Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data

Luigi Antelmi [1] **Nicholas Ayache** [1] **Philippe Robert** [2 3] **Marco Lorenzi** [1]
for the Alzheimer's Disease Neuroimaging Initiative [*]

Abstract

Interpretable modeling of heterogeneous data channels is essential in medical applications, for example when jointly analyzing clinical scores and medical images. Variational Autoencoders (VAE) are powerful generative models that learn representations of complex data. The flexibility of VAE may come at the expense of lack of interpretability in describing the joint relationship between heterogeneous data. To tackle this problem, in this work we extend the variational framework of VAE to bring parsimony and interpretability when jointly account for latent relationships across multiple channels. In the latent space, this is achieved by constraining the variational distribution of each channel to a common target prior. Parsimonious latent representations are enforced by variational dropout. Experiments on synthetic data show that our model correctly identifies the prescribed latent dimensions and data relationships across multiple testing scenarios. When applied to imaging and clinical data, our method allows to identify the joint effect of age and pathology in describing clinical condition in a large scale clinical cohort.

1. Introduction

Understanding the relationship among heterogeneous data is essential in medical applications, where performing a diagnosis, or understanding the dynamics of a pathology require to jointly analyze multiple data channels, such as demographic data, medical imaging data, and psychological tests.

Multivariate methods to jointly analyze heterogeneous data, such as Partial Least Squares (PLS), Reduced Rank Regression (RRR), or Canonical correlation analysis (CCA) (Hotelling, 1936) have successfully been applied in biomedical research (Liu & Calhoun, 2014), along with multi-channel (Kettenring, 1971; Luo et al., 2015) and non-linear variants (Huang et al., 2009; Andrew et al., 2013). These approaches are classified as *recognition* methods, as their common formulation consists in projecting the observations in a latent low dimensional space in which desired characteristics are enforced, such as maximum correlation (CCA), maximum covariance (PLS), or minimum regression error (RRR) (Haufe et al., 2014). In their classical formulation these models are not *generative* as they do not explicitly provide a mean to sample observations when the distribution of latent variables and parameters is known. *Bayesian-CCA* (Klami et al., 2013) actually goes in this direction: it is a generative formulation of CCA, where a transformation of a latent variable captures the shared variation between data channels. A limitation of this method for the application in real data scenarios is scalability, as inference on the posterior distribution results in $\mathcal{O}(D^3)$ complexity, being D the dimensionality of the data. Consequently, all the practical applications of Bayesian CCA in the earlier works were limited to very few dimensions and channels (Klami & Kaski, 2007).

Variational autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014) are models that couple a recognition function, or *encoder*, to infer a lower dimensional representation of the data, with a generative function, or *decoder*, which transforms the latent representation back to the original observation space. The VAE is a Bayesian model: the latent variables are inferred by estimating the associated posterior distributions. Inference is efficiently performed through

[*] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[1]University of Côte d'Azur, Inria, Epione Project-Team, France. [2]University of Côte d'Azur, CoBTeK, France. [3]Centre Mémoire, CHU of Nice, France. Correspondence to: Luigi Antelmi <luigi.antelmi@inria.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

amortized inference (Kim et al., 2018) by parametrizing the posterior moments with neural networks. The networks are optimized to maximize the associated evidence lower bound (ELBO). VAEs are flexible and can account for any kind of data. Within this setting, the joint analysis of heterogeneous channels can be performed through concatenation of the different data sources. However, modeling concatenated multi-channel data through a VAE may pose interpretability issues, as it is difficult to disentangle the contribution of a single channel in the description of the latent representation. Moreover, at test time, the model can usually be applied only to data presenting all the channels information. To tackle this problem, in this work we generalize the VAE by assuming that in a multi-channel scenario the latent representation associated to each channel must match a common target distribution. This is done by imposing a constraint on the latent representations in an information theoretical sense, where each latent representation is enforced to match a common target prior. We will show that this constraint can be optimized within a variational optimization framework, allowing efficient inference of channel encodings and latent representation.

Another limitation of the VAE concerns the interpretability of the latent space. In particular, we generally lack of a theoretical justification for the choice of the latent space dimension. This is a key parameter that can profoundly impact the interpretability of the estimated data representation. The optimization of the latent dimension through cross-validation may also pose generalization problems, especially when the data is scarce. To tackle this issue, in this work we investigate a principled theoretical framework for imposing parsimonious representations of the latent space through sparsity constraints. We argue that this kind of model may lead not only to improved interpretability, but also to optimal data representation. Indeed, it is known that VAEs suffer from the problem of *over-pruning*: the variational approximation leads to overly simplified representations, resulting in high model bias due to the impossibility to learn latent distribution different from the prior (Burda et al., 2015; Alemi et al., 2017). As discussed in (Yeung et al., 2017), over-pruning is a recurrent phenomenon ultimately leading to excessive regularization, even in cases when the model underfits the data. The authors tackle over-pruning with the introduction of a categorical sampler on the latent space dimensions. Another way to tackle over-pruning is to enforce sparsity on the latent space. Recently (Kingma et al., 2015; Molchanov et al., 2017) showed that *dropout*, a technique that regularize neural networks, can be naturally embedded in VAE to lead to a sparse representation of the variational parameters.

In our work, we leverage on these recent results to enforce sparsity on the proposed multi-channel VAE. In the variational formulation, the dropout parameters are not hyperparameters anymore, and can be directly learned through the optimization of the variational constraint. Code developed in Pytorch (Paszke et al., 2017) is publicly available at `https://gitlab.inria.fr/epione_ML/mcvae`.

The rest of the paper is organized as follows. In Section 2 we first describe the Multi-Channel Variational Autoencoder and mathematically derive the variational constraint as an extension of the VAE framework. The sparse representation of the latent space is further analyzed and discussed. In Section 3 we show results on extensive synthetic experiments comparing our model to standard non-sparse VAE formulations. We conclude the Section with the application of our model to real data, related to clinical cases of brain neurodegeneration. We show how the learned dropout parameter can be used to automatically identify meaningful latent effect of age and pathology, allowing to predict clinical diagnosis in Alzheimer's Disease (AD). Finally, we summarize our work and propose future extensions.

2. Method

We first describe the proposed Multi-Channel Variational Autoencoder (§2.1). In §2.2 we present the sparse formulation of our method.

2.1. Multi-Channel Variational Autoencoder

Let $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_C\}$ be an observation set of C channels, where each $\mathbf{x}_c$ is a d-dimensional vector. Also, let $\mathbf{z}$ denote the l-dimensional latent variable commonly shared by each $\mathbf{x}_c$. We assume the following generative process for the observation set:

$$\begin{aligned} &\mathbf{z} \sim p\left(\mathbf{z}\right), \\ &\mathbf{x}_c \sim p\left(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c\right), \qquad \text{for } c \text{ in } 1 \dots C, \end{aligned} \tag{1}$$

where $p\left(\mathbf{z}\right)$ is a prior distribution for the latent variable and $p\left(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c\right)$ is a likelihood distribution for the observations conditioned on the latent variable. We assume that the likelihood functions belong to a distribution family $\mathcal{P}$ parametrized by the set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_C\}$.

In the scenario depicted so far, solving the inference problem allows the discovery of the common latent space from which the observed data in each channel is generated. The solution of the inference problem is given by deriving the posterior $p\left(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}\right)$, that is not always computable analytically. In this case, *Variational Inference* can be applied to compute an approximate posterior (Blei et al., 2016).

Our working hypothesis is that every channel brings by itself some information about the latent variable distribution. As such, it makes sense to approximate the posterior distribution with $q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right)$, by conditioning it on the single channel $\mathbf{x}_c$ and on its variational parameters $\boldsymbol{\phi}_c$. Since each channel provides a different approximation, we can impose a constraint enforcing each $q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)$ to be as close as possible to the target posterior distribution. Being the mismatch measured in terms of Kullback-Leibler (KL) divergence, we specify this constraint as:

$$\underset{q\in\mathcal{Q}}{arg\,min}\ \mathbb{E}_c\left[\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)||\,p\left(\mathbf{z}|\mathbf{x}_1,\dots,\mathbf{x}_C,\boldsymbol{\theta}\right)\big)\right], \tag{2}$$

where the approximate posteriors $q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)$ belong to a distribution family $\mathcal{Q}$ parametrized by the set of parameters $\boldsymbol{\phi}=\{\boldsymbol{\phi}_1,\dots,\boldsymbol{\phi}_C\}$, and represent the view on the latent space that can be inferred from each channel $\mathbf{x}_c$. The quantity $\mathbb{E}_c$ is the average over channels computed empirically. Practically, solving the objective in Eq. (2) allows to minimize the discrepancy between the variational approximations and the target posterior. In §2.1.1 we show that the optimization (2) is equivalent to the optimization of the following evidence lower bound $\mathcal{L}\left(\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{x}\right)$:

$$\mathcal{L}\left(\boldsymbol{\theta},\boldsymbol{\phi},\mathbf{x}\right)=\mathbb{E}_c\left[L_c-\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c\right)||\,p\left(\mathbf{z}\right)\big)\right], \tag{3}$$

where $L_c=\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c)}\sum_{i=1}^{C}\ln p\left(\mathbf{x}_i|\mathbf{z},\boldsymbol{\theta}_i\right)$ is the expected log-likelihood of decoding each channel from the latent representation of the channel $\mathbf{x}_c$ only. This formulation is valid for any distribution family $\mathcal{P}$ and $\mathcal{Q}$.

2.1.1. Derivation of the Evidence Lower Bound

In the following derivation we omit the variational and generative parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ to leave the notation uncluttered.

The formula in (2) states that variational inference is carried out by introducing a set of probability density functions $q\left(\mathbf{z}|\mathbf{x}_c\right)$, belonging to a distribution family $\mathcal{Q}$, that are as close as possible to the target posterior over the latent variable $p\left(\mathbf{z}|\mathbf{x}=\{\mathbf{x}_1,\dots,\mathbf{x}_C\}\right)$. Given the intractability of $p\left(\mathbf{z}|\mathbf{x}\right)$ for most complex models, we cannot solve directly this optimization problem. We look then for an equivalent problem, by rearranging the objective:

$$\begin{aligned}
&\mathbb{E}_c\left[\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c\right)||\,p\left(\mathbf{z}|\mathbf{x}\right)\big)\right]=\\
&=\mathbb{E}_c\int_{\mathbf{z}}q\left(\mathbf{z}|\mathbf{x}_c\right)\big(\ln q\left(\mathbf{z}|\mathbf{x}_c\right)-\ln p\left(\mathbf{z}|\mathbf{x}\right)\big)\,d\mathbf{z}\\
&=\mathbb{E}_c\int_{\mathbf{z}}q\left(\mathbf{z}|\mathbf{x}_c\right)\\
&\qquad\big(\ln q\left(\mathbf{z}|x_c\right)-\ln p\left(\mathbf{x}|\mathbf{z}\right)-\ln p\left(\mathbf{z}\right)+\ln p\left(\mathbf{x}\right)\big)\,d\mathbf{z}\\
&=\ln p\left(\mathbf{x}\right)+\\
&\qquad\mathbb{E}_c\left[\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c\right)||\,p\left(\mathbf{z}\right)\big)-\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)}\left[\ln p\left(\mathbf{x}|\mathbf{z}\right)\right]\right],
\end{aligned}$$

where we factorize the true posterior $p\left(\mathbf{z}|\mathbf{x}\right)$ using Bayes' theorem. We can reorganize the terms, such that:

$$\begin{aligned}
&\ln p\left(\mathbf{x}\right)-\underbrace{\mathbb{E}_c\left[\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c\right)||\,p\left(\mathbf{z}|\mathbf{x}\right)\big)\right]}_{\geq 0}=\\
&=\underbrace{\mathbb{E}_c\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)}\left[\ln p\left(\mathbf{x}|\mathbf{z}\right)\right]-\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c\right)||\,p\left(\mathbf{z}\right)\big)\right]}_{\text{lower bound }\mathcal{L}}.
\end{aligned} \tag{4}$$

Since the $\mathcal{D}_{KL}$ term on the left hand side is always non-negative, the right hand side is a lower bound of the log evidence. Thus, by maximizing the lower bound we also maximize the data log evidence while solving the minimization problem in (2).

We note that the lower bound (4) is composed by a regularization term and a data matching term. The $\mathcal{D}_{KL}$ term minimizing the mismatch between the approximate distribution and the target prior acts as a regularizer. The inner expectation term favors the approximate posterior that maximizes the data log-likelihood.

The hypothesis that every channel is conditionally independent from all the others given $\mathbf{z}$ allows to factorize the data likelihood as $p\left(\mathbf{x}|\mathbf{z}\right)=\prod_{i=1}^{C}p\left(\mathbf{x}_i|\mathbf{z}\right)$, so that the lower bound becomes:

$$\begin{aligned}
&\mathcal{L}=\mathbb{E}_c\left[L_c-\mathcal{D}_{KL}\big(q\left(\mathbf{z}|\mathbf{x}_c\right)||\,p\left(\mathbf{z}\right)\big)\right]\\
&\text{where}\quad L_c=\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)}\left[\textstyle\sum_{i=1}^{C}\ln p\left(\mathbf{x}_i|\mathbf{z}\right)\right].
\end{aligned}$$

2.1.2. Comparison with VAE

Our model extends the VAE: the novelty is in the log-likelihood terms L_c in Eq. (3), representing the reconstruction of the multi-channel data from a single channel only. In case $C=1$ the model collapses to a VAE. In the case $C>1$, the L_c terms considered altogether force each channel to the joint decoding of itself and every other channel at the same time. This characteristic allows to reconstruct missing channels $\{\hat{\mathbf{x}}_i\}$ from the available ones $\{\tilde{\mathbf{x}}_j\}$ as:

$$\hat{\mathbf{x}}_i=\mathbb{E}_j\left[\mathbb{E}_{q(\mathbf{z}|\tilde{\mathbf{x}}_j)}\left[p\left(\mathbf{x}_i|\mathbf{z}\right)\right]\right]. \tag{5}$$

An application of Eq. (5) is provided in §3.4. Our model is different from a VAE where all the channels are concatenated into a single one. In that case there cannot be missing channels if we want to infer the latent space variables, unless recurring to costly data imputation techniques (*cf.* App. F in (Rezende et al., 2014)). Our model is also different from a stack of C independent VAEs, in which the C latent spaces are no more related to each-other. The dependence between encoding and decoding across channels stems from the joint approximation of the posterior distribution (Formula (2)).

2.1.3. Gaussian linear case

Model (1) is completely general and can account for complex non-linear relationships modeled, for example, through deep neural networks. However, for simplicity of interpretation, in what follows we focus our multi-channel variational framework to the *Gaussian Linear Model*. This is a special case, analogous to Bayesian-CCA (Klami et al., 2013), where the members of the variational family $\mathcal{Q}$ and generative family $\mathcal{P}$ are Gaussian parametrized by linear transformations. We define the members of the families $\mathcal{Q}$ and $\mathcal{P}$ as:

$$q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) = \mathcal{N}\left(\mathbf{z}|\mathbf{V}_c^{(\mu)}\mathbf{x}_c, diag(\mathbf{V}_c^{(\sigma)}\mathbf{x}_c)\right), \qquad (6)$$

$$p\left(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c\right) = \mathcal{N}\left(\mathbf{x}_c|\mathbf{G}_c^{(\mu)}\mathbf{z}, diag(\mathbf{g}_c^{(\sigma)})\right), \qquad (7)$$

i.e. factorized multivariate Gaussian distributions whose moments are linear transformations depending on the conditioning variables. $\boldsymbol{\theta}_c = \{\mathbf{G}_c^{(\mu)}, \mathbf{g}_c^{(\sigma)}\}$ and $\boldsymbol{\phi}_c = \{\mathbf{V}_c^{(\mu)}, \mathbf{V}_c^{(\sigma)}\}$ are the parameters to be optimized by maximizing the lower bound in (3).

2.1.4. Optimization of the lower bound

The optimization starts with a random initialization of the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}$ and $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C\}$. The expectations L_c in the Eq. (3) can be computed by sampling from the variational distributions $q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right)$ and, when the prior $p\left(\mathbf{z}\right) = \mathcal{N}\left(\mathbf{0}; \mathbf{I}\right)$, the $\mathcal{D}_{KL}$ term in Eq. (3) can be computed analytically (*cf.* (Kingma & Welling, 2014), appendix 2.A). The maximization of $\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}\right)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is efficiently carried out through minibatch stochastic gradient descent implemented with the backpropagation algorithm. With *Adam* (Kingma & Ba, 2014) we compute adaptive learning rates for the parameters.

2.2. Inducing Sparse Latent Representations

In extensive synthetic experiments with the non-sparse version of the multi-channel model, we found that the lower bound (3) generally reaches the maximum value at convergence when the number of fitted latent dimensions coincide with the true one used to generate the data (*Sup. Mat.*). This procedure provides an heuristic for selecting the latent variable dimensions, and proved to work well in controlled scenarios. However, according to our experience, it fails in most complex cases (*Sup. Mat.*), and is time consuming. Moreover, our trust in the result depends on the tightness between the model evidence and its lower bound: a factor that is not easy to control. To address this issue, we propose here to automatically infer the latent variable dimensions via a sparsity constraint on $\mathbf{z}$. Having a sparse $\mathbf{z}$ as a direct result of one single optimization would be computationally advantageous and it would ease the interpretability of the observation model in (1), as the number of relationships to take into account decreases.

2.2.1. Regularization via Dropout

Dropout (Srivastava et al., 2014) and *DropConnect* (Wan et al., 2013) are techniques for regularizing neural networks. The basic block of a neural network is the *fully connected* layer, composed by a linear transformation of an input vector $\mathbf{z}$ into an output vector $\mathbf{x}$, and a non linearity applied to the components of $\mathbf{x}$. Given a generic linear transformation $\mathbf{x} = \mathbf{G}\mathbf{z}$, with $\mathbf{z}$ and $\mathbf{x}$ column vectors, regularization techniques are based on the multiplication of either $\mathbf{z}$ (dropout) or $\mathbf{G}$ (dropconnect) element-wise by independent Bernoulli random variables. The components of $\mathbf{x}$ are hence computed as:

$$x_i = \sum_k g_{ik}(\xi_k z_k), \qquad \text{(dropout)} \qquad (8)$$

$$x_i = \sum_k (\xi_{ik} g_{ik}) z_k, \qquad \text{(dropconnect)} \qquad (9)$$

where $\xi_k, \xi_{ik} \sim \mathcal{B}(1-p)$ with hyperparameter p known as *drop rate*. The elements x_i are approximately Gaussian for the Lyapunov's central limit theorem (Wang & Manning, 2013), and their distributions takes the form:

$$x_i \sim \mathcal{N}\left(\textstyle\sum_k \theta_{ik}; \alpha \sum_k \theta_{ik}^2\right), \qquad (10)$$

where $\alpha = {}^{p}/_{1-p}$ and $\theta_{ik} = g_{ik} z_k (1-p)$. In *Gaussian dropout* (Wang & Manning, 2013) the regularization is achieved by sampling directly from (10).

2.2.2. Variational Dropout and Sparsity

In the context of the *Variational Autoencoder* (VAE), posterior distributions on the encoder weights w that take the form $w \sim \mathcal{N}\left(\mu; \alpha\mu^2\right)$ are called *dropout posteriors* (Kingma et al., 2015). The authors of (Kingma et al., 2015) show that if the variational posteriors on the encoder weights are dropout posteriors, Gaussian dropout arises from the application of the *local reparameterization trick*, a method introduced to increase the stability of gradients estimation in training. The only prior on w consistent with the optimization of the lower bound is the improper log-scale uniform:

$$p\left(\ln|w|\right) = \text{const} \Leftrightarrow p\left(|w|\right) \propto \frac{1}{|w|}. \qquad (11)$$

With this prior, the $\mathcal{D}_{KL}$ of the dropout posterior depends only on α and can be numerically approximated. In (Molchanov et al., 2017) the authors provide an approximation of $\mathcal{D}_{KL}$, reported in (12), to allow this parameter to be learned through the optimization of the lower bound via gradient-based methods:

$$\begin{gathered} \mathcal{D}_{KL}\big(\mathcal{N}\left(w; \alpha w^2\right) || \, p\left(w\right)\big) \approx \\ \approx -k_1\sigma(k_2 + k_3 \ln\alpha) + 0.5\ln(1+\alpha^{-1}) + k_1 \qquad (12) \\ k_1 = 0.63576 \quad k_2 = 1.87320 \quad k_3 = 1.48695 \\ \sigma(\cdot) \text{ Sigmoid function.} \end{gathered}$$

While the optimization of $\mathcal{D}_{KL}$ promotes $\alpha \to \infty$, the implicit drop rate p tends to 1, meaning that the associated weight w can be discarded. Sparsity arises naturally: large values of w correspond to even larger uncertainty αw^2 because of the quadratic relationship and the tendency of the optimization objective to favors $\alpha \to \infty$; therefore, unless that weight is beneficial for the optimization objective, that is to maximize the data log-likelihood, it will be set to zero.

2.2.3. Sparse Multi-Channel VAE

Compatibly with standard dropout methods, in our Multi-Channel VAE we define a variational approximation of the latent code $\mathbf{z}$. We note that the local reparameterization trick cannot be straightforwardly applied, since its standard formulation would require to transfer the uncertainty to a lower dimensional variable, such as from $\mathbf{G}$ to $\mathbf{x}$ in §2.2.1. We notice however that by choosing a dropout posterior for the elements of $\mathbf{z}$, that is if $z_k \sim \mathcal{N}\left(\mu_k; \alpha\mu_k^2\right)$, the output of the first layer with weights g_{ik} of the decoding transformation, before the non-linearity is applied, follows a Gaussian distribution:

$$x_i \sim \mathcal{N}\left(\textstyle\sum_k g_{ik}\mu_k; \alpha \sum_k g_{ik}^2 \mu_k^2\right), \tag{13}$$

in which the first two moments are as follows:

$$\mathbb{E}\left[x_i\right] = \mathbb{E}\left[\textstyle\sum_k g_{ik} z_k\right] = \sum_k g_{ik}\mu_k, \tag{14}$$

$$\begin{aligned} \operatorname{Var}\left[x_i\right] &= \operatorname{Var}\left[\textstyle\sum_k g_{ik} z_k\right] \\ &= \textstyle\sum_k \operatorname{Var}\left[g_{ik} z_k\right] + \sum_{k,\, j\neq k} \operatorname{Cov}\left[\left(g_{ik} z_k, g_{ij} z_j\right)\right] \\ &= \textstyle\sum_k g_{ik}^2 \alpha \mu_k^2 = \alpha \sum_k g_{ik}^2 \mu_k^2, \end{aligned} \tag{15}$$

with the covariance terms vanishing for the hypothesis of independent elements of $\mathbf{z}$. The analogy with (10) holds when $\theta_{ik} = g_{ik}\mu_k$, and so we can establish a connection with the standard dropout techniques. Specifically, imposing a dropout posterior for the latent code $\mathbf{z}$ is analogous to perform dropout on the latent code itself, and dropconnect on the decoder weights. We therefore define the approximate posteriors $q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right)$ in Eq. (3) and parametrize them to be factorized dropout posteriors, that is, for c in $1 \ldots C$:

$$q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) = \mathcal{N}\left(\boldsymbol{\mu}_c; \operatorname{diag}(\sqrt{\boldsymbol{\alpha}} \odot \boldsymbol{\mu}_c)^2\right), \tag{16}$$

with $\boldsymbol{\mu}_c = \boldsymbol{\phi}_c \mathbf{x}_c$, where parameters $\boldsymbol{\phi} = \{\boldsymbol{\alpha}, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C\}$ include $\boldsymbol{\phi}_c$ linear transformations, specific to channel c, while $\boldsymbol{\alpha}$ is shared among all the channels. Following the considerations of (Kingma et al., 2015), the prior distribution $p\left(\mathbf{z}\right)$ is chosen to be fully factorized by scale-invariant log-uniform priors:

$$p\left(\mathbf{z}\right) = \prod p\left(|z_i|\right), \quad \text{such that} \quad p\left(\ln|z_i|\right) \propto \text{const.} \tag{17}$$

Because of these choices, the $\mathcal{D}_{KL}$ term in Eq. (3) can be easily computed by leveraging on Eq. (12). For the same considerations made in the previous section, we induce a sparse behavior on the components of $\mathbf{z}$ and on the associated decoder parameters (*cfr.* Fig. 1). The variational parameter $\boldsymbol{\alpha}$ can be learned and, as the connection with the dropout techniques is kept, we can leverage on the relationship between $\boldsymbol{\alpha}$ and the dropout rate p to interpret the relative importance of the latent dimensions.

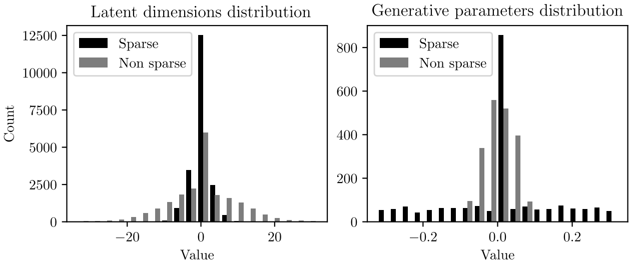


Figure 1: Effect of variational dropout on a synthetic experiment modeled with the Multi-Channel VAE. As expected, the minimum amount of non-zero components of $\mathbf{z}$ (left) and generative parameters $\mathbf{G}$ (right) is obtained with the sparse model.

3. Experiments

We first describe our results on extensive synthetic experiments performed with our non sparse model and with its sparse variant. We benchmark these models with respect to the VAE and conclude the Section with the application of our sparse model to real data, related to clinical cases of neurodegeneration.

3.1. Synthetic Experiments

Datasets $\mathbf{x} = \{\mathbf{x}_c\}$ with $c = 1 \ldots C$ channels where created according to the following model:

$$\begin{aligned} &\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}; \mathbf{I}_l\right), \\ &\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}; \mathbf{I}_{d_c}\right), \\ &\mathbf{G}_c = \mathit{diag}\left(\mathbf{R}_c \mathbf{R}_c^T\right)^{-1/2} \mathbf{R}_c, \\ &\mathbf{x}_c = \mathbf{G}_c \mathbf{z} + \mathit{snr}^{-1/2} \cdot \boldsymbol{\epsilon}, \end{aligned} \tag{18}$$

where for every channel c, $\mathbf{R}_c \in \mathbb{R}^{d_c \times l}$ is a random matrix with l orthonormal columns (*i.e.,* $\mathbf{R}_c^T \mathbf{R}_c = \mathbf{I}_l$), $\mathbf{G}_c$ is the linear generative law, and snr is the signal-to-noise ratio. With this choice, the diagonal elements of the covariance matrix of $\mathbf{x}_c$ are inversely proportional to snr, *i.e.,* $\mathit{diag}\left(\mathbb{E}\left[\mathbf{x}_c \mathbf{x}_c^T\right]\right) = (1 + \mathit{snr}^{-1})\mathbf{I}_{d_c}$. Scenarios where generated by varying one-at-a-time the dataset attributes, as listed in Tab. 1.

ELBO in non-sparse Multi-Channel VAE. For each generated scenario, we optimized multiple instances of a

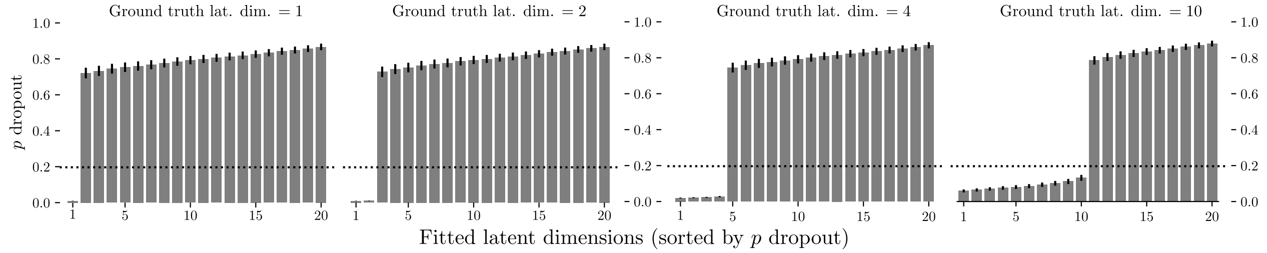


Figure 2: Estimated dropout rates for the latent dimensions when the initial latent dimensions of the Sparse Multi-Channel VAE was set to $l_{\text{fit}} = 20$ on data generated with respectively $l = 1, 2, 4,$ and 10 latent dimensions.

Gaussian Linear Multi-Channel model, as defined in §2.1.3. At convergence, the loss function (negative lower bound) has a minimum when the number of fitted latent dimensions l_{fit} corresponds to the number of the latent dimensions used to generate the data. When increasing the number of fitted latent dimensions, a sudden decrease of the loss (*elbow* effect) is indicative that the true number of latent dimensions has been found. These results are summarized in the *Supplementary Materials*, where we show also that the elbow effect becomes more evident when increasing the number of channels. Ambiguity in identifying the elbow usually arises for high-dimensional data channels.

Table 1: Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (18).

Attribute description	Iteration list
Total channels (C)	2 3 5 10
Channel dimension (d_c)	32
Latent space dimension (l)	1 2 4 10 20
Samples (training and testing)	100 1000
Signal-to-noise ratio (snr)	10 1
Seed (re-initialize $\mathbf{R}_c$)	1 2 3 4 5

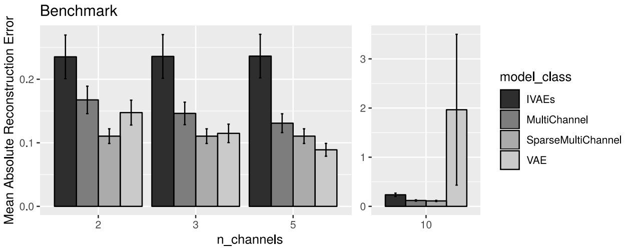


Figure 3: Testing benchmark of four variational methods applied to the multi-channel scenarios in Tab. 1 (cases snr $= 10$, $l_{\text{fit}} = l$). Sparse Multi-Channel models performs consistently better than non-sparse Multi-Channel ones.

Table 2: Benchmark with respect to VAE. (top) Bootstrapped 95% C.I. for the mean absolute error (MAE) difference between each model MAE and the reference MAE of the VAE. (bottom) Average compression factor.

	MCVAE	sMCVAE	IVAEs
95% CI	$[-.13; +.03]$	$[-.12; +.04]$	$[-.10; +.06]$
Compr. Factor	0%	**45%**	0%

3.2. Sparse Multi-Channel VAE Benchmark

This benchmark is based on the data scenarios illustrated in the previous section (Tab. 1). For each generated dataset, we optimized our Multi-Channel VAE with dropout posteriors (eq. 16) associated to log-uniform priors as in (eq. 17).

Results. In Fig. 1 we compare the latent space distributions and the generative parameters derived from the application of the sparse and non-sparse Multi-Channel VAE, after fitting the two models on the same data and by imposing the fitted dimension for the latent space to $l_{\text{fit}} = 20$. As expected, the number of zero elements is considerably higher in the sparse version. We note that the learned dropout rate is very low for the dimensions corresponding to the true latent dimensions used to generate the fitted scenario (Fig. 2). Because of this, model selection can be performed by retaining those latent dimensions satisfying an opportune threshold on the dropout rates. We can see that with the threshold $p < 0.2$, is possible to safely recover the true number of latent dimensions across all the testing scenarios.

3.3. Comparison with VAE

We compared the performance of four variational methods applied to the synthetic scenarios. Besides our sparse (sMCVAE) and non-sparse (MCVAE) Multi-Channel models, we considered a VAE, and a stack of independent VAEs (IVAEs). In the VAE cases, channels where concatenated feature-wise to form a single channel. In IVAEs experiments, every channel was independently modeled with a

Table 3: Proportion of correctly classified ADNI subjects belonging to the testing hold-out dataset. Classification done by means of *Linear Discriminant Analysis* using as training data the latent space inferred with the sparse and non sparse models. 10-fold cross validation mean results shown. Within the sparse framework, we selected the subspace generated by the most relevant latent dimensions identified by variational dropout ($p < 0.2$).

Model:	MCVAE				sMCVAE				IVAEs				VAE			
#layers:	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Normal	.82	.76	.76	.75	.89	.89	**.90**	.79	.78	.77	.77	.79	.81	.82	.78	.77
MCI	.58	.68	.70	.68	**.71**	.70	.68	.67	.65	.67	.69	.66	**.71**	**.71**	.63	**.71**
Dementia	**.88**	.68	.69	.70	.85	.84	.84	.82	.68	.71	.66	.51	.82	.82	.72	.73

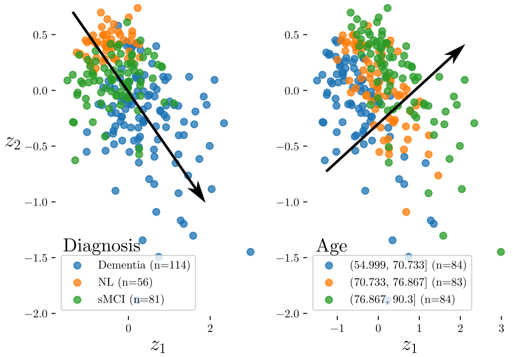


Figure 4: Stratification of the ADNI subjects (test data) in the sparse latent subspace inferred from the first two least dropped out dimensions. In the same subspace it is possible to stratify subjects in the test-set by disease status (left) and by age (right) in almost orthogonal directions. Classification accuracy for these subjects is given in the fifth numeric column of Tab. 3.

VAE. Each scenario was fitted multiple times, by varying the dimension of the fitted latent space l_{fit} in $\{1, 2, 4, 10, 20\}$. The comparison metric is the *mean absolute error* (MAE) between the generated testing data and the predictions from the inferred latent space.

Results. As depicted in Tab. 2, in general there is no significant difference between the average MAE for the different models (95% bootstrap confidence interval). However, when comparing the models in terms of number of parameters, our tests show that sMCVAE leads to equivalent reconstruction by pruning a consistent fraction of the parameters (on average 45%).

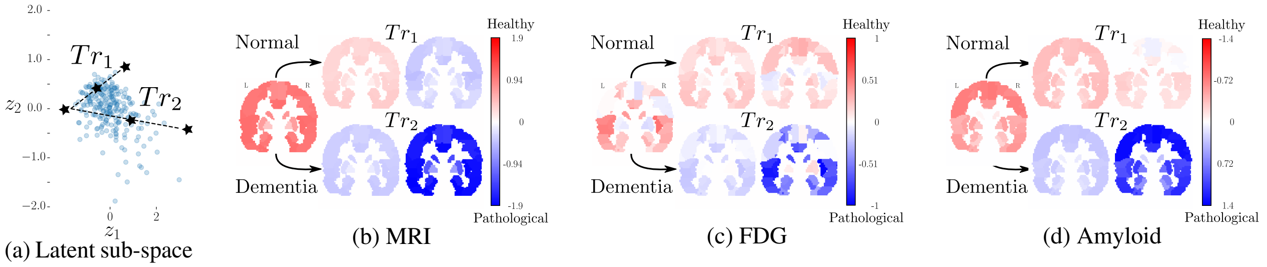
In Fig. 3 we restrict the visualization to the cases where snr $= 10$ and $l_{\text{fit}} = l$ (*cf.* Tab. 1). Sparse Multi-Channel models perform consistently better than the non-sparse ones. Although in some cases VAE seems to provide better results (*cf.* 5-channel case in Fig. 3), in complex cases with many channels the performance of VAE dramatically drops (*cf.* 10-channel case, *ibid.*). The IVAEs models leads to the worst performances in the majority of cases. This is expected, as the generated data variability depends on the joint information across channels. By modeling each channel independently, part of this variability is therefore mistaken as noise.

3.4. Medical Imaging data

3.4.1. Data preparation

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.

We analyzed clinical and imaging channels from 504 subjects of the ADNI cohort. We randomly assigned the subjects to a training and testing set through 10-fold cross validation. The clinical channel was composed by six continuous variables generally recorded in memory clinics: age; results to mini-mental state examination, adas-cog, cdr, and faq tests; scholarity level. The three imaging channels were structural MRI (gray matter only), functional FDG-PET, and Amyloid-PET. Raw data from the imaging channels were coregistered in a common geometric space by means of voxel-based morphometry methods (Ashburner & Friston, 2000). Visual quality check was performed to exclude processing errors. Image intensities were finally averaged over 90 brain regions mapped in the AAL atlas (Tzourio-Mazoyer et al., 2002) to produce 90 features arrays for each image. Lastly, data was centered and standardized across features. Our sparse multi-channel model (§2.2.3) was optimized on the resulting multi-channel dataset, along with MCVAE, IVAEs, and VAE models as described in §3.3. For each model class, multi-layer architectures were tested, ranging from 1 (linear) up to 4 layers for the encoding and decoding pathways, with a sigmoidal activation applied to all but last layer. **Results.** By applying the dropout threshold of 0.2 as identified in the synthetic experiments (Fig. 2), we identify 5 optimal latent dimensions. The encoding of the test set in the latent space given by our sMCVAE model is depicted in Fig. 4, where we limited the visualization to the



(a) Latent sub-space (b) MRI (c) FDG (d) Amyloid

Figure 5: Generation of imaging data from trajectories in the latent space. (a) Normal aging trajectory (Tr_1) *vs* Dementia aging trajectory (Tr_2) in the latent 2D sub-space (*cfr.* Fig. 4). Stars indicate the sampling points along trajectories. The trajectories share the same origin. MRIs (b), FDG (c), and Amyloid PET (d). All the trajectories show a plausible evolution across disease and healthy conditions.

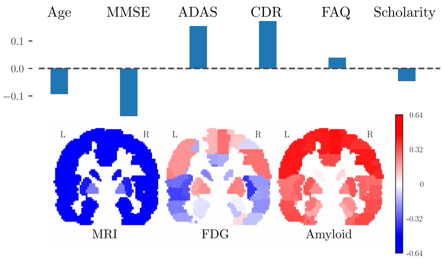


Figure 6: Generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (7)) of the four channels associated to the least dropout latent dimension in the sparse multi-channel model. (Top) Clinical channel parameters. (Bottom) Imaging ch. parameters.

2D subspace generated by the two most relevant dimensions. This subspace appears stratified by age and disease status, across roughly orthogonal directions. We note however that the model was agnostic to the disease status, and was able to correctly stratify the testing data only thanks to the learned latent representation. This is shown in Tab. 3, where the latent representation provided by our sparse Multi-Channel framework leads to competitive predictive performances in predicting the clinical status. Prediction was performed on the testing set via Linear Discriminant Analysis fitted on the training latent space. We note that the predictive accuracy is particularly high with the Multi-Channel framework.

We illustrate the ability of a single layer sMCVAE in reconstructing missing channels by using Eq. (5), to sample the imaging data from the latent dimensions obtained from the clinical channel. To this end, we sample points from two trajectories in the subspace shown in Fig. 4 to predict the imaging data channels. Trajectory 1 (Tr_1) follows an aging path centered on the healthy subject group. Trajectory 2 (Tr_2), starts from the same origin of Tr_1 and follows a path were aging is entangled with the pathological variability. We can see these trajectories and the generated imaging channels in Fig. 5. Fig. 6 shows the generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (7)) of the four channels associated to the most relevant latent dimension identified by dropout. These generative parameters show a plausible relationship across channels, describing a pattern of early onset AD, associated with abnormal scores (low MMSE, high ADAS and CDR), gray matter atrophy emerging from the MRI, low glucose uptake in the temporal lobes as emerging from the FDG-PET, and high amyloid deposits, coherently with the research literature on Alzheimer's Disease (Dubois et al., 2014; Jack et al., 2018).

4. Conclusion

This paper introduces the Sparse Multi-Channel VAE, an extension of variational autoencoders, to jointly account for latent relationships across heterogeneous data. Parsimonious and interpretable representations are enforced by variational dropout, leveraging on sparsity to provide an effective mean to model selection in the latent space. In extensive synthetic experiments, we compared the performance of our model against different configurations of the VAE. We found a generally equivalent or superior performance of our model with respect to the benchmark, associated to a compression factor close to 50% on the number of pruned parameters. In the real case scenario of Alzheimer's Disease modeling, our model allowed the unsupervised stratification of the latent space by disease status and age, providing evidence for a clinically sound interpretation of the latent space. Nonlinear parameterization of the model seemed not to bring clear advantages in the real case dataset, and needs further investigations. Given the scalability of our variational model, application to high resolution images may be also at reach, although this may require to account for full covariance matrices to take into account spatial relationships. To increase the model classification performance, supervised clustering of the latent space can be introduced, for example, through a categorical sampler in the latent space.Lastly, due to the general formulation, the proposed method can find various applications as a general data interpretation technique, not limited to the biomedical research area.

References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a Broken ELBO. nov 2017.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. Deep Canonical Correlation Analysis. *Proc. Mach. Learn. Res.*, 28(3):1247–1255, 2013.

Ashburner, J. and Friston, K. J. Voxel-based morphometry–the methods. *Neuroimage*, 11(6 Pt 1):805–21, jun 2000. ISSN 1053-8119. doi: 10.1006/nimg.2000.0582.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational Inference: A Review for Statisticians. 2016. doi: 10.1080/01621459.2017.1285773.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance Weighted Autoencoders. sep 2015.

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., DeKosky, S. T., Gauthier, S., Selkoe, D., Bateman, R., Cappa, S., Crutch, S., Engelborghs, S., Frisoni, G. B., Fox, N. C., Galasko, D., Habert, M.-o., Jicha, G. A., Nordberg, A., Pasquier, F., Rabinovici, G., Robert, P., Rowe, C., Salloway, S., Sarazin, M., Epelbaum, S., de Souza, L. C., Vellas, B., Visser, P. J., Schneider, L., Stern, Y., Scheltens, P., and Cummings, J. L. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet. Neurol.*, 13(6):614–29, jun 2014. ISSN 1474-4465. doi: 10.1016/S1474-4422(14)70090-0.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.10.067.

Hotelling, H. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321, dec 1936. ISSN 00063444.

Huang, S.-Y., Lee, M.-H., and Hsiao, C. K. Nonlinear measures of association with kernel canonical correlation analysis and applications. *J. Stat. Plan. Inference*, 139(7):2162–2174, 2009. ISSN 03783758. doi: 10.1016/j.jspi.2008.10.011.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., Sperling, R., Elliott, C., Masliah, E., Ryan, L., and Silverberg, N. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's Dement.*, 14(4):535–562, 2018. ISSN 15525260. doi: 10.1016/j.jalz.2018.02.018.

Kettenring, J. R. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971. ISSN 0006-3444. doi: 10.1093/biomet/58.3.433.

Kim, Y., Wiseman, S., Miller, A. C., Sontag, D., and Rush, A. M. Semi-Amortized Variational Autoencoders. feb 2018.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6, 2014.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014).*, dec 2014.

Kingma, D. P., Salimans, T., and Welling, M. Variational Dropout and the Local Reparameterization Trick. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Adv. Neural Inf. Process. Syst. 28*, pp. 2575–2583. Curran Associates, Inc., 2015.

Klami, A. and Kaski, S. Local dependent components. In Ghahramani, Z. (ed.), *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, pp. 425–432. Omnipress, 2007.

Klami, A., Seppo, V., and Kaski, S. Bayesian Canonical Correlation Analysis. *J. Mach. Learn. Res.*, 14:965–1003, 2013. ISSN 1532-4435.

Liu, J. and Calhoun, V. D. A review of multivariate analyses in imaging genetics. *Front. Neuroinform.*, 8:29, mar 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00029.

Luo, Y., Tao, D., Ramamohanarao, K., Xu, C., and Wen, Y. Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction. *IEEE Trans. Knowl. Data Eng.*, 27(11):3111–3124, 2015. ISSN 1041-4347. doi: 10.1109/TKDE.2015.2445757.

Molchanov, D., Ashukha, A., and Vetrov, D. Variational Dropout Sparsifies Deep Neural Networks. *arXiv*, 2017.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. jan 2014.

Srivastava, N., Geoffrey, H., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *Neuroimage*, 15(1):273–289, jan 2002. ISSN 10538119. doi: 10.1006/nimg.2001.0978.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R. Regularization of Neural Networks using DropConnect. In *Proc. 30th Int. Conf. Mach. Learn.*, pp. 1058—-1066, 2013.

Wang, S. and Manning, C. Fast dropout training. *Proc. 30th Int. Conf. Mach. Learn.*, 28(2):118–126, 2013.

Yeung, S., Kannan, A., Dauphin, Y., and Fei-Fei, L. Tackling Over-pruning in Variational Autoencoders. jun 2017.

Acknowledgments

This work has been supported by:

- the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01;
- the grant AAP Santé 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, ”NEF” computation cluster;
- Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.