Chapter 19

Integration of Multimodal Data

Marco Lorenzi, Marie Deprez, Irene Balelli, Ana L. Aguila, and Andre Altmann

Abstract

This chapter focuses on the joint modeling of heterogeneous information, such as imaging, clinical, and biological data. This kind of problem requires to generalize classical uni- and multivariate association models to account for complex data structure and interactions, as well as high data dimensionality.

Typical approaches are essentially based on the identification of latent modes of maximal statistical association between different sets of features and ultimately allow to identify joint patterns of variations between different data modalities, as well as to predict a target modality conditioned on the available ones. This rationale can be extended to account for several data modalities jointly, to define multi-view, or multi-channel, representation of multiple modalities. This chapter covers both classical approaches such as partial least squares (PLS) and canonical correlation analysis (CCA), along with most recent advances based on multi-channel variational autoencoders. Specific attention is here devoted to the problem of interpretability and generalization of such high-dimensional models. These methods are illustrated in different medical imaging applications, and in the joint analysis of imaging and non-imaging information, such as -omics or clinical data.

Key words Multivariate analysis, Latent variable models, Multimodal imaging, -Omics, Imaging-genetics, Partial least squares, Canonical correlation analysis, Variational autoencoders, Sparsity, Interpretability

1 Introduction

The goal of multimodal data analysis is to reveal novel insights on complex biological conditions. Through the combined analysis of multiple type of data, and the complementary views on pathophysiological processes they provide, we have the potential to improve our understanding of the underlying processes leading to complex and multifactorial disorders [1]. In medical imaging applications, multiple imaging modalities, such as structural magnetic resonance imaging (sMRI), functional MRI (fMRI), diffusion tensor imaging (DTI), or positron emission tomography (PET), can be jointly analyzed to better characterize pathological conditions affecting individuals [2]. Other typical multimodal analysis problems involve

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_19,

the joint analysis of heterogeneous data types, such as imaging and genetics data, where medical imaging is associated with the patient's genotype information, represented by genetic variants such as single-nucleotide polymorphisms (SNPs) [3]. This kind of application, termed *imaging-genetics*, is of central importance for the identification of genetic risk factors underlying complex diseases including age-related macular degeneration, obesity, schizophrenia, and Alzheimer's disease [4].

Despite the great potential of multimodal data analysis, the complexity of multiple data types and clinical questions poses several challenges to the researchers, involving scalability, interpretability, and generalization of complex association models.

1.1 Challenges of Multimodal Data Assimilation

Due to the complementary nature of multimodal information, there is great interest in combining different data types to better characterize the anatomy and physiology of patients and individuals. Multimodal data is generally acquired using heterogeneous protocols highlighting different anatomical, physiological, clinical, and biological information for a given individual [5].

Typical multimodal data integration challenges are:

- *Non-commensurability.* Since each data modality quantifies different physical and biological phenomena, multimodal data is represented by heterogeneous physical units associated to different aspects of the studied biological process (e.g., brain structure, activity, clinical scores, gene expression levels).
- *Spatial heterogeneity.* Multimodal medical images are characterized by specific spatial resolution, which is independent from the spatial coordinate system on which they are standardized.
- *Heterogeneous dimensions.* The data type and dimensions of medical data can vary according to the modality, ranging from scalars and time series typical of fMRI and PET data to structured tensors of diffusion weighted imaging.
- *Heterogeneous noise.* Medical data modalities are characterized by specific and heterogeneous artifacts and measurement uncertainty, resulting from heterogeneous acquisition and processing routines.
- *Missing data.* Multimodal medical datasets are often incomplete, since patients may not undergo the same protocol, and some modalities may be more expensive to acquire than others.
- *Interpretability.* A major challenge of multimodal data integration is the interpretability of the analysis results. This aspect is impacted by the complexity of the analysis methods and generally requires important expertise in data acquisition, processing, and analysis.

Multimodal data analysis methods proposed in the literature have been focusing on different data complexity and integration, depending on the application of interest. Visual inspection is the typical initial step of multimodal studies, where single modalities are compared on a qualitative basis. For example, different medical imaging modalities can be jointly visualized for a given individual to identify common spatial patterns of signal changes. Data integration can be subsequently performed by jointly exploring unimodal features and unimodal analysis results. To this end, we may stratify the cohort of a clinical study based on some biomarkers extracted from different medical imaging modalities exceeding predefined thresholds. Finally, multivariate statistical and machine learning techniques can be applied for data-driven analysis of the joint relationship between information encoded in different modalities. Such approaches attempt to maximize the advantages of combining cross-modality information, dimensions, and resolution of the multimodal signal. The ultimate goal of such analysis methods is to identify the "mechanisms" underlying the generation of the observed medical data, to provide a joint representation of the common variation of heterogeneous data types.

The literature on multimodal analysis approaches is extensive, depending on the kind of applications and related data types. In this chapter we focus on general data integration methods, which can be classically related to the fields of *multivariate statistical analysis* and *latent variable modeling*. The importance of these approaches lies in the generality of their formulation, which makes them an ideal baseline for the analysis of heterogeneous data types. Furthermore, this chapter illustrates current extensions of these basic approaches to deep probabilistic models, which allow great modeling flexibility for current state-of-the-art applications.

In Subheading 1.2 we provide an overview of typical multimodal analyses in neuroimaging applications, while in Subheading 2 we introduce the statistical foundations of multivariate latent variable modeling, with emphasis on the standard approaches of partial least squares (PLS) and canonical correlation analysis (CCA). In Subheading 3, these classical methods are reformulated under the Bayesian lens, to define linear counterparts of latent variable models (Subheading 3.2) and their extension to multichannel and deep multivariate analysis (Subheadings 3.3 and 3.4). In Subheading 4 we finally address the problem of group-wise regularization to improve the interpretability of multivariate association models, with specific focus in imaging-genetics applications.

Box 1: Online Tutorial
The material covered in this chapter is available at the following online tutorial:
https://bit.ly/3y4RaIO

1.2 Motivation from Neuroimaging Applications

Multimodal analysis methods have been explored for their potential in automatic patient diagnosis and stratification, as well as for their ability to identify interpretable data patterns characterizing clinical conditions. In this section, we summarize state-of-the-art contributions to the field, along with the remaining challenges to improve our understanding and applications to complex brain disorders.

- *Structural-structural combination*. Methods combining sMRI and dMRI imaging modalities are predominant in the field. Such combined analysis has been proposed, for example, for the detection of brain lesions (e.g., strokes [6, 7]) and to study and improve the management of patients with brain disorders [8].
- *Functional-functional combination*. Due to the complementary nature of EEG and fMRI, research in brain connectivity analysis has focused in the fusion of these modalities, to optimally integrate the high temporal resolution of EEG with the high spatial resolution of the fMRI signal. As a result, EEG-fMRI can provide simultaneous cortical and subcortical recording of brain activity with high spatiotemporal resolution. For example, this combination is increasingly used to provide clinical support for the diagnosis and treatment of epilepsy, to accurately localize seizure onset areas, as well as to map the surrounding functional cortex in order to avoid disability [9–11].
- *Structural-functional combination*. The combined analysis of sMRI, dMRI, and fMRI has been frequently proposed in neuropsychiatric research due to the high clinical availability of these imaging modalities and due to their potential to link brain function, structure, and connectivity. A typical application is in the study of autism spectrum disorder and attention-deficit hyperactivity disorder (ADHD). The combined analysis of such modalities has been proposed, for example, for the identification of altered white matter connectivity patterns in children with ADHD [12], highlighting association patterns between regional brain structural and functional abnormalities [13].
- *Imaging-genetics*. The combination of imaging and genetics data has been increasingly studied to identify genetic risk factors (genetic variations) associated with functional or structural abnormalities (quantitative traits, QTs) in complex brain disorders [3]. Such multimodal analyses are key to identify the

underlying mechanisms (from genotype to phenotype) leading to neurodegenerative diseases, such as Alzheimer's disease [14] or Parkinson's disease [15]. This analysis paradigm paves the way to novel data integration scenarios, including imaging and transcriptomics, or multi-omic data [16].

Overall, multimodal data integration in the study of brain disorders has shown promising results and is an actively evolving field. The potential of neuroimaging information is continuously improving, with increasing resolution and improved image contrast. Moreover, multiple imaging modalities are increasingly available in large collections of multimodal brain data, allowing for the application of complex modeling approaches on representative cohorts.

2 Methodological Background

2.1 From Multivariate Regression to Latent Variable Models

The use of multivariate analysis methods for biomedical data analysis is widespread, for example, in neuroscience [17], genetics [18], and imaging-genetics studies [19, 20]. These approaches come with the potential of explicitly highlighting the underlying relationship between data modalities, by identifying sets of relevant features that are jointly associated to explain the observed data.

In what follows, we represent the multimodal information available for a given subject k as a collection of arrays $\boldsymbol{x}_i^k$, $i = 1, \ldots, M$, where M is the number of available modalities. Each array has dimension $dim(\boldsymbol{x}_i^k) = D_i$. A multimodal data matrix for N individuals is therefore represented by the collection of matrices $\boldsymbol{X}_i$, with $dim(\boldsymbol{X}_i) = N \times D_i$. For sake of simplicity, we assume that $\boldsymbol{x}_i^k \in \mathbb{R}^{D_i}$.

A first assumption that can be made for defining a multivariate analysis method is that a target modality, say $\boldsymbol{X}_j$, is generated by the combination of a set of given modalities $\{\boldsymbol{X}_i\}_{i \neq j}$. A typical example of this application concerns the prediction of certain clinical variables from the combination of imaging features. In this case, the underlying forward *generative model* for an observation $\boldsymbol{x}_j^k$ can be expressed as:

$$\boldsymbol{x}_j^k = g(\{\boldsymbol{x}_i^k\}_{i \neq j}) + \boldsymbol{\varepsilon}_j^k, \qquad (1)$$

where we assume that there exists an ideal mapping $g(\cdot)$ that transforms the ensemble of observed modalities for the individual k, to generate the target one $\boldsymbol{x}_j^k$. Note that we generally assume that the observations are corrupted by a certain noise $\boldsymbol{\varepsilon}_j^k$, whose nature depends on the data type. The standard choice for the noise is Gaussian, $\boldsymbol{\varepsilon}_j^k \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{Id})$.

Within this setting, a multimodal model is represented by a function $f(\{\boldsymbol{X}_i\}_{i=1}^M, \boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$, taking as input the ensemble of modalities across subjects. The model f is optimized

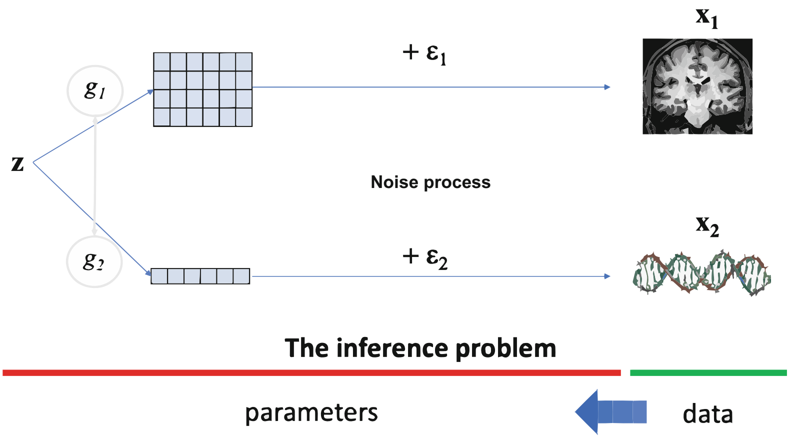

Fig. 1 Illustration of a generative process for the modeling of imaging and genetics data

with respect to $\boldsymbol{\theta}$ to solve a specific task. In our case, the set of input modalities can be used to predict a target modality j, in this case we have $f:\bigotimes_{i\neq j}\mathbb{R}^{D_i}\mapsto\mathbb{R}^{\mathrm{D_j}}$.

In its basic form, this kind of formulation includes standard multivariate linear regression, where the relationship between two modalities $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is modeled through a set a linear parameters $\boldsymbol{\theta}=\boldsymbol{W}\in\mathbb{R}^{D_2\times D_1}$ and $f(\boldsymbol{X}_2)=\boldsymbol{X}_2\cdot\boldsymbol{W}$. Under the Gaussian noise assumption, the typical optimization task is formulated as the least squares problem:

$$\boldsymbol{W}^*=\underset{\boldsymbol{W}}{\operatorname{argmin}}\;\|\boldsymbol{X}_1-\boldsymbol{X}_2\cdot\boldsymbol{W}\|^2. \tag{2}$$

When modeling jointly multiple modalities, the forward generative model of Eq. 1 may be suboptimal, as it implies the explicit dependence of the target modality upon the other ones. This assumption may be too restrictive, as often an explicit assumption of dependency cannot be made, and we are rather interested in modeling the *joint variation* between data modalities. This is the rationale of *latent variable models*.

In the latent variable setting, we assume that the multiple modalities are jointly dependent from a common latent representation $\boldsymbol{z}$ (Fig. 1) belonging to an ideal low-dimensional space of dimension $D\leq\min\{dim(D_i),\ i=1,\ \ldots,\ M\}$.[1] In this case, Eq. 1 can be extended to the generative process:

$$\boldsymbol{x}_i^k \quad = g_i(\boldsymbol{z}_k)+\boldsymbol{\varepsilon}_i^k, \qquad i=1,\ldots,M. \tag{3}$$

[1] Note that we could also consider *overcomplete* basis for the latent space such that $D>\min\{dim(D_i),\ i=1,\ldots,M\}$. This choice may be motivated by the need of accounting for modalities with particularly low dimension. The study of overcomplete latent data representations is focus of active research [21–23].

Equation 3 is the forward process governing the data generation. The goal of latent variable modeling is to make inference on the latent space and on the generative process from the observed data modalities, based on specific assumptions on the transformations from the latent to the data space, and on the kind of noise process affecting the observations (Box 2). In particular, the inference problem can be tackled by estimating inverse mappings, $f_j(\boldsymbol{x}_j^k)$, from the data space of the observed modalities to the latent space.

Based on this framework, in the following sections, we illustrate the standard approaches for solving the inference problem of Eq. 1.

Box 2: Online Tutorial—Generative Models

The forward model of Eq. 3 for multimodal data generation can be easily coded in Python to generate a synthetic multimodal dataset:

```
# N subjects
n = 500
# here we define 2 Gaussian latents variables
# z = (l_1, l_2)
l1 = np.random.normal(size=n)
l2 = np.random.normal(size=n)

latents = np.array([l1, l2]).T

# We define two random transformations from the latent
#  space to the 5D space of X1 and X2 respectively
transform_x = \
    np.random.randint(-8,8, size = 10).reshape([2,5])
transform_y = \
    np.random.randint(-8,8, size = 10).reshape([2,5])

# We compute data X = z w_x, and Y = z w_y
X1 = latents.dot(transform_x)
X2 = latents.dot(transform_y)

# We add some random Gaussian noise
X1 = X1 + 2*np.random.normal(size = n*5).reshape((n, 5))
X2 = X2 + 2*np.random.normal(size = n*5).reshape((n, 5))
```

2.2 Classical Latent Variable Models: PLS and CCA

Classical latent variable models extend the standard linear regression to analyze the joint variability of different modalities. Typical formulation of latent variable models include *partial least squares* (PLS) and *canonical correlation analysis* (CCA) [24], which have successfully been applied in biomedical research [25], along with multimodal [26, 27] and nonlinear [28, 29] variants.

Box 3: Online Tutorial—PLS and CCA with sklearn

```
from sklearn.cross_decomposition import PLSCanonical, CCA

########################################
# We fit PLS and CCA as provided by scikit-learn

#Defining PLS object, no scaling of input X1 and X2
plsca = PLSCanonical(n_components=2, scale = False)
cca = CCA(n_components=2, scale = False)

#Fitting on train data
plsca.fit(X1, X2)
cca.fit(X1, X2)

#We project the training data in the latent dimension
X1_pls_r, X2_pls_r = \
    plsca.transform(X1, X2)
X1_cca_r, X2_cca_r = \
    cca.transform(X1, X2)
```

The basic principle of these multivariate analysis techniques relies on the identification of *linear transformations* of modalities $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ into a lower dimensional subspace of dimension $D \leq \min \{dim(D_i), dim(D_j)\}$, where the projected data exhibits the desired statistical properties of similarity. For example, PLS aims at maximizing the covariance between these combinations (or projections on the modes' directions), while CCA maximizes their statistical correlation (Box 3). For simplicity, in what follows we focus on the joint analysis of two modalities $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and the multimodal model can be written as

$$
\begin{aligned}
f(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{\theta}) &= [f_1(\boldsymbol{X}_1, \boldsymbol{u}_1), f_2(\boldsymbol{X}_2, \boldsymbol{u}_2)] && (4)\\
&= [\boldsymbol{z}_1, \boldsymbol{z}_2], && (5)
\end{aligned}
$$

where $\boldsymbol{\theta} = \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ are linear projection operators for the modalities, $\boldsymbol{u}_i \in \mathbb{R}^{D_i}$, while $\boldsymbol{z}_i = \boldsymbol{X}_i \cdot \boldsymbol{u}_i \in \mathbb{R}^N$ are the latent projections for each modality $i = 1, 2$. The optimization problem can thus be formulated as:

$$
\begin{aligned}
\boldsymbol{u}_1^*, \boldsymbol{u}_2^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad Sim(\boldsymbol{z}_1, \boldsymbol{z}_2) && (6)\\
&= \underset{\boldsymbol{u}_1, \boldsymbol{u}_2}{\operatorname{argmax}} \quad Sim(\boldsymbol{X}_1 \cdot \boldsymbol{u}_1, \boldsymbol{X}_2 \cdot \boldsymbol{u}_2), && (7)
\end{aligned}
$$

where *Sim* is a suitable measure of statistical similarity, depending on the envisaged methods (e.g., variance for PLS, or correlation for CCA) (Fig. 2).

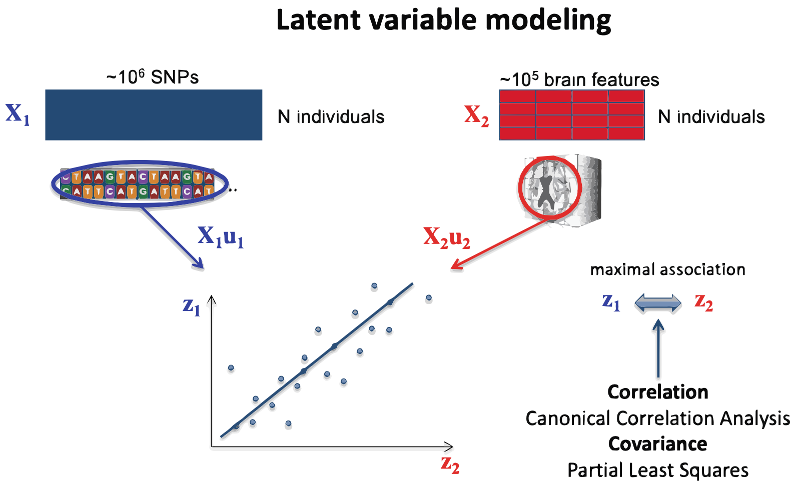

Fig. 2 Illustration of latent variable modeling for an idealized application to the modeling of genetics and imaging data

2.3 Latent Variable Models Through Eigen-Decomposition

2.3.1 Partial Least Squares

For PLS, the problem of Eq. 6 requires the estimation of projections $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ maximizing the *covariance* between the latent representation of the two modalities $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$:

$$\boldsymbol{u}_1^*, \boldsymbol{u}_2^* \quad = \underset{\boldsymbol{u}_1, \boldsymbol{u}_2}{\operatorname{argmax}} \quad \operatorname{Cov}(\boldsymbol{X}_1 \cdot \boldsymbol{u}_1, \boldsymbol{X}_2 \cdot \boldsymbol{u}_2), \tag{8}$$

where

$$\operatorname{Cov}(\boldsymbol{X}_1 \cdot \boldsymbol{u}_1, \boldsymbol{X}_2 \cdot \boldsymbol{u}_2) = \frac{\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_2}{\sqrt{\boldsymbol{u}_1^T \boldsymbol{u}_1}\sqrt{\boldsymbol{u}_2^T \boldsymbol{u}_2}}, \tag{9}$$

and $\boldsymbol{S} = \boldsymbol{X}_1^T \boldsymbol{X}_2$ is the sample covariance between modalities.

Without loss of generality, the maximization of Eq. 9 can be considered under the orthogonality constraint $\sqrt{\boldsymbol{u}_1^T \boldsymbol{u}_1} = \sqrt{\boldsymbol{u}_2^T \boldsymbol{u}_2} = 1$. This constrained optimization problem can be expressed in the Lagrangian form:

$$\mathcal{L}(\boldsymbol{u}_1, \boldsymbol{u}_2, \lambda_x, \lambda_y) = \boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_2 - \lambda_x(\boldsymbol{u}_1^T \boldsymbol{u}_1 - 1) - \lambda_y(\boldsymbol{u}_2^T \boldsymbol{u}_2 - 1), \tag{10}$$

whose solution can be written as:

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{S} \\ \boldsymbol{S}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}. \tag{11}$$

Equation 11 corresponds to the *primal* formulation of PLS and shows that the PLS projections maximizing the latent covariance are the left and right eigen-vectors of the sample covariance matrix across modalities. This solution is known as PLS-SVD and has been widely adopted in the field of neuroimaging [30, 31], for the study of common patterns of variability between multimodal imaging data, such as PET and fMRI.

It is worth to notice that classical *principal component analysis* (PCA) is a special case of PLS when $\boldsymbol{X}_1 = \boldsymbol{X}_2$. In this case the latent projections maximize the data variance and correspond to the eigen-modes of the sample covariance matrix $\boldsymbol{S} = \boldsymbol{X}_1^T \boldsymbol{X}_1$.

2.3.2 Canonical Correlation Analysis

In canonical correlation analysis (CCA), the problem of Eq. 6 is formulated by optimizing linear transformations such that $\boldsymbol{X}_1 \boldsymbol{u}_1$ and $\boldsymbol{X}_2 \boldsymbol{u}_2$ are maximally correlated:

$$\boldsymbol{u}_1^*, \boldsymbol{u}_2^* = \underset{\boldsymbol{u}_1, \boldsymbol{u}_2}{\operatorname{argmax}} \operatorname{Corr}(\boldsymbol{X}_1 \boldsymbol{u}_1, \boldsymbol{X}_2 \boldsymbol{u}_2), \tag{12}$$

where

$$\operatorname{Corr}(\boldsymbol{X}_1 \boldsymbol{u}_1, \boldsymbol{X}_2 \boldsymbol{u}_2) = \frac{\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_2}{\sqrt{\boldsymbol{u}_1^T \boldsymbol{S}_1 \boldsymbol{u}_1} \sqrt{\boldsymbol{u}_2^T \boldsymbol{S}_2 \boldsymbol{u}_2}}. \tag{13}$$

where $\boldsymbol{S}_1 = \boldsymbol{X}_1^T \boldsymbol{X}_1$ and $\boldsymbol{S}_2 = \boldsymbol{X}_2^T \boldsymbol{X}_2$ are the sample covariances of modality 1 and 2, respectively.

Proceeding in a similar way as for the derivation of PLS, it can be shown that CCA is associated to the generalized eigen-decomposition problem [32]:

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{S} \\ \boldsymbol{S}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{S}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}, \tag{14}$$

It is common practice to reformulate the CCA problem of Eq. 14 with a *regularized* version aimed to avoid numerical instabilities due to the estimation of the sample covariances $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$:

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{S} \\ \boldsymbol{S}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{S}_1 + \delta \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_2 + \delta \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_1 \end{bmatrix}. \tag{15}$$

In this latter formulation, the right hand side of Eq. 14 is regularized by introducing a constant diagonal term δ, proportional to the regularization strength (with $\delta = 0$ we obtain Eq. 14). Interestingly, for large value of δ, the diagonal term dominates the sample covariance matrices of the right-hand side, and we retrieve the standard eigen-value problem of Eq. 11. This shows that PLS can be interpreted as an infinitely regularized formulation of CCA.

2.4 Kernel Methods for Latent Variable Models

In order to capture nonlinear relationships, we may wish to project our input features into a high-dimensional space prior to performing CCA (or PLS):

$$\phi : \boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N) \mapsto [\phi(\boldsymbol{x}^1), \ldots, \phi(\boldsymbol{x}^{\mathrm{N}})] \tag{16}$$

where ϕ is a nonlinear feature map. As derived by Bach et al. [33], the data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ can be replaced by the Gram matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ such that we can achieve a nonlinear feature mapping via the kernel trick [34]:

$$\boldsymbol{K}_1\left(\boldsymbol{x}_1^i, \boldsymbol{x}_1^j\right) = \left\langle \phi(\boldsymbol{x}_1^i), \phi\left(\boldsymbol{x}_1^j\right)\right\rangle \text{ and } \boldsymbol{K}_2\left(\boldsymbol{x}_2^i, \boldsymbol{x}_2^j\right) = \left\langle \phi(\boldsymbol{x}_2^i), \phi\left(\boldsymbol{x}_2^j\right)\right\rangle \tag{17}$$

where $\mathbf{K}_1 = [K_1\left(\boldsymbol{x}_1^i, \boldsymbol{x}_1^j\right)]_{N\times N}$ and $\mathbf{K}_2 = [K_2\left(\boldsymbol{x}_2^i, \boldsymbol{x}_2^j\right)]_{N\times N}$. In this case, kernel CCA canonical directions correspond to the solutions of the updated generalized eigen-value problem:

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{K}_1^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}. \tag{18}$$

Similarly to the primal formulation of CCA, we can apply an ℓ_2-norm regularization penalty on the weights α_1 and α_2 of Eq. 18, giving rise to regularized kernel CCA:

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{K}_1\boldsymbol{K}_2 \\ \boldsymbol{K}_2\boldsymbol{K}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{K}_1^2 + \delta\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{K}_2^2 + \delta\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}, \tag{19}$$

2.5 Optimization of Latent Variable Models

The *nonlinear iterative partial least squares* (NIPALS) is a classical scheme proposed by H. Wold [35] for the optimization of latent variable models through the iterative computation of PLS and CCA projections. Within this method, the projections associated with the modalities $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are obtained through the iterative solution of simple least squares problems.

The principle of NIPALS is to identify projection vectors $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}$ and corresponding latent representations $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ to minimize the functionals

$$\mathcal{L}_i \quad = \|X_i - \boldsymbol{z}_i \boldsymbol{u}_i^T\|^2, \tag{20}$$

subject to the constraint of maximal similarity between representations $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ (Fig. 3).

Following [37], the NIPALS method is optimized as follows (Algorithm 1). The latent projection for modality 1 is first initialized as $\boldsymbol{z}_1^{(0)}$ from randomly chosen columns of the data matrix $\boldsymbol{X}_1$. Subsequently, the linear regression function

$$\mathcal{L}_2^{(0)} = \|X_2 - \boldsymbol{z}_1^{(0)} \boldsymbol{u}_2^T\|^2$$

is optimized with respect to $\boldsymbol{u}_2$, to obtain the projection $\boldsymbol{u}_2^{(0)}$. After unit scaling of the projection coefficients, the new latent representation is computed for modality 2 as $\boldsymbol{z}_2^{(0)} = \boldsymbol{X}_2 \cdot \boldsymbol{u}_2^{(0)}$. At this point, the latent projection is used for a new optimization step of the linear regression problem

$$\mathcal{L}_1^{(0)} = \|X_1 - \boldsymbol{z}_2^{(0)} \boldsymbol{u}_1^T\|^2,$$

this time with respect to $\boldsymbol{u}_1$, to obtain the projection parameters $\boldsymbol{u}_1^{(0)}$ relative to modality 1. After unit scaling of the coefficients, the new latent representations is computed for modality 1 as $\boldsymbol{z}_1^{(1)} = \boldsymbol{X}_1 \cdot \boldsymbol{u}_1^{(0)}$. The whole procedure is then iterated.

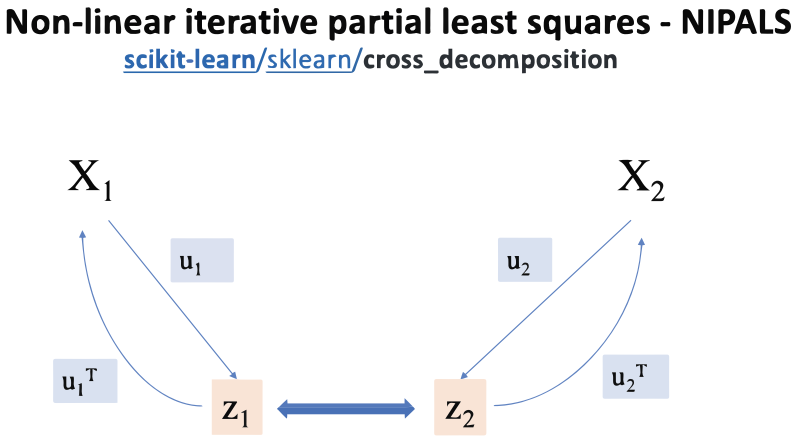
Non-linear iterative partial least squares - NIPALS
scikit-learn/sklearn/cross_decomposition
X_1
X_2
u_1
u_2
u_1^T
u_2^T
z_1
z_2

Fig. 3 Schematic of NIPALS algorithm (Algorithm 1). This implementation can be found in standard machine learning packages such as scikit-learn [36]

It can be shown that the NIPALS method of Algorithm 1 converges to a stable solution for projections and latent parameters and the resulting projection vectors correspond to the first left and right eigen-modes associated to the covariance matrix $\boldsymbol{S} = \boldsymbol{X}_1^T \cdot \boldsymbol{X}_2$.

Algorithm 1 NIPALS iterative computation for PLS components [37]

```
Initialize $\boldsymbol{z}_1^{(0)}$, $i = 0$.
Until not converged do:
    1. Estimate the projection $\boldsymbol{u}_2^{(i)}$ by minimizing $\mathcal{L}_2^{(i)} = \|X_2 - \boldsymbol{z}_1^{(i)} \boldsymbol{u}_2^T\|^2$:
        $\boldsymbol{u}_2^{(i)} = X_2^T \boldsymbol{z}_1^{(i)} \left( \boldsymbol{z}_1^{(i)T} \boldsymbol{z}_1^{(i)} \right)^{-1}$
    2. Normalize $\boldsymbol{u}_2^{(i)} \leftarrow \frac{\boldsymbol{u}_2^{(i)}}{\|\boldsymbol{u}_2^{(i)}\|}$.
    3. Estimate the latent representation for modality 2:
        $\boldsymbol{z}_2^{(i)} = \boldsymbol{X}_2 \cdot \boldsymbol{u}_2^{(i)}$
    4. Estimate the projection $\boldsymbol{u}_1^{(i)}$ by minimizing $\mathcal{L}_1^{(i)} = \|X_1 - \boldsymbol{z}_2^{(i)} \boldsymbol{u}_1^T\|^2$:
        $\boldsymbol{u}_1^{(i)} = X_1^T \boldsymbol{z}_2^{(i)} \left( \boldsymbol{z}_2^{(i)T} \boldsymbol{z}_2^{(i)} \right)^{-1}$
    5. Normalize $\boldsymbol{u}_1^{(i)} \leftarrow \frac{\boldsymbol{u}_1^{(i)}}{\|\boldsymbol{u}_1^{(i)}\|}$.
    6. Update the latent representation for modality 1:
        $\boldsymbol{z}_1^{(i+1)} = \boldsymbol{X}_1 \cdot \boldsymbol{u}_1^{(i)}$.
```

After the first eigen-modes are computed through Algorithm 1, the higher-order components can be subsequently computed by *deflating* the data matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. This can be done by regressing out the current projections in the latent space:

$$\boldsymbol{X}_i \quad \leftarrow \boldsymbol{X}_i - \boldsymbol{z}_i \frac{\boldsymbol{z}_i^T \boldsymbol{X}_i}{\boldsymbol{z}_i^T \boldsymbol{z}_i} \tag{21}$$

NIPALS can be seamlessly used to optimize the CCA problem. Indeed, it can be shown that the CCA projections and latent representations can be obtained by estimating the linear projections $\boldsymbol{u}_2$ and $\boldsymbol{u}_1$ in **steps 1** and **4** of Algorithm 1 via the linear regression problems

$$\mathcal{L}_2^{(i)} = \|X_2\boldsymbol{u}_2 - \boldsymbol{z}_1^{(i)}\|^2 \qquad \text{(step 1 for CCA)},$$

and

$$\mathcal{L}_1^{(i)} = \|X_1\boldsymbol{u}_1 - \boldsymbol{z}_2^{(i)}\|^2 \qquad \text{(step 4 for CCA)}.$$

Box 4: Online Tutorial—NIPALS Implementation
The online tutorial provides an implementation of the NIPALS algorithm for both CCA and PLS, corresponding to Algorithm 1. It can be verified that the numerical solution is equivalent to the one provided by sklearn and to the one obtained through the solution of the eigen-value problem.

3 Bayesian Frameworks for Latent Variable Models

Bayesian formulations for latent variable models have been developed in the past, including for PLS [38] and CCA [39]. The advantage of employing a Bayesian framework to solve the original inference problem is that it provides a natural setting to quantify the parameters' variability in an interpretable manner, coming with their estimated distribution. In addition, these methods are particularly attractive for their ability of integrating prior knowledge on the model's parameters.

3.1 Multi-view PPCA

Recently, the seminal work of Tipping and Bishop on probabilistic PCA (PPCA) [40] has been extended to allow the joint integration of multimodal data [41] (*multi-view PPCA*), under the assumption of a common latent space able to explain and generate all modalities.

Recalling the notation of Subheading 2.1, let $\boldsymbol{x} = \{\boldsymbol{x}_i^k\}_{i=1}^M$ be an observation of M modalities for subject k, where each $\boldsymbol{x}_i^k$ is a vector of dimension D_i. We denote by $\boldsymbol{z}^k$ the D-dimensional latent variable commonly shared by each $\boldsymbol{x}_i^k$. In this context, the forward process underlying the data generation of Eq. 1 is linear, and for each subject k and modality i, we write (*see* Fig. 4a):

$$\boldsymbol{x}_i^k = W_i(\boldsymbol{z}^k) + \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i, \tag{22}$$

$$i = 1,\ldots,M; \quad k = 1,\ldots,N; \quad dim(\boldsymbol{z}^k) < \min(D_i), \tag{23}$$

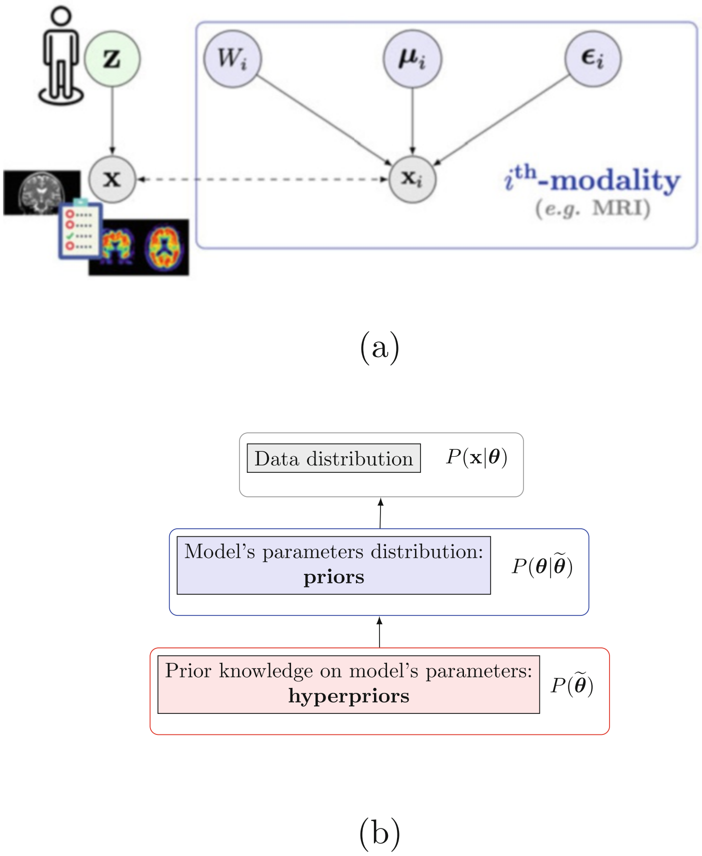
Z
x
W_i
μ_i
ϵ_i
x_i
i^{th}-modality
(e.g. MRI)
(a)
Data distribution $P(\mathbf{x}|\boldsymbol{\theta})$
Model's parameters distribution: priors $P(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}})$
Prior knowledge on model's parameters: hyperpriors $P(\widetilde{\boldsymbol{\theta}})$
(b)

Fig. 4 (**a**) Graphical model of multi-view PPCA. The green node represents the latent variable able to jointly describe all observed data explaining the patient status. Gray nodes denote original multimodal data, and blue nodes the view-specific parameters. (**b**) Hierarchical structure of multi-view PPCA: prior knowledge on model's parameters can be integrated in a natural way when the model is embedded in a Bayesian framework

where W_i represents the linear mapping from the ith-modality to the latent space, while $\boldsymbol{\mu}_i$ and $\boldsymbol{\varepsilon}_i$ denote the common intercept and error for modality i. Note that the modality index i does not appear in the latent variable $\boldsymbol{z}^k$, allowing a compact formulation of the generative model of the whole dataset (i.e., including all modalities) by simple concatenation:

$$\boldsymbol{x}^k := \begin{bmatrix} \boldsymbol{x}_1^k \\ \vdots \\ \boldsymbol{x}_M^k \end{bmatrix} = \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} \boldsymbol{z}^k + \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_M \end{bmatrix} =: W\boldsymbol{z}^k + \boldsymbol{\mu} + \boldsymbol{\varepsilon}. \tag{24}$$

Further hypotheses are needed to define the probability distributions of each element appearing in Eq. 22, such as $\boldsymbol{z}^k \sim p(\boldsymbol{z}^k)$, the standard Gaussian prior distribution for the latent variables, and $\boldsymbol{\varepsilon}_i \sim p(\boldsymbol{\varepsilon}_i)$, a centered Gaussian distribution. From these assumptions, one can finally derive the likelihood of the data given latent variables and model parameters, $p(\boldsymbol{x}_i^k|\boldsymbol{z}^k, \boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i = \{W_i, \boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i\}$ and, by using Bayes theorem, also the posterior distribution of the latent variables, $p(\boldsymbol{z}^k|\boldsymbol{x}_i^k)$.

Box 5: Online Tutorial—Multi-view PPCA

```
from Model.mvPPCA import MVPPCA

### Data in mv-PPCA is specified by:
# 1 - number of views, views' dimensions
# and latent dimension
n_views = 2 # X1 and X2
n_components = n_components
dim_views = [X.shape[1], Y.shape[1]]
# 2 - a dataframe containing all views
data = pd.DataFrame(np.hstack((X, Y)))

### Here we create an instance of the model
#and a dataframe to store results during training
n_iterations=200
results = pd.DataFrame()
# Multi-views PPCA
mvPPCA = MVPPCA(data=data, norm=False,
                dim_views=dim_views,
                n_components=n_components,
                n_iterations=n_iterations)
###################
## Model Fitting ##
###################
results = results.append(mvPPCA.fit(), ignore_index=True)
# Optimized parameters can be recovered as follows:
muk, Wk, Sigma2k = mvPPCA.local_params
```

3.1.1 Optimization

In order to solve the inference problem and estimate the model's parameters in $\boldsymbol{\theta}$, the classical expectation-maximization (EM) scheme can be deployed. EM optimization consists in an iterative process where each iteration is composed of two steps:

- Expectation step (E): Given the parameters previously optimized, the expectation of the log-likelihood of the joint distribution of $\boldsymbol{x}_i$ and $\boldsymbol{z}^k$ with respect to the posterior distribution of the latent variables is evaluated.
- Maximization step (M): The functional of the E step is maximized with respect to the model's parameters.

It is worth noticing that prior knowledge on the model's parameters distribution can be easily integrated in this Bayesian framework (Fig. 4b), with minimal modification of the optimization scheme, consisting in a penalization of the functional to be maximized in the M-step forcing the optimized parameters to remain close to their priors. In this case we talk about maximum a posteriori (MAP) optimization.

3.2 Bayesian Latent Variable Models via Autoencoding

Autoencoders and variational autoencoders have become very popular approaches for the estimation of latent representation of complex data, which allow powerful extensions of the Bayesian models presented in Subheading 3.1 to account for nonlinear and deep data representations.

Autoencoders (AEs) extend classical latent variable models to account for complex, potentially highly nonlinear, projections from the data space to the latent space (encoding), along with reconstruction functions (decoding) mapping the latent representation back to the data space. Since typical encoding (f_e) and decoding (f_d) functions of AEs are parameterized by feedforward neural networks, inference can be efficiently performed by means of stochastic gradient descent through backpropagation. In this sense, AEs can be seen as a powerful extension of classical PCA, where encoding into the latent representations and decoding are jointly optimized to minimize the reconstruction error of the data:

$$\mathcal{L} = \|\boldsymbol{X} - f_d(f_e(\boldsymbol{X}))\|_2^2 \tag{25}$$

The *variational autoencoder* (VAE) [42, 43] introduces a Bayesian formulation of AEs, akin to PPCA, where the latent variables are inferred by estimating the associated posterior distributions. In this case, the optimization problem can be efficiently performed by *stochastic variational inference* [44], where the posterior moments of the variational posterior of the latent distribution are parameterized by neural networks.

In the same way PLS and CCA extend PCA for multimodal analysis, research has been devoted to define equivalent extensions for the VAEs to identify common latent representations of multiple data modalities, such as the multi-channel VAE [23], or deep CCA [29]. These approaches are based on a similar formulation, which is provided in the following section.

3.3 Multi-channel Variational Autoencoder

The multi-channel variational autoencoder (mcVAE) assumes the following generative process for the observation set:

$$\begin{aligned} z^k &\sim p(z^k) \\ x_i^k &\sim p(x_i^k|z^k, \theta_i) \qquad i = 1, \ldots, M, \end{aligned} \tag{26}$$

where $p(z^k)$ is a prior distribution for the latent variable. In this case, $p(x_i^k|z, \theta_i)$ is the likelihood of the observed modality i for subject k, conditioned on the latent variable and on the generative parameters θ_i parameterizing the decoding from the latent space to the data space of modality i.

Solving this inference problem requires the estimation of the posterior for the latent distribution $p(z|X_1, \ldots, X_M)$, which is generally an intractable problem. Following the VAE scheme, *variational inference* can be applied to compute an approximate posterior [45].

3.3.1 Optimization

The inference problem of mcVAE is solved by identifying variational posterior distributions specific to each data modality $q(z^k|x_i^k, \varphi_i)$, by conditioning them on the observed modality x_i and on the corresponding variational parameters φ_i parameterizing the encoding of the observed modality to the latent space.

In this way, since each modality provides a different approximation, a similarity constraint is imposed in the latent space to enforce each modality-specific distribution $q(z^k|x_i^k, \varphi_i)$ to be as close as possible to the common target posterior distribution. The measure of "proximity" between distributions is the Kullback-Leibler (KL) divergence. This constraint defines the following functional:

$$\underset{q}{argmin} \sum_i D_{\mathrm{KL}}\left[q(z^k|x_i^k, \varphi_i) \| p(z|x_1^k, \ldots, x_M^k)\right] \tag{27}$$

where the approximate posteriors $q(z|x_i, \varphi_i)$ represent the view on the latent space that can be inferred from the modality x_i. In [23] it was shown that the optimization of Eq. 27 is equivalent to the optimization of the following evidence lower bound (ELBO):

$$\mathcal{L} = D - R \tag{28}$$

where $R = \sum_i \mathrm{KL}\left[q(z^k|x_i^k, \varphi_i) \| p(z)\right]$, and $D = \Sigma_i L_i$, with

$$L_i = \underset{q(z^k|x_i^k, \varphi_i)}{\mathbb{E}} \sum_{j=1}^{M} \ln p\left(x_j|z, \theta_j\right)$$

is the expected log-likelihood of each data channel x_j quantifying the reconstruction obtained by decoding from the latent representation of the remaining channels x_i. Therefore, optimizing the term D in Eq. 28 with respect to encoding and decoding parameters $\{\theta_i, \varphi_i\}_{i=1}^{M}$ identifies the optimal representation of each modality in the latent space which can, on average, jointly reconstruct all the other channels. This term thus enforces a coherent latent representation across different modalities and is balanced by the regularization term R, which constrains the latent representation of each modality to the common prior $p(\boldsymbol{z})$. As for standard VAEs, encoding and decoding functions can be arbitrarily chosen to parameterize respectively latent distributions and data likelihoods. Typical choices for such functions are neural networks, which can provide extremely flexible and powerful data representation (Box 6). For example, leveraging the modeling capabilities of deep convolutional networks, mcVAE has been used in a recent cardiovascular study for the prediction of cardiac MRI data from retinal fundus images [46].

Box 6 Online Tutorial—mcVAE with PyTorch

```
import torch
from mcvae.models import Mcvae
from mcvae.models.utils import DEVICE, load_or_fit

### Data in mcvae is specified by:
# 1 - a dictionary with the data characteristics
init_dict = {
    'n_channels': 2, # X1 and X2
    'lat_dim': n_components,
    'n_feats': tuple([X1.shape[1], X2.shape[1]]),
}
# 2 - a list with the different data channels
data = []
data.append(torch.FloatTensor(X1))
data.append(torch.FloatTensor(X2))

# Here we create an instance of the model
adam_lr = 1e-2
n_epochs = 4000
# Multi-Channel VAE
torch.manual_seed(24)
model = Mcvae(**init_dict)
model.to(DEVICE)
###################
## Model Fitting ##
###################
model.optimizer = torch.optim.Adam(model.parameters(),\
                                       lr=adam_lr)
load_or_fit(model=model, data=data, epochs=n_epochs,\
                ptfile='model.pt', force_fit=FORCE_REFIT)
```

3.4 Deep CCA

The mcVAE uses neural network layers to learn nonlinear representations of multimodal data. Similarly, Deep CCA [29] provides an alternative to kernel CCA to learn nonlinear mappings of multimodal information. Deep CCA computes representations by passing two views through functions f_1 and f_2 with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively, which can be learnt by multilayer neural networks. The parameters are optimized by maximizing the correlation between the learned representations $f_1(\boldsymbol{X}_1;\boldsymbol{\theta}_1)$ and $f_2(\boldsymbol{X}_2;\boldsymbol{\theta}_2)$:

$$\left(\boldsymbol{\theta}_{1opt}, \boldsymbol{\theta}_{2opt}\right) = \operatorname{argmax} \operatorname{Corr}(f_1(\boldsymbol{X}_1;\boldsymbol{\theta}_1), f_2(\boldsymbol{X}_2;\boldsymbol{\theta}_2))(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \quad (29)$$

In its classical formulation, the correlation objective given in Eq. 29 is a function of the full training set, and as such, mini-batch optimization can lead to suboptimal results. Therefore, optimization of classical deep CCA must be performed with full-batch optimization, for example, through the L-BFGS (limited Broyden-Fletcher-Goldfarb-Shanno) scheme [47]. For this reason, with this vanilla implementation, deep CCA is not computationally viable for large datasets. Furthermore, this approach does not provide a model for generating samples from the latent space. To address these issues, Wang et al. [48] introduced deep variational CCA (VCCA) which extends the probabilistic CCA framework introduced in Subheading 3 to a nonlinear generative model. In a similar approach to VAEs and mcVAE, deep VCCA uses variational inference to approximate the posterior distribution and derives the following ELBO:

$$\mathcal{L} = -D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_1) \| p(\boldsymbol{z})\right) + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_1)}\left[\log p_{\boldsymbol{\theta}_1}(\boldsymbol{x}_1 \mid \boldsymbol{z}) + \log p_{\boldsymbol{\theta}_2}(\boldsymbol{x}_2 \mid \boldsymbol{z})\right] \quad (30)$$

where the approximate posterior, $q_\phi(\boldsymbol{z}|\boldsymbol{x}_1)$, and likelihood distributions, $p_{\boldsymbol{\theta}_1}(\boldsymbol{x}_1 \mid \boldsymbol{z})$ and $p_{\boldsymbol{\theta}_2}(\boldsymbol{x}_2 \mid \boldsymbol{z})$, are parameterized by neural networks with parameters ϕ, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$.

We note that, in contrast to mcVAE, deep VCCA is based on the estimation of a single latent posterior distribution. Therefore, the resulting representation is dependent on the reference modality from which the joint latent representation is encoded and may therefore bias the estimation of the latent representation. Finally Wang et al. [48] introduce a variant of deep VCCA, VCCA-private, which extracts the private, in addition to shared, latent information. Here, private latent variables hold view-specific information which is not shared across modalities.

4 Biologically Inspired Data Integration Strategies

Medical imaging and -omics data are characterized by nontrivial relationships across features, which represent specific mechanisms underlying the pathophysiological processes.

For example, the pattern of brain atrophy and functional impairment may involve brain regions according to the brain connectivity structure [49]. Similarly, biological processes such as gene expression are the result of the joint contribution of several SNPs acting according to *biological pathways*. According to these processes, it is possible to establish relationships between genetics features under the form of relation networks, represented by ontologies such as the KEGG pathways[2] and the Gene Ontology Consortium.[3]

When applying data-driven multivariate analysis methods to this kind of data, it is therefore relevant to promote interpretability and plausibility of the model, by enforcing the solution to follow the structural constraints underlying the data. This kind of model behavior can be achieved through *regularization* of the model parameters.

In particular, group-wise regularization [50] is an effective approach to enforce structural patterns during model optimization, where related features are jointly penalized with respect to a common parameter. For example, group-wise constraints may be introduced to account for biological pathways in models of gene association, or for known brain networks and regional interactions in neuroimaging studies. More specifically, we assume that the D_i features of a modality $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD_i})$ are grouped in subsets $\{\mathcal{S}_l\}_{l=1}^{L}$, according to the indices $\mathcal{S}_l = (s_1, \ldots, s_{N_l})$. The regularization of the of the general multivariate model of Eq. 2 according to the group-wise constraint can be expressed as:

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}}{\operatorname{argmin}} \, \|\boldsymbol{X}_1 - \boldsymbol{X}_2 \cdot \boldsymbol{W}\|^2 + \lambda \sum_{l=1}^{L} \beta_l R(\boldsymbol{W}_l), \tag{31}$$

where $R(\boldsymbol{W}_l) = \sum_{j=1}^{D_1} \sqrt{\sum_{s \in \mathcal{S}_l} \boldsymbol{W}[s,j]^2}$ is the penalization of the entries of $\boldsymbol{W}$ associated with the features of $\boldsymbol{X}_2$ indexed by $\mathcal{S}_l$. The total penalty is achieved by the sum across the D_1 columns.

Group-wise regularization is particularly effective in the following situations:

- To compensate for large data dimensionality, by reducing the number of "free parameters" to be optimized by aggregating the available features [51].

[2] https://www.genome.jp/kegg/pathway.html.

[3] http://geneontology.org/.

- To account for the small effect size of each independent features, to combine features in order to increase the detection power. For example, in genetic analysis, each SNP accounts for below 1% of the variance in brain imaging quantitative traits when considered individually [52, 53].
- To meaningfully integrate complementary information to introduce biologically inspired constraints into the model.

In the context of group-wise regularization in neural networks, several optimization/regularization strategies have been proposed to allow the identification of compressed representation of multimodal data in the bottleneck layers, such as by imposing sparsity of the model parameters or by introducing grouping constraints motivated by prior knowledge [54].

For instance, the Bayesian Genome-to-Phenome Sparse Regression (G2PSR) method proposed in [55] associates genomic data to phenotypic features, such as multimodal neuroimaging and clinical data, by constraining the transformation to optimize relevant group-wise SNPs-gene associations. The resulting architecture groups the input SNP layer into corresponding genes represented in the intermediate layer $\boldsymbol{L}$ of the network (Fig. 6). Sparsity at the gene level is introduced through variational dropout [56], to estimate the relevance of each gene (and related SNPs) in reconstructing the output phenotypic features.

In more detail, to incorporate biological constraints in G2PSR framework, a group-wise penalization is imposed with nonzero weights $\boldsymbol{W}^{g}$ mapping the input SNPs to their common gene g. The idea is that during optimization the model is forced to jointly discard all the SNPs mapping to genes which are not relevant to the predictive task. Following [56], the variational approximation is

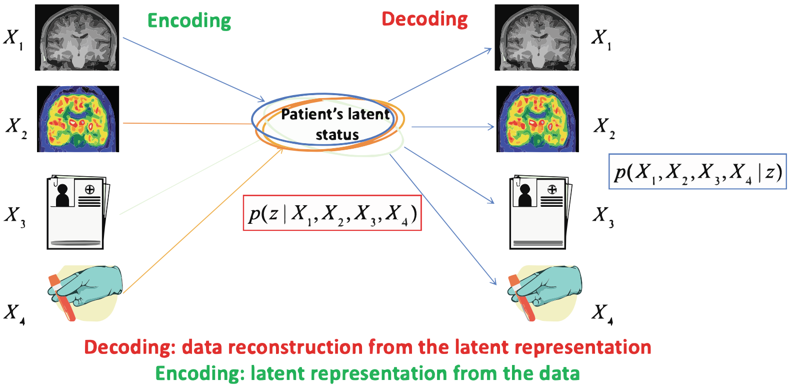
Encoding
Decoding
X_1
X_2
X_3
X_4
Patient's latent status
$p(z \mid X_1, X_2, X_3, X_4)$
$p(X_1, X_2, X_3, X_4 \mid z)$
X_1
X_2
X_3
X_4
Decoding: data reconstruction from the latent representation
Encoding: latent representation from the data

Fig. 5 The multi-channel VAE (mcVAE) for the joint modeling of multimodal medical imaging, clinical, and biological information. The mcVAE approximates the latent posterior $p(\boldsymbol{z}|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4)$ to maximize the likelihood of the data reconstruction $p(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4|\boldsymbol{z})$ (plus a regularization term)

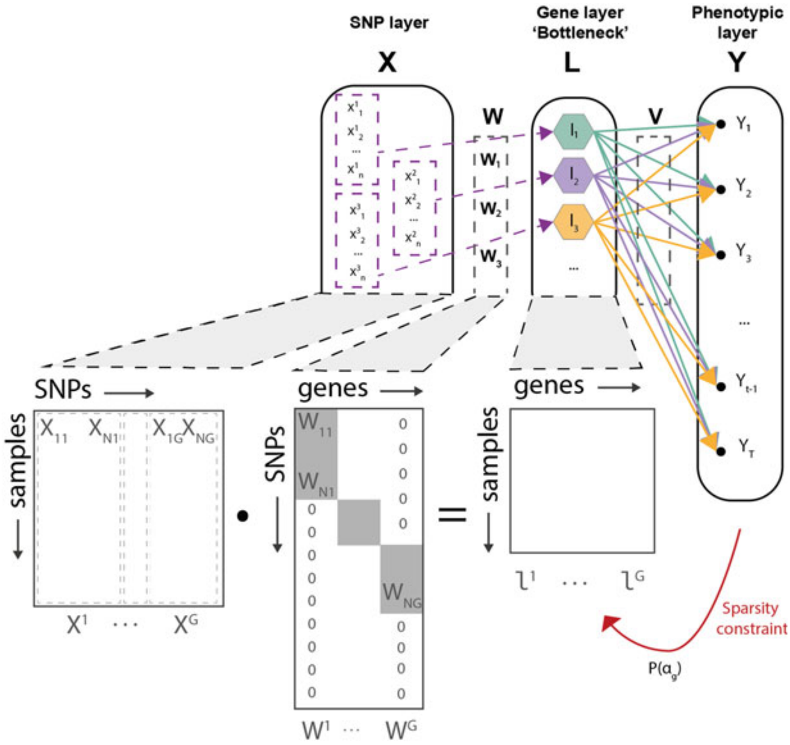

Fig. 6 Illustration of G2PSR SNP-gene grouping constraint and overall neural network architecture

parametrized as $q(W^g)$, such that each element of the input layer is defined as $W_i^g \sim \mathcal{N}(\mu_i^g; \alpha_g.\mu_i^{g^2})$ [57], where the parameter α_g is optimized to quantify the common uncertainty associated with the ensemble of SNPs contributing to the gene g.

5 Conclusions

This chapter presented an overview of basic notions and tools for multimodal analysis. The set of frameworks introduced here represents an ideal starting ground for more complex analysis, either based on linear multivariate methods [58, 59] or on neural network architectures, extending the modeling capabilities to account for highly heterogeneous information, such multi-organ data [46], text information, and data from electronic health records [60, 61].

Acknowledgements

This work was supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) (ANR-19-P3IA-0002).

References

1. Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. Nat Rev Gen 15(1):34–48. https://doi.org/10.1038/nrg3575
2. Liu S, Cai W, Liu S, Zhang F, Fulham M, Feng D, Pujol S, Kikinis R (2015) Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders. Brain Inform 2(3):167–180. https://doi.org/10.1007/s40708-015-0019-x
3. Shen L, Thompson PM (2020) Brain imaging genomics: Integrated analysis and machine learning. Proc IEEE Inst Electr Electron Eng 108(1):125–162. https://doi.org/10.1109/JPROC.2019.2947272
4. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101(1):5–22. https://doi.org/10.1016/j.ajhg.2017.06.005
5. Lahat D, Adali T, Jutten C (2014) Challenges in multimodal data fusion. In: EUSIPCO 2014—22th European signal processing conference, Lisbonne, Portugal, pp 101–105. https://hal.archives-ouvertes.fr/hal-01062366
6. Menon BK, Campbell BC, Levi C, Goyal M (2015) Role of imaging in current acute ischemic stroke workflow for endovascular therapy. Stroke 46(6):1453–1461. https://doi.org/10.1161/STROKEAHA.115.009160
7. Zameer S, Siddiqui AS, Riaz R (2021) Multimodality imaging in acute ischemic stroke. Curr Med Imaging 17(5):567–577
8. Liu X, Lai Y, Wang X, Hao C, Chen L, Zhou Z, Yu X, Hong N (2013) A combined DTI and structural MRI study in medicated-naïve chronic schizophrenia. Magn Reson Imaging 32(1):1–8
9. Rastogi S, Lee C, Salamon N (2008) Neuroimaging in pediatric epilepsy: a multimodality approach. Radiographics 28(4):1079–1095
10. Abela E, Rummel C, Hauf M, Weisstanner C, Schindler K, Wiest R (2014) Neuroimaging of epilepsy: lesions, networks, oscillations. Clin Neuroradiol 24(1):5–15
11. Fernández S, Donaire A, Serès E, Setoain X, Bargalló N, Falcén C, Sanmartí F, Maestro I, Rumià J, Pintor L, Boget T, Aparicio J, Carreño M (2015) PET/MRI and PET/MRI/SISCOM coregistration in the presurgical evaluation of refractory focal epilepsy. Epilepsy Research 111:1–9. https://doi.org/10.1016/j.eplepsyres.2014.12.011
12. Hong SB, Zalesky A, Fornito A, Park S, Yang YH, Park MH, Song IC, Sohn CH, Shin MS, Kim BN, Cho SC, Han DH, Cheong JH, Kim JW (2014) Connectomic disturbances in attention-deficit/hyperactivity disorder: a whole-brain tractography analysis. Biol Psychiatry 76(8):656–663
13. Mueller S, Keeser D, Samson AC, Kirsch V, Blautzik J, Grothe M, Erat O, Hegenloh M, Coates U, Reiser MF, Hennig-Fast K, Meindl T (2013) Convergent findings of altered functional and structural brain connectivity in individuals with high functioning autism: a multimodal mri study. PLOS ONE 8(6):1–11. https://doi.org/10.1371/journal.pone.0067329
14. Lorenzi M, Altmann A, Gutman B, Wray S, Arber C, Hibar DP, Jahanshad N, Schott JM, Alexander DC, Thompson PM, Ourselin S, null null (2018) Susceptibility of brain atrophy to *TRIB3* in Alzheimer's disease, evidence from functional prioritization in imaging genetics. Proc Natl Acad Sci 115(12):3162–3167. https://doi.org/10.1073/pnas.1706100115
15. Kim M, Kim J, Lee SH, Park H (2017) Imaging genetics approach to Parkinson's disease and its correlation with clinical score. Sci Rep 7(1):46700. https://doi.org/10.1038/srep46700
16. Martins D, Giacomel A, Williams SC, Turkheimer F, Dipasquale O, Veronese M, Group PTW, et al. (2021) Imaging transcriptomics: convergent cellular, transcriptomic, and molecular neuroimaging signatures in the healthy adult human brain. Cell Rep 37(13):110173
17. Schrouff J, Rosa MJ, Rondina JM, Marquand AF, Chu C, Ashburner J, Phillips C, Richiardi J, Mourão-Miranda J (2013) PRoNTo: pattern recognition for neuroimaging toolbox. Neuroinformatics 11(3):319–337
18. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV (2009) Machine learning in genome-wide association studies. Genetic Epidemiol 33(S1):S51–S57
19. Liu J, Calhoun VD (2014) A review of multivariate analyses in imaging genetics. Front Neuroinform 8:29
20. Lorenzi M, Altmann A, Gutman B, Wray S, Arber C, Hibar DP, Jahanshad N, Schott JM, Alexander DC, Thompson PM, Ourselin S (2018) Susceptibility of brain atrophy to trib3 in Alzheimer's disease, evidence from

functional prioritization in imaging genetics. Proc Natl Acad Sci 115(12):3162–3167. https://doi.org/10.1073/pnas.1706100115

21. Shashanka M, Raj B, Smaragdis P (2007) Sparse overcomplete latent variable decomposition of counts data. In: Advances in neural information processing systems, vol 20
22. Anandkumar A, Ge R, Janzamin M (2015) Learning overcomplete latent variable models through tensor methods. In: Conference on learning theory, PMLR, pp 36–112
23. Antelmi L, Ayache N, Robert P, Lorenzi M (2019) Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In: International conference on machine learning, PMLR, pp 302–311
24. Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3/4):321
25. Liu J, Calhoun V (2014) A review of multivariate analyses in imaging genetics. Front Neuroinform 8:29. https://doi.org/10.3389/fninf.2014.00029
26. Kettenring JR (1971) Canonical analysis of several sets of variables. Biometrika 58(3): 433–451. https://doi.org/10.1093/biomet/58.3.433
27. Luo Y, Tao D, Ramamohanarao K, Xu C, Wen Y (2015) Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Trans Knowl Data Eng 27(11):3111–3124. https://doi.org/10.1109/TKDE.2015.2445757
28. Huang SY, Lee MH, Hsiao CK (2009) Nonlinear measures of association with kernel canonical correlation analysis and applications. J Stat Plan Inference 139(7):2162–2174. https://doi.org/10.1016/j.jspi.2008.10.011
29. Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In: Dasgupta S, McAllester D (eds) Proceedings of the 30th international conference on machine learning, PMLR, Atlanta, Georgia, USA, Proceedings of Machine Learning Research, vol 28, pp 1247–1255. https://proceedings.mlr.press/v28/andrew13.html
30. McIntosh A, Bookstein F, Haxby JV, Grady C (1996) Spatial pattern analysis of functional brain images using partial least squares. Neuroimage 3(3):143–157
31. Worsley KJ (1997) An overview and some new developments in the statistical analysis of pet and fmri data. Hum Brain Mapp 5(4):254–258
32. De Bie T, Cristianini N, Rosipal R (2005) Eigenproblems in pattern recognition. In: Handbook of geometric computing, pp 129–167
33. Bach F, Jordan M (2003) Kernel independent component analysis. J Mach Learn Res 3:1–48. https://doi.org/10.1162/153244303768966085
34. Theodoridis S, Koutroumbas K (2008) Pattern recognition, 4th edn. Academic Press, New York
35. Wold H (1975) Path models with latent variables: the nipals approach. In: Quantitative sociology. Elsevier, Amsterdam, pp 307–357
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
37. Tenenhaus M (1999) L'approche pls. Revue de statistique appliquée 47(2):5–40
38. Vidaurre D, van Gerven MA, Bielza C, Larrañaga P, Heskes T (2013) Bayesian sparse partial least squares. Neural Comput 25(12): 3318–3339
39. Klami A, Virtanen S, Kaski S (2013) Bayesian canonical correlation analysis. J Mach Learn Res 14(4):965–1003
40. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Soc Series B (Statistical Methodology) 61(3): 611–622
41. Balelli I, Silva S, Lorenzi M (2021) A probabilistic framework for modeling the variability across federated datasets of heterogeneous multi-view observations. In: Information processing in medical imaging: proceedings of the...conference.
42. Kingma DP, Welling M (2014) Auto-Encoding Variational Bayes. In: Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014) 1312.6114
43. Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. PMLR, pp 1278–1286
44. Kim Y, Wiseman S, Miller A, Sontag D, Rush A (2018) Semi-amortized variational autoencoders. In: International conference on machine learning. PMLR, pp 2678–2687
45. Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: a review for statisticians. J Am Stat Assoc 112(518):859–877
46. Diaz-Pinto A, Ravikumar N, Attar R, Suinesiaputra A, Zhao Y, Levelt E, Dall'Armellina E, Lorenzi M, Chen Q, Keenan TD et al (2022) Predicting myocardial infarction through retinal scans and minimal personal information. Nat Mach Intell 4:55–61

47. Nocedal J, Wright S (2006) Numerical optimization. Springer nature, pp 1–664. Springer series in operations research and financial engineering
48. Wang W, Lee H, Livescu K (2016) Deep variational canonical correlation analysis. http://arxiv.org/abs/1610.03454
49. Hafkemeijer A, Altmann-Schneider I, Oleksik AM, van de Wiel L, Middelkoop HA, van Buchem MA, van der Grond J, Rombouts SA (2013) Increased functional connectivity and brain atrophy in elderly with subjective memory complaints. Brain Connectivity 3(4): 353–362
50. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Series B (Statistical Methodology) 68(1):49–67
51. Zhang Y, Xu Z, Shen X, Pan W, Initiative ADN (2014) Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. NeuroImage 96:309–325. https://doi.org/10.1016/j.neuroimage.2014.03.061
52. Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ, Potkin SG, Jack Jr CR, Weiner MW, Toga AW, Thompson PM, Initiative ADN (2011) Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. NeuroImage 56(4): 1875–1891. https://doi.org/10.1016/j.neuroimage.2011.03.077
53. Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE (2012) Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. NeuroImage 63:858–873
54. Schmidt W, Kraaijveld M, Duin R (1992) Feedforward neural networks with random weights. In: Proceedings of the 11th IAPR international conference on pattern recognition. Vol. II. Conference B: pattern recognition methodology and systems, pp 1–4. https://doi.org/10.1109/ICPR.1992.201708
55. Deprez M, Moreira J, Sermesant M, Lorenzi M (2022) Decoding genetic markers of multiple phenotypic layers through biologically constrained genome-to-phenome Bayesian sparse regression. Front Mol Med. https://doi.org/10.3389/fmmed.2022.830956
56. Molchanov D, Ashukha A, Vetrov D (2017) Variational dropout sparsifies deep neural networks. arXiv 1701.05369
57. Kingma DP, Welling M (2014) Auto-encoding variational bayes. CoRR abs/1312.6114
58. Pearlson GD, Liu J, Calhoun VD (2015) An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. Front Genetics 6:276
59. Le Floch É, Guillemot V, Frouin V, Pinel P, Lalanne C, Trinchera L, Tenenhaus A, Moreno A, Zilbovicius M, Bourgeron T et al (2012) Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. Neuroimage 63(1):11–24
60. Rodin I, Fedulova I, Shelmanov A, Dylov DV (2019) Multitask and multimodal neural network model for interpretable analysis of x-ray images. In: 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 1601–1604
61. Huang SC, Pareek A, Zamanian R, Banerjee I, Lungren MP (2020) Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep 10(1):1–9