\tau^{(\mathrm{m_t})}\rangle I + 2\Lambda_{\boldsymbol{\xi}_{\mathrm{n},:}^{(\mathrm{m_t})}}$

Table 3: Mean-field update rule for the $q\Big(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\Big)$ distribution in (24), where $\Lambda_{\boldsymbol{\xi}_{n,:}}$ is a diagonal matrix for which the diagonal elements are $\lambda(\xi_{n,1}), \lambda(\xi_{n,2}), \ldots, \lambda(\xi_{n,D_m})$. This distribution only affects the views modelled as multidimensional binary data.

served data. This case can be directly reformulated from the semi-supervised SSHIBA model presented in Section 3.4, and hence we omit it from here.

3.3. Heterogeneous data: Categorical observations

This section presents how SSHIBA works with categorical observations.

3.3.1. Generative model

We incorporate the multinomial probit in [31]. In this case, the structure is similar to the one followed by the multidimensional binary case of Figure 3 but, in the categorical case, $\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})}$ (assuming that the m-th view corresponds to a categorical variable) is an integer scalar that takes values in the set $\{0, \ldots, D_m - 1\}$, being D_m the number of classes. The multinomial probit relates $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ with $\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})}$ as follows:

$$\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})} = i \qquad \text{if} \qquad \mathrm{x}_{\mathrm{n,i}}^{(\mathrm{m})} = \max_{1 \leq \mathrm{d} \leq \mathrm{D_m}} \Big(\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})}\Big). \tag{25}$$

If we set the noise parameter $\tau^{(\mathrm{m})} = 1$, in [31] it is shown that we can express $p\Big(\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})} = i \mid \mathbf{z}_{\mathrm{n},:}, \mathbf{W}^{(\mathrm{m})}\Big)$ as follows:

$$p\big(\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})} = i \mid \mathbf{z}_{\mathrm{n},:}, \mathbf{W}^{(\mathrm{m})}\big) = \mathbb{E}_{p(u)}\left[\prod_{j \neq i}\Big(\Phi\Big(u + \mathbf{y}_{\mathrm{n,i}}^{(\mathrm{m})} - \mathbf{y}_{\mathrm{n,j}}^{(\mathrm{m})}\Big)\Big)\right] \tag{26}$$

where $\mathbf{y}_{\mathrm{n},:}^{(\mathrm{m})} = \mathbf{z}_{\mathrm{n},:}\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}$, $p(u) \sim \mathcal{N}(0,1)$, and $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function (cdf). Expectations w.r.t. $p(u)$ can be effectively approximated using Monte Carlo, as they only require sampling from an uni-dimensional standard Gaussian.

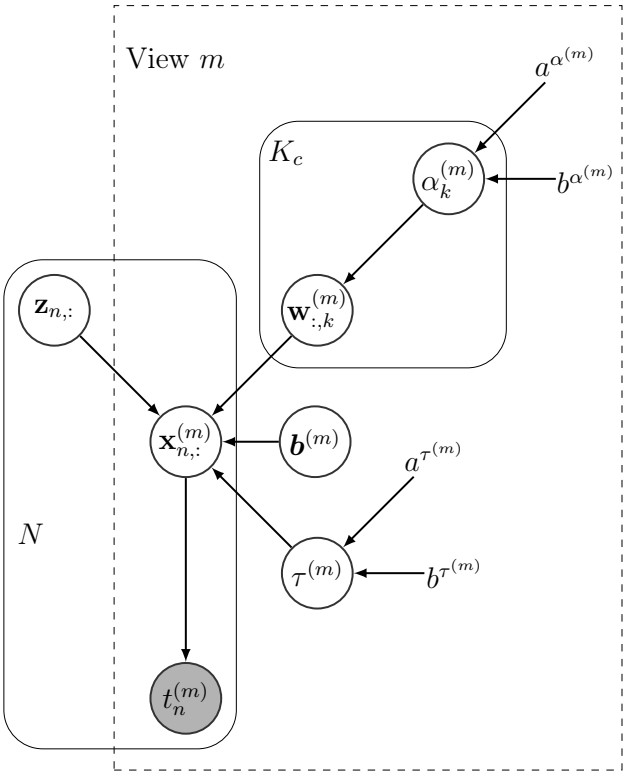


Figure 4: SSHIBA graphical model for categorical views.

3.3.2. Variational inference

Deriving mean-field update for the categorical views closely follows the methodology in [31], so we omit further details from here. Given the mean-field variational family in (24) (assuming now that $\mathcal{M}_t$ is the set of views that correspond to categorical observations), the mean-field update of the term $q\left(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\right)$ is summarized in Table 3. The mean-field update for the rest of the terms are provided in previous sections (as in the multi-dimensional binary case, we replace $\mathbf{X}^{(\mathrm{m})^{\mathrm{T}}}$ by $\langle \mathbf{X}^{(\mathrm{m})^{\mathrm{T}}} \rangle$). Observe that, given $\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})}$, $q\left(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\right)$ corresponds to a truncated Gaussian distribution. Again, we note that a predictive model will be easily formulated from the semi-supervised case presented in the next subsection.

Variable	q distribution	Parameters
$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}$	$\frac{1}{\boldsymbol{\xi}_{n,:}}\mathcal{N}\Big(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\,\vert\langle\mathbf{y}_{\mathrm{n},:}^{(\mathrm{m_t})}\rangle, I\Big)\times$ $\delta\Big(\mathrm{x}_{\mathrm{n,i}}^{(\mathrm{m_t})} > \mathrm{x}_{\mathrm{n,j}}^{(\mathrm{m_t})}\,\forall i \neq j\Big)$	$\langle\mathrm{x}_{\mathrm{n,i}}^{(\mathrm{m_t})}\rangle = \langle\mathbf{y}_{\mathrm{n,i}}^{(\mathrm{m_t})}\rangle - \sum_{j\neq i}\Big(\langle\mathbf{y}_{\mathrm{n,j}}^{(\mathrm{m_t})}\rangle - \langle\mathrm{x}_{\mathrm{n,j}}^{(\mathrm{m_t})}\rangle\Big)$ $\langle\mathrm{x}_{\mathrm{n,j}}^{(\mathrm{m_t})}\rangle = \langle\mathbf{y}_{\mathrm{n,j}}^{(\mathrm{m_t})}\rangle - \frac{1}{\boldsymbol{\xi}_{n,:}}\mathbb{E}_{p(u)}\Big[\mathcal{N}_u\Big(\langle\mathbf{y}_{\mathrm{n,j}}^{(\mathrm{m_t})}\rangle - \langle\mathbf{y}_{\mathrm{n,i}}^{(\mathrm{m_t})}\rangle, 1\Big)$ $\prod_{k\neq i\neq j}\Big(\Phi\Big(u + \langle\mathbf{y}_{\mathrm{n,i}}^{(\mathrm{m_t})}\rangle - \langle\mathbf{y}_{\mathrm{n,k}}^{(\mathrm{m_t})}\rangle\Big)\Big)\Big]$

Table 4: q distribution of the different rv of the graphical model for the categorical scheme, where $\langle\mathbf{y}_{\mathrm{n},:}^{(\mathrm{m_t})}\rangle = \langle\mathbf{z}_{\mathrm{n},:}\rangle\langle\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}\rangle + \langle\boldsymbol{b}^{(\mathrm{m})}\rangle$ and $\boldsymbol{\xi}_{n,:} = \mathbb{E}_{p(u)}\Big[\prod_{j\neq i}\Big(\Phi\Big(u + \langle\mathbf{y}_{\mathrm{n,i}}^{(\mathrm{m_t})}\rangle - \langle\mathbf{y}_{\mathrm{n,j}}^{(\mathrm{m_t})}\rangle\Big)\Big)\Big]$ and assuming that $\mathrm{t}_{\mathrm{n}}^{(\mathrm{m})} = i$. This distribution only affects the views modelled as categorical data.

3.4. Semi-supervised SSHIBA

The last main contribution of the paper is to show how missing-views can be incorporated into SSHIBA training (e.g. variational inference) following an unsupervised fashion, in which there is no need for a predictive distribution since both "training" and "test" data are jointly fused by the model, which simply considers as unobserved both the views in the test data that we aim at predicting and the missing values in both "training" and "test" sets.

In the case the m-th view corresponds to a real-variable, we denote by $\tilde{\mathbf{X}}^{(\mathrm{m})}$ (in contrast to $\mathbf{X}^{(\mathrm{m})}$) to the set of data points for which this view is missing. Similarly, if the m-th corresponds to a multi-dimensional binary variable or categorical variable, the set of data points for which this view is missing is denoted by $\tilde{\mathbf{T}}^{(\mathrm{m})}$ (in contrast to $\mathbf{T}^{(\mathrm{m})}$). Note that the SSHIBA graphical model summarized in Figures 2, 3, and 4 remains unaltered, we simply have white dots instead of grey dots for those data points for which the corresponding view is unobserved.

3.4.1. Variational inference

Missing views are handled as any other rv in the model and hence during variational inference our goal is now to approximate the joint posterior distribution of the parameters of the model Θ and the missing data views ($\tilde{\mathbf{X}}^{(\mathrm{m})}$ or $\tilde{\mathbf{T}}^{(\mathrm{m})}$). Following the mean-field method, we again assume a variational family that factorizes across all elements in Θ and all data points in $\tilde{\mathbf{X}}^{(\mathrm{m})}$ or

$\tilde{\mathbf{T}}^{(\mathrm{m})}$:

$$p(\Theta, \tilde{\mathbf{T}}^{\{\mathcal{M}_\mathrm{t}\}}, \tilde{\mathbf{X}}^{\{\mathcal{M}_\mathrm{r}\}} | \, \mathbf{T}^{\{\mathcal{M}_\mathrm{t}\}}, \mathbf{X}^{\{\mathcal{M}_\mathrm{r}\}}) \approx q(\mathbf{Z}) \prod_{m_t \in \mathcal{M}_t} \left(q(\tilde{\mathbf{T}}^{\{\mathcal{M}_\mathrm{t}\}}) \prod_{n=1}^{N} q\big(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m_t})}\big) \right) \times \prod_{m_r \in \mathcal{M}_r} q(\tilde{\mathbf{X}}^{\{\mathcal{M}_\mathrm{r}\}}) \prod_{\mathrm{m}=1}^{\mathrm{M}} \big(q\big(\mathbf{W}^{(\mathrm{m})}\big) q\big(\boldsymbol{\alpha}^{(\mathrm{m})}\big) q\big(\tau^{(\mathrm{m})}\big) q\big(\boldsymbol{\gamma}^{(\mathrm{m})}\big) \big). \tag{27}$$

The mean-field update for all factors in (27) be found in Appendix C and the final distributions are shown in in Table 5.

Version	Variable	q distribution	Parameters
Regression	$\tilde{\mathbf{X}}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}} \mathcal{N}\Big(\mathbf{x}_{*,:}^{(\mathrm{m})} \, \vert \mu_{\mathbf{x}_{*,:}^{(\mathrm{m})}}, \Sigma_{\tilde{\mathbf{X}}^{(\mathrm{m})}}\Big)$	$\mu_{\tilde{\mathbf{X}}^{(\mathrm{m})}} = \langle \tilde{\mathbf{Z}} \rangle \langle \mathbf{W}^{(\mathrm{m})} \rangle^T$ $\Sigma_{\tilde{\mathbf{X}}^{(\mathrm{m})}} = \langle \tau^{(\mathrm{m})} \rangle^{-1} I_{D_m}$
Multidimensional	$\tilde{\mathbf{T}}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}} \mathcal{N}\Big(\tilde{\mathbf{t}}_{\mathrm{n},:}^{(\mathrm{m})} \, \vert \langle \tilde{\mathbf{t}}_{\mathrm{n},:}^{(\mathrm{m})} \rangle, \Sigma_{\tilde{\mathbf{T}}^{(\mathrm{m})}}\Big)$	$\mu_{\tilde{\mathbf{t}}_{\mathrm{n},:}^{(\mathrm{m})}} = \sigma\Big(\langle \tilde{\mathbf{X}}^{(\mathrm{m})} \rangle\Big)$ $\Sigma_{\tilde{\mathbf{T}}^{(\mathrm{m})}} = \frac{e^{\langle \tilde{\mathbf{X}}^{(\mathrm{m})} \rangle}}{\Big(1+e^{\langle \tilde{\mathbf{X}}^{(\mathrm{m})} \rangle}\Big)^2}$
Categorical	$\tilde{\mathbf{t}}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}} \mathcal{N}\Big(\tilde{\mathrm{t}}_{\mathrm{n}}^{(\mathrm{m})} \, \vert \langle \tilde{\mathrm{t}}_{\mathrm{n}}^{(\mathrm{m})} \rangle, \Sigma_{\tilde{\mathbf{t}}^{(\mathrm{m})}}\Big)$	$\langle \tilde{\mathrm{t}}_{\mathrm{n}}^{(\mathrm{m})} \rangle = \langle \tilde{\mathbf{y}}_{\mathrm{n,j}}^{(\mathrm{m_t})} \rangle - \frac{1}{\xi_{n,:}} \mathbb{E}_{p(u)} \big[\mathcal{N}_u \Big(\langle \tilde{\mathbf{y}}_{\mathrm{n,j}}^{(\mathrm{m_t})} \rangle - \langle \tilde{\mathbf{y}}_{\mathrm{n,i}}^{(\mathrm{m_t})} \rangle, 1 \Big)$ $\prod_{k \neq i \neq j} \Big(\Phi \Big(u + \langle \tilde{\mathbf{y}}_{\mathrm{n,i}}^{(\mathrm{m_t})} \rangle - \langle \tilde{\mathbf{y}}_{\mathrm{n,k}}^{(\mathrm{m_t})} \rangle \Big) \Big) \big]$

Table 5: q distribution of the different rv of the graphical model for the semi-supervised scheme. The table shows what are the different parameters of the distributions. The first parameter is the mean and the second one is the variance. Where $\langle \tilde{\mathbf{y}}_{\mathrm{n},:}^{(\mathrm{m_t})} \rangle = \langle \tilde{\mathbf{z}}_{\mathrm{n},:} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \rangle + \langle \boldsymbol{b}^{(\mathrm{m})} \rangle$.

4. Results

In this section we present experimental results that demonstrate the ability of SSHIBA to capture the statistical properties of real databases, while comparing it with some state-of-the-art algorithms. The implementation of this project was done using *Python 3.7* and the different baselines where implemented using packages from *Scikit-learn* [32].

Regarding SSHIBA, in all experiments we implement automatic latent factor selection, also referred as pruning. For this purpose, during the inference learning we remove the k-th column of $\mathbf{W}^{(\mathrm{m})}$, $\forall m$, if all the elements of $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})}$, across all the views, are lower than the pruning threshold set to 10^{-6}.

To determine the number of iterations of the inference process, we used a convergence criteria based on the evolution of the lower bound. In particular, we stop the algorithm either when $LB[-2] > LB[-1](1 - 10^{-8})$, where $LB[-1]$ is the lower bound at the last iteration and $LB[-2]$ at the previous one, or when it reaches $5 * 10^4$ iterations. SSHIBA is trained 10 times with random initialisation and we kept the model with the best lower bound.

4.1. Database description

As the presented model is appealing in a wide range of contexts, we included several databases of different nature (different sizes, dimensions, types of variables, ...) to demonstrate its potential.

First of all, we used three multi-label databases from different domains available in the Mulan repository [33]: *yeast* is a genetic database [34], *scene* is a landscape image database [35] and *birds* is an audio recordings database [36]. For these three databases, we stuck the multi-label information in one single view and the rest of numeric features in another view. We also worked with the a database that includes categorical observations, the *AVIRIS* database [37] is composed of 220 Band Hyperspectral Image of agronome farms, where we used the categorical labels as one view and the rest of numeric features as another

Additionally, we use the Labeled Faces in the Wild (LFW) dataset [38] consisting of face photographs of different people. We used a version of the dataset with aligned faces obtained by [39] to work with images under the same conditions. At the same time, the images have been cropped to eliminate undesirable information and resized to reduce the computational cost of training the models, having images of 60×40 pixels. Once the images were processed we decided to work with two different problems:

- ⋄ **Face recognition**: It consists in identifying the person, between the 7 people with most images in the dataset, to whom the image corresponds. We will refer to this version as LFW.

- ⋄ **Multi-label attributes**: Here we need to determine whether an image has certain attributes or not. These attributes, obtained by [40], correspond to different physical information related to the people in the photographs, such as gender, hair colour or wearing glasses. We will refer to this version as LFWA.

Database	Domain	Samples	Features	Labels
yeast	genes	2,417	103	14
scene	landscapes images	2,407	294	6
AVIRIS	hyperspectral images	21,025	220	16
birds	audio	645	260	19
LFW	faces images	1,277	2,400	7
LFWA	faces images	22,343	2,400	73

Table 6: Summary of the main characteristic of the databases used in this work.

The characteristics of all the databases mentioned above are summarized in Table 6.

We used training and test partitions to train the model and measure the performance respectively. In particular, both the *scene* and *yeast* databases are already divided into train and test sets, around a 50% and 60% train data respectively. In the case of the *LFW* databases as well as *AVIRIS*, both were split using 70% train / 30% test partitions.

4.2. Baseline or state-of-art methods

To analyse the different versions of the method in comparison to some contextual results, we decided to include some state-of-the-art algorithms to obtain reference scoring. In particular, we have used the following methods:

- Canonical Correlation Analysis (**CCA**) is a supervised feature extraction method which finds a latent space for the data. Due to the parallelisms with our method, we decided to used this algorithm as one of the baselines to compare to.

- Principal Component Analysis (**PCA**) is a non-supervised feature extraction algorithm that we decided to combine with Logistic Regression to carry out feature extraction and predictions.

- As all the problems to solve involve classification tasks, we have included Logistic regression (**LR**) as a state-of-the-art method widely used as a classifier.

- To compare our results to those of a neural network, we used a Multi-Layer Perceptron (**MLP**) with one hidden layer.

- We also included the base method presented in [27], **BIBFA**. As they indicate in the paper, we added a final thresholding process to obtain a label prediction.

10 folds Cross-Validation (CV) was used to adjust the regularization parameter for the logistic regression, MLP and ridge regression. The number of latent factors (K_c) of the PCA has been set to those who explain 95% of the variance and for CCA K_c is $C - 1$ (where C is the number of classes).

Finally, as we work with both multi-label and multiclass datasets, we decided to use the balanced Multiclass Area Under the Curve (AUC) metric to compare the performance of the different methods. It is calculated as $AUC_{mc} = \frac{1}{N}\sum_c(N_c \times AUC_c)$, where N is the total number of samples, N_c is the number of samples of class c and AUC_c is the AUC of class c with respect to the rest of the classes.

4.3. SSHIBA for heterogeneous prediction

In this first set of experiments, we use *yeast*, *scene*, *birds* (multi-label), and *AVIRIS* (categorical) to test the ability of SSHIBA to perform prediction over a multi-label/categorical view. To carry out the estimation of the test labels we used the standard predictive approach described in Section 2.3. Furthermore, all these results have been calculated using the complete dataset (Table 7), as well as a reduced version consisting of a 20% of the original data (Table 8). For the reduced version, we use the iterative stratifier presented in [41] to have splits with the minority categories properly represented.

In Table 7 we can see the performance and the number of latent factors obtained with the different databases using all the available data. The results provide an insight on the method, where we can see that the algorithm is capable of providing a dimensionality reduction of the input features while maintaining the prediction performance compared to the rest of the discriminative approaches as well as extra capabilities, as we demonstrate in the rest of experiments (feature selection, missing data imputation, multi-view learning). Furthermore, we observed that for the *AVIRIS* database we obtained an AUC improvement of *0.01* when treating the data as categorical w.r.t. binarizing the labels and using the multi-label version. Furthermore, we can see that the inclusion of a more restrictive pruning criteria greatly reduces the number of latent factors, while maintaining the performance in terms of AUC.

	yeast		*scene*		*AVIRIS*		*birds*	
	AUC	K_c	AUC	K_c	AUC	K_c	AUC	K_c
SSHIBA	0.66	66	**0.92**	137	**0.89**	197	0.83	75
SSHIBA (lax prun.)	0.66	19	**0.92**	40	0.88	75	**0.84**	69
BIBFA	**0.69**	66	0.90	119	**0.89**	197	0.67	10
CCA	0.61	13	0.88	5	0.88	72	0.56	18
CCA + LR	0.66	13	0.87	5	**0.89**	72	0.56	18
PCA + LR	0.68	73	**0.92**	121	0.81	252	0.82	87
MLP	0.61	300	0.82	900	0.85	50	0.68	100
LR	0.67	-	**0.92**	-	**0.89**	-	0.81	-
RR	0.68	-	0.91	-	**0.89**	-	0.83	-

Table 7: Results of the predictive SSHIBA and the different methods under study on multi-label and categorical databases. Results include the performance in terms of AUC and the number K_c of latent factors. We also included a version of SSHIBA with a less restrictive pruning criteria (lax pruning) to analyse its effect on the number of latent factors.

	yeast		*scene*		*AVIRIS*		*birds*	
	AUC	K_c	AUC	K_c	AUC	K_c	AUC	K_c
SSHIBA	$\mathbf{0.65 \pm 0.01}$	20 ± 2	$\mathbf{0.90 \pm 0.01}$	128 ± 2	0.87 ± 0.01	78 ± 82	$\mathbf{0.66 \pm 0.02}$	63 ± 5
BIBFA	0.63 ± 0.01	29 ± 1	$\mathbf{0.90 \pm 0.01}$	32 ± 1	0.87 ± 0.01	180 ± 10	0.62 ± 0.03	8 ± 1
CCA	0.56 ± 0.01	13	0.65 ± 0.03	5	0.87 ± 0.01	72	0.56 ± 0.06	18
CCA + LR	0.60 ± 0.01	13	0.65 ± 0.03	5	0.87 ± 0.01	72	0.56 ± 0.07	18
PCA + LR	$\mathbf{0.65 \pm 0.01}$	66 ± 1	$\mathbf{0.90 \pm 0.01}$	87 ± 2	0.82 ± 0.01	18 ± 1	0.57 ± 0.05	22 ± 2
MLP	0.59 ± 0.01	220 ± 8	0.79 ± 0.01	350 ± 167	0.77 ± 0.01	210 ± 37	0.53 ± 0.04	290 ± 120
LR	$\mathbf{0.65 \pm 0.01}$	-	$\mathbf{0.90 \pm 0.01}$	-	$\mathbf{0.88 \pm 0.01}$	-	0.55 ± 0.06	-
RR	$\mathbf{0.65 \pm 0.01}$	-	0.89 ± 0.01	-	0.87 ± 0.01	-	0.59 ± 0.06	-

Table 8: Results of the predictive SSHIBA and the different methods under study on multi-label and categorical databases. Results include the performance in terms of AUC and the number K_c of latent factors when 20% of the training samples are used. These results have been calculated with a 5-fold CV, so their standard deviations are also included.

Table 8 shows the results using 20% of the training data. Even in this challenging setup, SSHIBA achieves competitive (if not the best) results in all cases while still providing data dimensionality reduction. In particular, we highlight the results in *yeast*, where SSHIBA achieves the smallest latent dimension among the best performing methods, and in *birds*, where SSHIBA stands out of all methods. Also note how SSHIBA is able to significantly achieve a smaller latent space w.r.t. BIBFA in *AVIRIS*. We conjecture that the ability of SSHIBA to treat each data type according to its true nature (binary/categorical) explains the robustness of the method in the low sample-size regime.

For the previous results, we have set a less conservative pruning criteria in the *SSHIBA* model, resulting in a number of latent factors that in some cases could be reduced without harming the performance. If we wanted to analyse the complexity of the model as a predictor/classifier, we would have to take into account two things:

1. The pruning criteria: in our previous results we have set a latent factor can be pruned when there is no value higher than 10^{-6} for any feature k in any projection matrix $\mathbf{W}^{(\mathrm{m})}$, resulting in a smaller fraction of pruned factors. We can increase this pruning criteria by setting it to a higher value and, therefore, reducing the number of latent factors.

2. The latent factors that affect the prediction/classification: we can analyse the sparse weight matrices $\mathbf{W}^{(\mathrm{m})}$ to find which latent factors are relevant for each view. From this premise we can state that the latent variables that are not relevant for the output view, $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m_{out}})} \approx 0$, will not influence the prediction. Hence only a subset of the latent variables are going to be used for the prediction, namely the latent variables common to the input and output view and the ones private to the output view.

In Figure 5 we analyze the results on these two issues. Firstly, we have set the pruning threshold to a considerably lower value, a *restrictive pruning*. This reduces the number of latent factors (for instance, in *scene* is reduced from 137 to 40) while maintaining the performance in terms of AUC in the four databases, as we saw in Table 7. Secondly, by inspecting the weight matrices $\mathbf{W}^{(\mathrm{m})}$, we can analyze the relevance of each latent factor for the different views. In Figure 5 we represent the relevance of each latent factor as well as which view it is relevant to. We can conclude that not all the

latent factors are used for both views, some are only relevant for a specific view. This can translate into a further reduction of the latent factors needed for the prediction of the test data, as we will only need the common factors and those related to the output view. Therefore we have that for *yeast* the effective prediction latent factors are reduced from 66 to 12, *scene* from 137 to 21, for *AVIRIS* from 197 to 63 and for *birds* from 75 to 18.

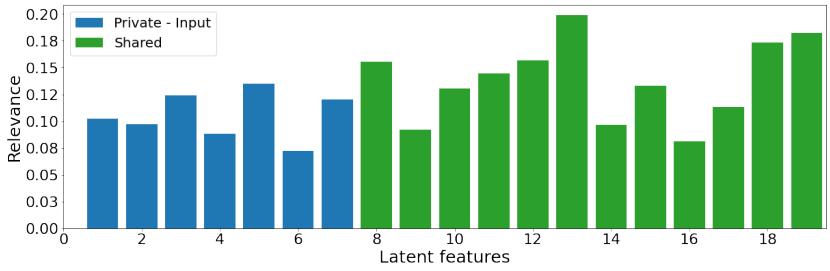
Finally, we can conclude that the results obtained in this section prove that the model provides a performance equivalent to the best of the baselines we are comparing to. This is done while also providing a selection of the most relevant features, automatic imputation of any missing value in the data as well as allowing to combine and correctly model different types of data in different views to combine their information and enhance the performance of the model.

4.4. Feature selection with SSHIBA

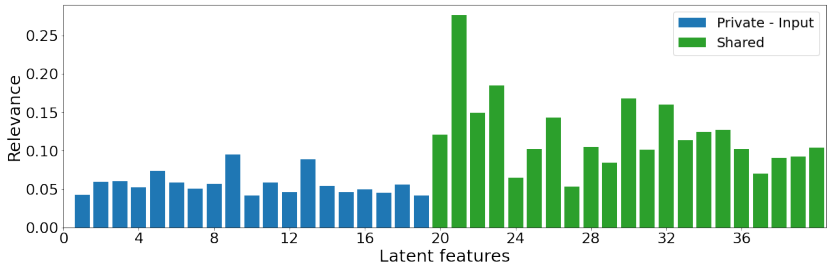
This section focuses on the extension of our model to allow feature selection, as presented in Section 3.1. To do so, we use the categorical and multi-label databases *LFW* and *LFWA*. With these experiments we aim to visually analyse the feature relevances, as well as the latent space learnt by the model and how it describes the data.

Figure 6 shows each of the columns of the matrix $\mathbf{W}^{(1)}$ learned by the model in both databases (recall that each column of this matrix has the same dimension as the images). The columns are ordered using the value of the variable $\boldsymbol{\alpha}^{(1)}$, since it provides the relevance of each latent factor. Note that each column of the matrix is capturing a face shape, and these faces will be combined for data reconstruction. In both Figure 6a and Figure 6b, we can see how, as we advance through the faces, we reach a point in which the images become more blurry and less informative, around the sixth row and corresponding to a value of $\boldsymbol{\alpha}^{(1)} \approx 0.3$. It is around this point that we could start pruning and removing the irrelevant latent factors which do not provide significant information.

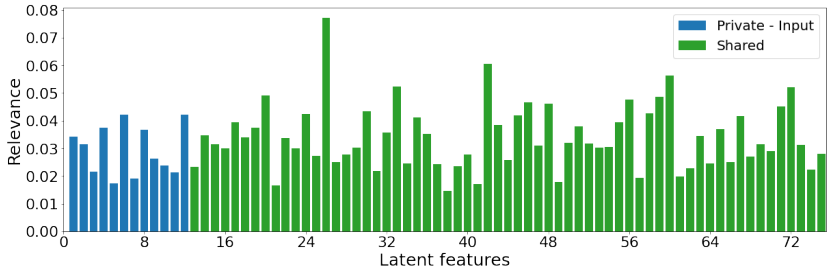
Besides, these images reveal how the model adapts to the learning task. E.g., in the case of Figure 6a we can see how the model pay more attention to the different individuals and some latent-faces can be related to some labels: the first latent-face seems to be dedicated to *George W. Bush* and the second one to *Hugo Chavez*. On the other hand, in Figure 6b latent faces tend to focus on face regions associated to different attributes, such as, the eyes area or the forehead.



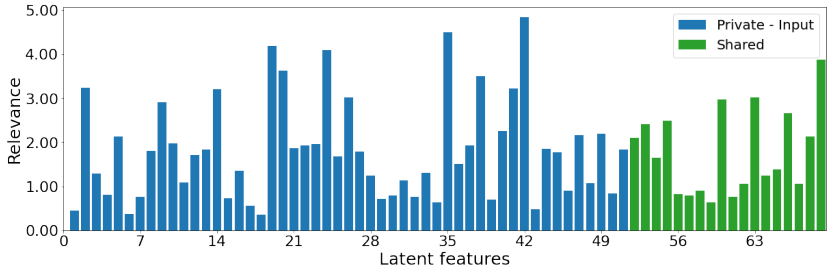
(a) *yeast*



(b) *scene*



(c) *AVIRIS*



(d) *birds*

Figure 5: Analysis of the learnt latent factors over the different views.

(a) LFW database. (b) LFWA database.

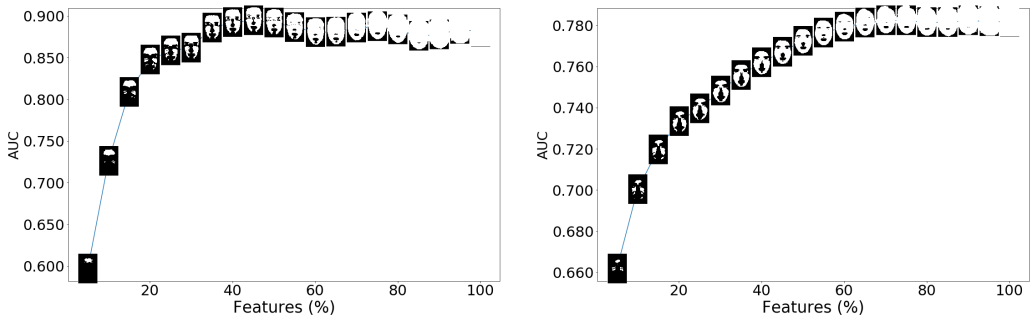
Figure 6: $\mathbf{W}^{(1)}$ matrix learnt by the sparse version of SSHIBA using two different databases. Each latent face is a column of this matrix $\mathbf{W}^{(1)}$. The images include the latent faces learned by the model and are ordered using the latent relevance variable $\boldsymbol{\alpha}^{(1)}$.

In Figure 7 we can see the representation of the sparsity variable, $\boldsymbol{\gamma}^{(1)}$, which indicates the relevance of each feature, i.e. the relevance of each pixel. These two figures provide an insight into how the model adapts the results to the problem, e.g. for the identification of 7 different subjects, Figure 7a shows how the algorithm focuses on some specific areas, such as the forehead. However, when looking at Figure 7b we can see that the model focuses on different regions, relevant the characteristics the each person has, such as his eye colour, his race or the strength of his nose lines.

Finally, we can analyse the effectiveness of the feature selection by ordering the features (pixels) by relevance and calculating the final model perfor-

(a) LFW (b) LFWA

Figure 7: Gamma masks learnt by the sparse version of SSHIBA using two different databases. The masks represent the importance of each pixel: lighter colours imply the pixel is more relevant while darker ones represent the pixel is less relevant.



(a) LFW database. (b) LFWA database.

Figure 8: AUC results on the *LFW* and *LFWA* databases using the sparse version of the method. These images show the AUC results using different percentages of the most relevant values in the learnt mask. Each face shows the mask with different numbers of features.

mance for different percentages of selected features. Figure 8 shows this AUC evolution, where the results prove that using only around 50% of the pixels, the model is capable of obtaining a good classification AUC. In particular, Figure 8a shows an improvement in the performance using only 40% of the original pixels.

4.5. Missing data imputation with SSHIBA

This section presents the experiments we carried out using the semi-supervised approach for the imputation of missing values. In this case, we included random patterns of missing values in four different databases and used SSHIBA to impute such values using semi-supervised approach. For this experiment we used the *yeast*, *scene*, *AVIRIS* and *birds* databases. We compare the semi-supervised approach with both the predictive method (assuming no missing data in the train set), and with the results obtained when the train missing pattern is first imputed using some common imputation techniques. In Table 9 we include the obtained results. First, note that in the case of no missing values, the semi-supervised method (which jointly processes the test and training data) is able to improve the predictive method, obtaining a 0.68 AUC, which achieves the best result in Table 7. Furthermore, when we include a 50% of missing data in the train set, the use of the semi-supervised SSHIBA with no pre-imputation method achieves the best results, as the probabilistic model is able to handle the uncertainty of the missing entries with no artificial imputation. This result certainly demonstrate the superior ability of the method to capture hidden correlations in our data, boosted by a proper modelling of each data type.

Missing Pattern	Imputation Method	SSHIBA	AUCs			
			yeast	*scene*	*AVIRIS*	*birds*
No missing in train.	–	Predictive	0.66	0.92	0.88	0.83
		SS	**0.68**	0.92	0.88	0.83
50% missing in train.	Semi-Supervised	SS	**0.64**	**0.89**	**0.87**	**0.79**
	Mean		0.61	0.87	0.78	0.77
	Median		0.55	0.70	0.78	0.75
	Most frequent value		0.48	0.52	0.77	0.74

Table 9: Results on *yeast*, *scene*, *AVIRIS* and *birds* databases of the semi-supervised and predictive SSHIBA in comparison to different imputation techniques. Results include the AUC values with the complete dataset and when there is a 50% of missing input data.

4.6. Multiview learning with SSHIBA

As a final experiment on the proposed SSHIBA algorithm, we tested its potential on a multiview problem. In this case, we decided to combine the information of the LFW and LFWA databases to have information of both the person identity and their characteristics. This problem was also

solved with the previously defined baselines to compare the results. As these methods are not compatible with multiview, we decided to incorporate the extra information as an extra input feature.

	Two views		Three views	
	AUC	K_c	AUC	K_c
SSHIBA	**0.68**	39	**0.69**	35
CCA	0.60	62	0.60	62
CCA + Log. Reg.	0.60	62	0.60	62
PCA + Log. Reg.	0.65	187	0.66	187
MLP	0.60	375	0.60	375
Logistic reg.	0.65	-	0.65	-
Ridge reg.	0.67	-	0.67	-

Table 10: Results on *LFWA* database using the data of the *LFW* database as an extra view. Results include the performance in terms of AUC and the number of latent factors when the complete *LFWA* dataset is used (*two views*) and when the data from the *LFW* database is included (*three views*).

In Table 10 we can see the results obtained. Results include the AUC values for all methods under study when the complete *LFWA* dataset is used (*Two views*) and when the data from the *LFW* database is also incorporated (*Three views*). First of all, we can notice that the SSHIBA algorithm is not only outperforming the rest of the baseline results, but also having a significantly lower number of latent features than the FE algorithms. Equivalently, we can see that the addition of a new view with further information on the data leads to a reduction on the latent features as well as an improvement of the performance of the algorithm. The inclusion of additional information allows the model to capture more accurate data correlations with a smaller hidden dimensionality.

5. Conclusions

In this article we generalize the BIBFA model to create a new FA framework, called SSHIBA, capable of adapting to the particularities of any learning problem. In particular, this new model includes new functionalities, such as, being able to carry out a selection of the most relevant features while extracting latent features, modelling not only real problem but also multilabel

and categorical ones and, at the same time, work in a semi-supervised way with unlabelled data and missing values.

The results with SSHIBA show that, in the worst case, the performance of the method is similar to the state-of-the-art algorithms while being able to find a reduced latent space, having less extracted features than classical feature extraction methods. Combining this with feature selection capabilities and an adequate data modelling (multilabel, categorical, ...), we obtain more compacted models with a gain of interpretability. Besides, the semi-supervised version of the algorithm has been proven to perform like the predictive or even outperform it, while providing the online imputation of any possible missing value in the data, allowing the algorithm to use a greater amount of datasets that might have some unclassified data.

These utilities are of special interest for problems in which we are not only interested in performance, but also in the interpretability of the results (e.g. medical applications). Furthermore, the ability of working with multiple views combined with modelling the data according to their characteristics might benefit databases with different types of heterogeneous data.

In the future, this model can be adapted to work with high-dimensionality databases by working with kernels in the dual space. This change in the formulation will be able to not only enhance the computational time of datasets with a high number of samples, but also include non-linearities in the input data.

6. Acknowledgments

The authors wish to thank Irene Santos, for fruitful discussions and help during the earlier stages of our work. The work of Pablo M. Olmos is supported by Spanish government MEC under grant PID2019-108539RB-C22 and RTI2018-099655-B-10, by Comunidad de Madrid under grants IND2017/TIC-7618, IND2018/TIC-9649, IND2020/TIC-17372, and Y2018/TCS-4705, by BBVA Foundation under the Deep-DARWiN project, and by the European Union (FEDER and the European Research Council (ERC) through the European Unions Horizon 2020 research and innovation program under Grant 714161). C. Sevilla-Salcedo and V. Gómez-Verdejo's work has been partly funded by the Spanish MINECO grants TEC2017-83838-R and PID2020-115363RB-I00.

References

[1] H. Suk, S. Lee, A novel bayesian framework for discriminative feature extraction in brain-computer interfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2) (2012) 286–299.

[2] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in statistics, Springer, 1992, pp. 162–190.

[3] O. Kursun, E. Alpaydin, O. V. Favorov, Canonical correlation analysis using within-class coupling, Pattern Recognition Letters 32 (2) (2011) 134–144.

[4] C. Sevilla-Salcedo, V. Gómez-Verdejo, J. Tohka, A. D. N. Initiative, Regularized bagged canonical component analysis for multiclass learning in brain imaging, Neuroinformatics (2020).

[5] H. Tan, X. Zhang, L. Lan, X. Huang, Z. Luo, Nonnegative constrained graph based canonical correlation analysis for multi-view feature learning, Neural Processing Letters 50 (2) (2019) 1215–1240.

[6] Y. Li, F. Wu, A. Ngom, A review on machine learning principles for multi-view biological data integration, Briefings in bioinformatics 19 (2) (2018) 325–340.

[7] X. Xing, K. Wang, T. Yan, Z. Lv, Complete canonical correlation analysis with application to multi-view gait recognition, Pattern Recognition 50 (2016) 107–117.

[8] J. Chen, L. Du, H. He, Y. Guo, Convolutional factor analysis model with application to radar automatic target recognition, Pattern Recognition 87 (2019) 140–156.

[9] X. Zhang, B. Chen, H. Liu, L. Zuo, B. Feng, Infinite max-margin factor analysis via data augmentation, Pattern Recognition 52 (2016) 17–32.

[10] M. Pearce, J. Branke, Continuous multi-task bayesian optimisation with correlation, European Journal of Operational Research 270 (3) (2018) 1074–1085.

[11] D. Hernández-Lobato, J. M. Hernández-Lobato, T. Helleputte, P. Dupont, Expectation propagation for bayesian multi-task feature selection, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2010, pp. 522–537.

[12] A. F. Marquand, M. Brammer, S. C. Williams, O. M. Doyle, Bayesian multi-task learning for decoding multi-subject neuroimaging data, NeuroImage 92 (2014) 298–311.

[13] K. Adachi, Sparse factor analysis, in: Matrix-Based Introduction to Multivariate Data Analysis, Springer, 2020, pp. 361–382.

[14] E. J. Min, C. Chang, Q. Long, Generalized bayesian factor analysis for integrative clustering with applications to multi-omics data, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2018, pp. 109–119.

[15] W. Yu, J. T. Ormerod, M. Stewart, Variational discriminant analysis with variable selection, Statistics and Computing (2020) 1–19.

[16] P. Connor, P. Hollensen, O. Krigolson, T. Trappenberg, A biological mechanism for bayesian feature selection: Weight decay and raising the lasso, Neural Networks 67 (2015) 121–130.

[17] D. Pauger, H. Wagner, et al., Bayesian effect fusion for categorical predictors, Bayesian Analysis 14 (2) (2019) 341–369.

[18] E. Terzi, M. A. Cengiz, Bayesian hierarchical modeling for categorical longitudinal data from sedation measurements, Computational and mathematical methods in medicine 2013 (2013).

[19] M. Gönen, Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning, Pattern recognition letters 38 (2014) 132–141.

[20] M. Gönen, Bayesian supervised multilabel learning with coupled embedding and classification, in: Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM, 2012, pp. 367–378.

[21] G. Zhang, J. Yin, X. Su, Y. Huang, Y. Lao, Z. Liang, S. Ou, H. Zhang, Augmenting multi-instance multilabel learning with sparse bayesian

models for skin biopsy image analysis, BioMed research international 2014 (2014).

[22] K. Toutanova, M. Johnson, A bayesian lda-based model for semi-supervised part-of-speech tagging, in: Advances in neural information processing systems, 2008, pp. 1521–1528.

[23] W. Lian, P. Rai, E. Salazar, L. Carin, Integrating features and similarities: Flexible models for heterogeneous multiview data., in: AAAI, Citeseer, 2015, pp. 2757–2763.

[24] J. Gordon, J. M. Hernández-Lobato, Bayesian semisupervised learning with deep generative models, arXiv preprint arXiv:1706.09751 (2017).

[25] Z. Ge, Z. Song, Semisupervised bayesian method for soft sensor modeling with unlabeled data samples, AIChE Journal 57 (8) (2011) 2109–2119.

[26] P. Zhu, X. Liu, Y. Wang, X. Yang, Mixture semisupervised bayesian principal component regression for soft sensor modeling, IEEE Access 6 (2018) 40909–40919.

[27] A. Klami, S. Virtanen, S. Kaski, Bayesian canonical correlation analysis, Journal of Machine Learning Research 14 (Apr) (2013) 965–1003.

[28] R. M. Neal, Bayesian learning for neural networks, Vol. 118, Springer Science & Business Media, 2012.

[29] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, Journal of the American Statistical Association 112 (518) (2017) 859–877.

[30] T. Jaakkola, M. Jordan, A variational approach to bayesian logistic regression models and their extensions, in: Sixth International Workshop on Artificial Intelligence and Statistics, Vol. 82, 1997, p. 4.

[31] M. Girolami, S. Rogers, Variational bayesian multinomial probit regression with gaussian process priors, Neural Computation 18 (8) (2006) 1790–1817.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay,

Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[33] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, I. Vlahavas, Mulan: A java library for multi-label learning, Journal of Machine Learning Research 12 (2011) 2411–2414.

[34] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, in: Advances in neural information processing systems, 2002, pp. 681–687.

[35] M. R. Boutell, J. Luo, X. Shen, C. M. Brown, Learning multi-label scene classification, Pattern recognition 37 (9) (2004) 1757–1771.

[36] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern, et al., The 9th annual mlsp competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment, in: 2013 IEEE international workshop on machine learning for signal processing (MLSP), IEEE, 2013, pp. 1–8.

[37] M. Baumgardner, L. Biehl, D. Landgrebe, 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. purdue university research repository. 2015 (1992).

[38] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007).

[39] L. Wolf, T. Hassner, Y. Taigman, Effective unconstrained face recognition by combining multiple descriptors and learned background statistics, IEEE transactions on pattern analysis and machine intelligence 33 (10) (2010) 1978–1990.

[40] N. Kumar, A. C. Berg, P. N. Belhumeur, S. K. Nayar, Attribute and simile classifiers for face verification, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 365–372.

[41] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 145–158.