Statistical learning on heterogeneous medical data with bayesian latent variable models : application to neuroimaging dementia studies

Luigi Antelmi

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Apprentissage statistique sur des données médicales hétérogènes avec des modèles Bayésiens à variables latentes : application aux études de neuroimagerie pour les maladies neurodégénératives

Luigi ANTELMI

INRIA, Équipe EPIONE

Thèse dirigée par Nicholas AYACHE
et co-dirigée par Philippe ROBERT et Marco LORENZI

Soutenue le 7 Juillet 2021

Présentée en vue de l'obtention du grade de DOCTEUR EN AUTOMATIQUE, TRAITEMENT DU SIGNAL ET DES IMAGES de l'UNIVERSITÉ CÔTE D'AZUR.

Devant le jury composé de :

Gloria MENEGAZ	University of Verona	Rapporteur
Pietro MICHIARDI	EURECOM	Rapporteur
Valentina GARIBOTTO	Hôpitaux Universitaires de Genève	Présidente
Nicholas AYACHE	Inria Sophia Antipolis	Directeur de thèse
Philippe ROBERT	Université Côte d'Azur (EA Cobtek)	Co-directeur de thèse
Marco LORENZI	Inria Sophia Antipolis	Co-encadrant

informatics mathematics Inria Epione e-patient / e-medicine

UCA JEDI INITIATIVE D'EXCELLENCE

médecine Numérique mNC3 Cerveau Cognition Comportement

Apprentissage statistique sur des données médicales hétérogènes avec des modèles Bayésiens à variables latentes : application aux études de neuroimagerie pour les maladies neurodégénératives

Statistical Learning on Heterogeneous Medical Data with Bayesian Latent Variable Models: Application to Neuroimaging Dementia Studies

Présidente du jury

Valentina GARIBOTTO Professeur Hôpitaux Universitaires de Genève

Rapporteurs

Gloria MENEGAZ	Professeur	University of Verona
Pietro MICHIARDI	Professeur	EURECOM

Examinateurs

Valentina GARIBOTTO	Professeur	Hôpitaux Universitaires de Genève
Gloria MENEGAZ	Professeur	University of Verona
Pietro MICHIARDI	Professeur	EURECOM
Nicholas AYACHE	Professeur	Inria Sophia Antipolis
Philippe ROBERT	Professeur	Université Côte d'Azur (EA Cobtek)
Marco LORENZI	Chargé de recherche	Inria Sophia Antipolis

Abstract

This thesis presents new computational tools for the joint modeling of multi-modal biomedical data, robust to missing data, with application to neuroimaging studies in dementia. The theoretical base for this work is the Variational Autoencoder (VAE), a latent variable generative model well suited for working with complex data as it forces them into a simpler low-dimensional space, able to model data non-linearities.

The core of this Thesis consists in the Multi-Channel Variational Autoencoder (MCVAE), an extension of the VAE to jointly model latent relationships across multi-modal observations. This is achieved by: 1) constraining the latent distribution of each data modality to a common target prior, 2) forcing these latent distribution to generate all the data modalities through their associated generative functions.

Moreover, we adapt the MCVAE to a Multi-Task setting, where the problem of dealing with missing data is addressed with a specific optimization scheme following these steps: 1) defining tasks across datasets based on the identification of data subsets presenting compatible modalities, 2) stacking multiple instances of the MCVAE, where each instance models a specific task, 3) sharing the models parameters of common modalities between modeling tasks. Thanks to these actions, the Multi-Task MCVAE allows to learn a joint model for all the data points leveraging on all the available information.

Overall, this thesis provides a novel investigation of flexible approaches to account for data heterogeneity in the analysis of biomedical information. This work enables new research directions in which medical information can be consistently modeled within a joint probabilistic framework accounting for multiple data modalities, missing information, and biases across different datasets.

Lastly, thanks to their general formulation, the methodologies here proposed can find applications beyond the neuroimaging research field.

Keywords: Alzheimer's Disease (AD); Neuro-imaging (NI); Magnetic Resonance Imaging (MRI); Positron Emission Tomography (PET); Variational Autoencoder (VAE); Multi-Task Learning (MTL); High Dimensionality; Heterogeneous Data.

Résumé

Cette thèse présente de nouveaux outils informatiques pour la modélisation conjointe de données biomédicales multimodales, robustes aux données manquantes, avec une application aux études de neuro-imagerie dans les maladies neurodégénératives. La base théorique de notre travail est l'auto-encodeur variationnel (VAE), un modèle de variables latentes bien adapté pour travailler avec des données complexes car il les projette dans un espace plus simple et de faible dimension, capable de modéliser les non-linéarités des données.

Le cœur de cette thèse consiste en l'autoencodeur variationnel multicanal (MCVAE), une extension du VAE pour modéliser conjointement les relations latentes entre les observations multimodales. Ceci est réalisé 1) en contraignant la distribution latente de chaque modalité de données à une distribution a priori commune, 2) en forçant chaque distribution latente à générer toutes les modalités de données à travers leurs fonctions génératives associées.

De plus, nous adaptons le MCVAE à un contexte multi-tâches, où le problème du traitement des données manquantes est traité avec un schéma d'optimisation spécifique qui suit les étapes suivantes : 1) définition des tâches à travers les ensembles de données basée sur l'identification des sous-ensemble présentant des modalités compatibles, 2) empilement de plusieurs instances du MCVAE, où chaque instance modélise une tâche spécifique, 3) partage des paramètres communes entre les tâches de modélisation. Grâce à ces actions, le MCVAE multi-tâches permet d'apprendre un modèle conjoint pour tous les points de données en s'appuyant sur toutes les informations disponibles.

Dans son ensemble, cette thèse fournit un nouvel ensemble d'approches flexibles pour tenir compte de l'hétérogénéité des données dans l'analyse des informations biomédicales. Ce travail permet de nouvelles directions de recherche dans lesquelles l'information médicale peut être modélisée de manière cohérente dans un cadre probabiliste conjoint tenant compte des canaux de données multiples, des informations manquantes et des biais dans différents ensembles de données.

Enfin, grâce à leur formulation générale, les méthodologies proposées ici peuvent trouver des applications au-delà du domaine de la recherche en neuro-imagerie.

Mots-clés: Maladie d'Alzheimer (MA); Neuro-imagerie (NI); Imagerie par résonance magnétique (IRM); Tomographie par émission de positrons (TEP); Auto-encodeur variationnel (AEV); Apprentissage multi-tâche (AMT); Données de haute dimension; Données multimodales.

Acknowledgement

This PhD Thesis is the result of a long journey involving many people and their important direct and indirect contributions.

First of all I would like to thank Pr. Nicholas Ayache for being actively present in all the most important moments of the PhD. I want to thank him for his scientific insights and writing advices, but mostly for the constant reassurance and encouragement. Pr. Philippe Robert and Dr. Valeria Manera, for the insightful discussions about the importance of translating the research findings by taking into account the human clinical environment and, on the other way round, to enrich fundamental research with observations from clinical research. Dr. Marco Lorenzi, to whom I just can't say thank you enough for his tireless scientific support over these years, for his contagious enthusiasm, and for having taught me that hard work and attention to details eventually pay you back.

I would like to thank the jury members, in particular Pr. Gloria Menegaz and Pr. Pietro Michiardi for their voluntary contribution in reviewing this manuscript, their comments and advises. I equally thank Pr. Valentina Garibotto who accepted the role of jury member with manifested joy and interest.

I am very happy for having shared this important journey with wonderful colleagues, ending up becoming dear friends: Clément, Jaume, Sara, Santiago, few lines here cannot express adequately my gratitude, so I will just say: thank you!

A very special thank you goes to Silvia, who made me aware of what does it really mean to care for something special.

I reserve these final lines to thank my parents Antonio and Maria Concetta and my little brother Pierpaolo for their interest, support, love, and affection. I never get used to be far from them, but knowing that they are always there for me makes me feel strong and safe in any circumstance.

Thank you all, sincerely.

Luigi.

Financial Support

This work has been supported by:

- the French government, through the UCA^{JEDI} Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01;

- the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002;

- the OPAL infrastructure from Université Côte d'Azur, providing computational resources and support.

Contents

1 Introduction **1**

1.1 Context . . . 1
1.1.1 Alzheimer's Disease . . . 1
1.1.2 Biomarkers . . . 2
1.2 Integrating biomarkers . . . 3
1.2.1 Current approaches and open challenges . . . 4
1.3 Beyond dementia studies . . . 6
1.4 Objectives and organization of this Thesis . . . 7
1.5 Publications . . . 9

2 Multi-Channel Variational Autoencoder **11**

2.1 Introduction . . . 12
2.2 Method . . . 14
2.2.1 Multi-Channel Variational Autoencoder (MCVAE) . . . 14
2.2.2 Inducing Sparse Latent Representations . . . 18
2.3 Experiments . . . 20
2.3.1 Synthetic Experiments . . . 21
2.3.2 Sparse Multi-Channel VAE Benchmark . . . 21
2.3.3 Comparison with VAE . . . 23
2.3.4 Medical Imaging data . . . 24
2.4 Conclusion . . . 26
2.5 Supplementary Material . . . 29

3 Multi-Task Multi-Channel Variational Autoencoder **31**

3.1 Introduction . . . 32
3.2 Method . . . 35
3.2.1 Generative Model . . . 36
3.2.2 Inference Model . . . 36
3.2.3 Optimization . . . 37
3.2.4 Comparison with VAE and MCVAE . . . 38
3.2.5 Parameterization . . . 40
3.3 Experiments . . . 41
3.3.1 Illustration on a simplified brain imaging dataset . . . 41
3.3.2 Synthetic Experiments . . . 43

3.3.3 Experiments on Brain Imaging Data . . . 46
3.4 Discussion . . . 58
3.5 Conclusions . . . 60
3.6 Supplementary Material . . . 60
3.6.1 Derivation of the Lower Bound . . . 61
3.6.2 Data Generation . . . 62

4 Benchmark of Harmonization Methods 63
4.1 Introduction . . . 64
4.2 Methods . . . 66
4.2.1 ComBat . . . 66
4.2.2 CovBat . . . 67
4.2.3 Domain Invariant Variational Autoencoder . . . 67
4.2.4 Qualitative benchmark . . . 68
4.3 Experiments . . . 71
4.3.1 Synthetic data generation procedure . . . 71
4.3.2 Quantitative benchmark . . . 73
4.3.3 Results . . . 73
4.4 Discussion . . . 74
4.5 Conclusion . . . 77

5 `mcvae`: an Open Source Python Toolbox 79
5.1 Introduction . . . 79
5.2 Implementation overview . . . 80
5.3 Usage: example of multi-modal learning . . . 82
5.4 Supplementary documentation . . . 86
5.4.1 Main model classes (`mcvae/models`) . . . 86
5.4.2 Utilities . . . 87

6 Conclusion 91
6.1 Summary of the main contributions . . . 91
6.2 Future developments . . . 93
6.2.1 Disjoint representations . . . 93
6.2.2 Domain shift compensation . . . 94
6.2.3 Temporal modeling . . . 96
6.2.4 Other applications . . . 96
6.3 Final remarks . . . 97

Acronyms 99

Glossary 103

Bibliography 105

1 Introduction

Contents

1.1 Context **1**
1.1.1 Alzheimer's Disease 1
1.1.2 Biomarkers 2
1.2 Integrating biomarkers **3**
1.2.1 Current approaches and open challenges 4
1.3 Beyond dementia studies **6**
1.4 Objectives and organization of this Thesis **7**
1.5 Publications **9**

1.1 Context

Dementia is an umbrella term for several diseases affecting memory, other cognitive abilities and behavior that interfere significantly with a person's ability to maintain their activities of daily living. Although age is the strongest known risk factor for dementia, it is not a normal part of aging. [1] Alzheimer's Disease (AD) is the most common cause of dementia and accounts for 60% to 80% of the cases [Alzheimer Association Report, 2020].

1.1.1 Alzheimer's Disease

Alzheimer's Disease was firstly reported in 1906 by a clinical psychiatrist and neuroanatomist named Alois Alzheimer, who described a 50 year-old woman whom he had followed from her admission for paranoia, progressive sleep and memory disturbance, aggression, and confusion, until her death 5 years later. His report noted distinctive protein plaques and neurofibrillary tangles in the brain *post-mortem* histology [Hippius, 2003],

Nowadays, a definitive diagnosis of AD can only be established through postmortem brain tissue biopsies, aiming at finding plaques of amyloid proteins in the extracellular space and tangles of tau proteins in the intracellular space. However, clinical interest currently focuses on identifying persons with cognitive impairment who will probably

[1] `https://www.who.int/health-topics/dementia`

progress to dementia, rather than on identifying the exact underlying pathology, and on the research for pharmacological and non-pharmacological interventions to slow down the degenerative process. Indeed, as researchers are developing an increasing awareness of the complexity of AD and related disorders, the clinical evaluation of progression to dementia through the integration of clinical, imaging, and biological biomarkers is considered as a key step towards the accurate definition of the pathology [Boccardi, 2021].

One of the major challenges for understanding AD is that the pathology evolves unnoticed for a long period (up to 20 years) before the manifestation of clinically recognizable cognitive [Frisoni, 2003; Solomon, 2011] and behavioral symptoms [Scarmeas, 2007; Fostinelli, 2020]. Clinicians refer to this stage as to the *pre-clinical* phase of AD. Therefore, efforts have focused on finding a set of biomarkers that would allow an early detection and follow-up monitoring of the AD hallmarks along the disease progression. In 2011, the National Institute on Aging and Alzheimer's Association (NIA-AA) created separate diagnostic recommendations for the preclinical, mild cognitive impairment, and dementia stages of AD [McKhann, 2011]. Scientific progress in the interim led to an initiative by the NIA-AA to update and unify the 2011 guidelines. These efforts resulted in the definition of the A/T/N Research Framework, in which the acronym comes for the three main biomarkers categories involved in AD, namely: amyloid, tau and neurodegeneration [Jack, 2018].

1.1.2 Biomarkers

From what we reported in the previous section, it follows that Alzheimer's Disease can be tracked via biomarkers, accordingly with the A/T/N categories in the Research Framework, which indicate the abnormality of specific physiological processes. For example, measurements of concentration of Aβ_{42} and tau proteins in the cerebrospinal fluid (CSF) allow to detect respectively levels of amyloid-beta and tau. The A/T/N criteria define a clear role for tau and amyloid biomarkers in the diagnostic procedure of patients complaining about cognition. In particular, tau-positiveness is necessary but not sufficient to define clinical AD, and tau-positiveness associated to amyloid-negativity denotes the presence of a neurodegenerative disorder belonging to a non-AD continuum. Imaging techniques, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), are also suited to measure many pathophysiological changes involved in AD. For instance, abnormal deposition of amyloid proteins can be also measured via PET with the ^{18}F-Florbetapir (AV45) radioactive tracer [Clark, 2011]. Accumulation of neurofibrillary tangles is quantified via PET with the ^{18}F-Flortaucipir (AV1451) tracer [Barthel, 2020]. Finally, neurodegeneration is indicated by cerebral atrophy from MRI scans [Fox, 2004] and glucose hypo-metabolism from ^{18}F-Fluorodeoxyglucose-PET [Herholz, 2012; Garibotto, 2017].

1.2 Integrating biomarkers

Although most of the existing researches in dementia focus on only a single modality of biomarkers, there is a general agreement that combining biomarkers improves diagnostic accuracy [Chételat, 2021]. Fostered by many research initiatives aiming at collecting and sharing data for a better understanding of Alzheimer's Disease and other forms of dementia, the literature on this subject is growing. For example, in [Shaffer, 2013] the authors, after studying subjects with Mild Cognitive Impairment, found that combining MRI, FDG-PET, and CSF data with routine clinical tests significantly increased the accuracy of predicting conversion to AD compared with clinical testing alone. Specifically, MRI-derived gray matter probability maps and FDG-PET images were analyzed by using Independent Component Analysis (ICA), a data-driven method to extract independent sources of information from whole-brain data. The ICA loading parameters for all MRI and FDG components, along with CSF proteins, were entered into logistic regression models. A variety of models were considered, including all combinations of MRI, PET, and CSF biomarkers with the age, education years, Apolipoprotein-E (ApoE), Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog) as covariates. Similar results were confirmed in [Gupta, 2019], where authors proposed a machine learning-based framework to discriminate subjects with AD or MCI utilizing a combination of four different biomarkers: FDG-PET, MRI, CSF protein levels, and ApoE genotype. Here, a kernel-based multi-class support vector machine (SVM) classifier with a grid-search method was applied to optimally select features from the input biomarkers.

Over the last 20 years, governments, universities, charities and pharmaceutical companies have devoted increasingly significant resources, in terms of funding, time, and effort, to foster knowledge advancements. For example, neuroimaging studies in dementia, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [Weiner, 2013], the Open Access Series of Imaging Studies (OASIS) [LaMontagne, 2019], the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) [Malone, 2013], have produced huge amounts of heterogeneous, multi-modal, and high-dimensional data, including those coming from MRI and PET Imaging. All these data were collected from subjects with different cognitive conditions, with the aim to find a set of biomarkers that would allow to detect and monitor patho-physiological stages along the disease path.

Different biomarkers may be combined to provide better insights [Apostolova, 2010]. Hence, the joint analysis of biomedical data in Dementia studies is important for better clinical diagnosis and to understand the relationship between biomarkers. However, jointly accounting for heterogeneous measures poses important challenges related to the modeling of heterogeneity and to the interpretability of the results. Moreover, when pooling together observations from different studies in order to take advantage of the increased variability and sample size, the joint analysis requires to consistently analyze

high-dimensional and heterogeneous information in presence of often non-overlapping acquisition and data processing protocols, with missing data across data samples. The next section addresses these analysis challenges by describing current approaches to multi-modal data modeling.

1.2.1 Current approaches and open challenges

As described in the previous section, as tackling a complex problem like AD requires to establish links between high-dimensional heterogeneous data sources, a variety of approaches have been proposed in the literature. As this subject will be widely discussed in the introductory sections of Chapters 2 and 3, here we describe the general aspects of current approaches to link data modalities.

Multivariate methods

Probably the simplest approach is the massive univariate correlation analysis [Nathoo, 2019]. Unfortunately this approach is too limited in modeling power, and prone to false positives when the data dimension is high. To overcome the limitations of mass-univariate analysis, more advanced methods, such as Canonical Correlation Analysis (CCA), Independent Component Analysis (ICA), Partial Least Squares (PLS), and Reduced Rank Regression (RRR), have successfully been applied in biomedical research (see [Liu, 2014] for a comprehensive review), along with multi-sources [Kettenring, 1971; Luo, 2015] and non-linear variants [Huang, 2009; Andrew, 2013a]. The common formulation of these approaches consists in projecting the observations in a latent low dimensional space where desired statistical properties are enforced, such as maximum correlation (CCA), maximum covariance (PLS), or minimum regression error (RRR). However, since they are not generative, these methods are limited in providing information on how this latent representation is expressed in the observations [Haufe, 2014]. Generative modeling attempts have been made, such as with the Bayesian-CCA [Klami, 2013], where a transformation of a latent variable captures the shared variation between two data sources. Unfortunately, due to scalability issues in the computation of posterior distributions, all the practical applications are limited to model data with very few dimensions. Variational Inference (VI) is a popular approach to compute posterior distributions when the usual integrations are intractable. Variational Inference (VI) has been successfully applied in the recent seminal work on the Variational Autoencoder (VAE) [Kingma, 2014b; Rezende, 2014], a powerful generative model for high-dimensional single-modality observations. The work developed in this Thesis is largely inspired by the VAE.

The Variational Autoencoder

The VAE is composed by two main elements: the encoder and the decoder. The encoder can be seen as a Bayesian version of the Principal Component Analysis (PCA) [Rolinek, 2019], that is it transforms input data, usually high dimensional, to produce a compressed and informative version of the data distribution: for each high-dimensional observation we can associate an expected value and a probability interval into a lower dimensional latent space that captures the main variability in the original data. The decoder works in the opposite sense of the encoder: it is a generator function that, given a compressed low dimensional representation of a data point, produces a distribution maximizing the original data likelihood. The training of a VAE aims at finding the optimal encoder and decoder pair associated with the maximum of information of the original high-dimensional data when encoding, with the minimum reconstruction error when decoding. To prevent overfitting, the encoding distributions are regularized during training, to ensure that the encoded latent space preserves smoothness properties allowing us to generate new data. The regularization is achieved by minimizing an information theoretic measure, the Kullback-Leibler divergence function, between the encoding distributions and the associate prior, usually chosen to be the factorized isotropic multivariate Gaussian. This is the variational inference of statistics. The term *variational* comes from the calculus of variation that originally was used to estimate distributions instead of variables.

Since its first introduction in 2014, research involving the VAE increased steadily, and many research groups are currently involved in improving the performances and capabilities of the VAE. For example, in [Rossi, 2019], the authors show that, given the usually high number of parameters required to fit a VAE based model, initialization plays a huge role in their convergence to a good solution, and propose a method to prevent the problem of *posterior collapse* [Lucas, 2019], that is to avoid the trivial solutions of encoding distributions being equal to the prior. Another area of research aims at obtaining more informative encoding distributions. Indeed, the factorized Gaussian usually adopted as regularizer is not enough to guarantee the factorization of the associated encoding distributions. This constitutes a problem as a desirable property of the latent space is to have each dimension disentangled from the others. This is still a very active area of research, with solutions proposed in the context of supervised learning [Lopez-Martin, 2017], unsupervised learning [Higgins, 2018], and semi-supervised learning [Mita, 2020].

With the VAE is possible to model multi-modal data by stacking all the modalities into a single one. This represents a limit, as modeling stacked data through a VAE may pose interpretability issues. Indeed, it would be generally difficult to disentangle the contribution of a single modality in the description of the latent representation, especially with non-linear encoder and decoder architectures.

These limitations are crucial when applying VAE, and more general machine learning models, to clinical data. To address this issue, in this Thesis we focused on the extension of VAE approaches to model multi-modal data in a more interpretable manner, by introducing independent encoders and decoders which are jointly linked in the latent space in an information theoretical sense.

Data missingness

Data scarcity is a critical issue when modeling observations in dementia studies, as some data modalities are costly to obtain and not always available. This is the case, for example, of PET images with radioactive tracers for Aβ_{42} and Tau proteins, which are known to play an important role in AD [Jack, 2018]. As fitting multi-modal models requires to establish correspondences between modalities, subjects with at least one missing modality are generally discarded, yielding to potentially severe loss of available information. An appropriate approach to increase the sample size and take advantage of all the available data, is to gather observations from different studies, although this approach does not solve the problem of missing data. Indeed, according to the cohort study design, there may be views which are specifically absent for a given dataset (*i.e.,* missing not at random). This potential mismatch across datasets hampers their interoperability, and prevents the gathering of all the available observations into a single, robust and generalizable joint model accounting for the global data variability. This challenge is typically addressed in machine learning in the field of Multi-Task Learning (MTL), where each dataset is associated to a specific modeling task. MTL is usually achieved with specific output layers for every task, and by including a shared latent representation for all of them [DoradoMoreno, 2020]. It has been successfully applied in classification [Choi, 2019; Zhou, 2019a] and in feature prediction problems [Gondara, 2018; Wei, 2020].

In this Thesis we develop a generative and probabilistic statistical learning model for the joint analysis of high-dimensional heterogeneous biomedical data, to simultaneously learn from multiple datasets, even in the presence of non fully compatible datasets, and missing data.

1.3 Beyond dementia studies

The problem of coherently modeling heterogeneous data sources is common to many application fields, well beyond the study of Alzheimer's Disease and neurological disorders.

In oncology studies, for example, the identification of cancer subtypes [Khan, 2020] plays an important role in revealing useful insights into disease pathogenesis [Li, 2020] and advancing personalized therapy [Valieris, 2020]. Although approaches have been proposed to identify cancer subtypes from multiple genomic data sources [Zhang, 2012], very few of them are particularly designed to exploit cross-modality correlations. In [Méndez, 2015], authors proposed a multi-view consensus clustering methodology for the integration of multimodal MRI images into a unified segmentation framework, aiming at heterogeneity assessment in tumoral lesions. In that work, the modalities adopted for tissue characterization are Dynamic Contrast Enhanced MRI (DCE-MRI), that uses serial acquisition of images during and after the intravenous injection of a contrast agent to assess organ perfusion, and Diffusion Tensor Imaging (DTI), sensitive to the tissue microstructure. The task is particularly challenging as the DCE-MRI is a 4-dimensional acquisition modality of space and time, while the result of a Diffusion Tensor Imaging (DTI) acquisition consists is a second-order tensor for every voxel. Given this important level of data complexity, the advent on novel methods for the joint modeling of high-dimensional multi-modal observations are likely to produce further advancements in the oncology research field.

Another example encompasses the new Information and Communication Technologies (ICT), as they are starting to have a role in the monitoring and behavioral assessment of frail people [König, 2015; Manera, 2020]. Serious Games (SG), for example, are digital applications specially adapted for purposes other than entertaining; such as rehabilitation, training and education [Robert, 2014]. As they are likely to produce new forms of data that could be integrated with the classic instruments to better assess the disease severity and progression, multi-modal methods for the joint modeling of heterogeneous data, such as the ones developed in this Thesis, could play an important role in this field, too.

1.4 Objectives and organization of this Thesis

In this Thesis we develop a general framework to solve the problem of jointly modeling heterogeneous data in the presence of missing observations for both prediction and classification tasks.

We benchmarked our framework in synthetically generated scenario to assess its general properties, and on real data coming from neuroimaging research studies in Dementia.

Throughout this work we will adopt the words *channels*, *views*, and *modalities* interchangeably, to refer to an homogeneous set of quantitative measurements.

The manuscript is organized as follows.

We present in **Chapter 2** the Multi-Channel Variational Autoencoder (MCVAE), an extension of the VAE to jointly model latent relationships across multiple channels, consisting in groups of heterogeneous observations generated from a single source of information. In the latent space, this is achieved by constraining the variational distribution of each channel to a common target prior. Moreover, we show how sparse and parsimonious latent representations can be 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.

We introduce in **Chapter 3** the Multi-Task Multi-Channel Variational Autoencoder (MT-MCVAE), an extension of the MCVAE to modeling multi-task (that is multi-dataset) multi-channel observations, where the non-trivial problem of dealing with missing data arises. This problem has been addressed in our current work via a specific optimization scheme requiring an extension of our previous formulation to account for dataset- and channel-specific observations. Simulations on synthetic data show that our method is able to identify a common latent representation of multi-channel datasets, even when the compatibility across datasets is minimal. When jointly analyzing multi-modal neuroimaging and clinical data from real independent dementia studies, the MT-MCVAE is able to mitigate the absence of modalities without having to discard any available information. Moreover, by slightly changing the architecture of the MT-MCVAE, the inferred latent representation can be used to define robust classifiers gathering the combined information across different datasets.

Chapter 4 and **Chapter 5** are of different nature with respect to the previous ones. Here, we do not propose new methodologies: we introduce, instead, important work of practical utility for the methods developed in the earlier chapters.

Specifically, in **Chapter 4** we benchmark existing harmonization methodologies for correcting the bias induced by the data domain. Indeed, when integrating data across different studies and datasets to increase the sample size, such as with the MT-MCVAE, the bias induced by the *domain shift*, that is the existence of different protocols between studies, multiple imaging machine manufacturers, image reconstruction software, and preprocessing algorithms, creates barriers to the integration of multi-centric datasets.

In **Chapter 5** we present the open-source Python package `mcvae`, where we publicly released the source code of the methods presented in Ch. 2 and Ch. 3 of this thesis along with the necessary documentation, to foster and promote research in joint modeling of heterogeneous data in other domains.

Finally, we conclude the manuscript with **Chapter 6** by summarizing the main contributions of this work. We also present potential extensions built upon the acknowledged limitations to propose future research perspectives.

1.5 Publications

The contributions of this manuscript led to the following publications in conferences and peer-reviewed journals.

- *Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data.* Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. In ICML 2019 - 36th International Conference on Machine Learning, Long Beach, United States, June 2019. Proceedings of Machine Learning Research, (97):302-311, 2019. [Antelmi, 2019]

- *Combining Multi-Task Learning and Multi-Channel Variational Auto-Encoders to Exploit Datasets with Missing Observations - Application to Multi-Modal Neuroimaging Studies in Dementia.* Luigi Antelmi, Nicholas Ayache, Philippe Robert, Federica Ribaldi, Valentina Garibotto, Giovanni B Frisoni, and Marco Lorenzi. Under review at NeuroImage 2021. [Antelmi, 2021]

- *Multi-Chanel Stochastic Variational Inference for the Joint Analysis of Heterogeneous Biomedical Data in Alzheimer's Disease*. Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. In Understanding and Interpreting Machine Learning in Medical Image Computing Applications, Granada, Spain, September 2018. Lecture Notes in Computer Science, 11038:15-23, 2018. [Antelmi, 2018a]

- *A method for statistical learning in large databases of heterogeneous imaging, cognitive and behavioral data*. Luigi Antelmi, Marco Lorenzi, Valeria Manera, Philippe Robert, and Nicholas Ayache. EPICLIN 2018, Poster session. [Antelmi, 2018b]

2 Multi-Channel Variational Autoencoder

Contents

2.1 Introduction **12**
2.2 Method **14**
2.2.1 Multi-Channel Variational Autoencoder (MCVAE) 14
2.2.2 Inducing Sparse Latent Representations 18
2.3 Experiments **20**
2.3.1 Synthetic Experiments 21
2.3.2 Sparse Multi-Channel VAE Benchmark 21
2.3.3 Comparison with VAE 23
2.3.4 Medical Imaging data 24
2.4 Conclusion **26**
2.5 Supplementary Material **29**

In this chapter we present the Multi-Channel Variational Autoencoder (MCVAE), a latent variable framework to jointly model complex heterogeneous observations. Here we call *channel* a group of homogeneous observations. We argue that our framework can be of interest for the neuroimaging community as it can be adopted to model the joint relationship between multi-modal neuroimaging data, such as those coming from Alzheimer's Disease Neuroimaging Initiative (ADNI). Indeed, in this context of high heterogeneity due to the presence of, among many others, Magnetic Resonance Imaging (MRI) data and Positron Emission Tomography (PET) imaging data, that is channels with their own informative content, there is a rational need for methods to establish relationships between observations. To do so, we postulate a single source of information for all the channels, and we use Variational Inference (VI) to infer this single source from them. This is achieved in the latent space by constraining the variational distribution of each channel to a common target prior. Moreover, we show how sparse and parsimonious latent representations can be enforced by variational dropout. This chapter is published in the *Proceedings of Machine Learning Research* [Antelmi, 2019] and is based on a previous works presented at the fisrt *Workshop on Machine Learning in Clinical Neuroimaging* [Antelmi, 2018a], and at the 12^{th} *EPICLIN Conference* [Antelmi, 2018b].

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.

2.1 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, 2014], along with multi-channel [Kettenring, 1971; Luo, 2015] and non-linear variants [Huang, 2009; Andrew, 2013a]. 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, 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, 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, 2007].

Variational Autoencoder (VAE) [Kingma, 2014b; Rezende, 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 *amortized inference* [Kim, 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, 2015; Alemi, 2017]. As discussed in [Yeung, 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, 2015; Molchanov,

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, 2017] is publicly available at `https://gitlab.inria.fr/epione_ML/mcvae`.

The rest of this chapter is organized as follows. In § (2.2) 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 § (2.3) 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.

2.2 Method

We first describe the proposed Multi-Channel Variational Autoencoder (§2.2.1). In §2.2.2 we present the sparse formulation of our method.

2.2.1 Multi-Channel Variational Autoencoder (MCVAE)

Let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_C\}$ 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:

$$\begin{aligned} &\mathbf{z} \sim p(\mathbf{z}), \\ &\mathbf{x}_c \sim p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c), \qquad \text{for } c \text{ in } 1 \ldots C, \end{aligned} \tag{2.1}$$

where $p(\mathbf{z})$ is a prior distribution for the latent variable and $p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c)$ 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 $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\}$.

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 $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, that is not always computable analytically. In this case, Variational Inference (VI) can be applied to compute an approximate posterior [Blei, 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 $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$, 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 $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ to be as close as possible to the target posterior distribution. Being the mismatch measured in terms of Kullback-Leibler ($\mathcal{D}_{\mathrm{KL}}$) divergence, we specify this constraint as:

$$\underset{q \in \mathcal{Q}}{arg\,min}\ \mathbb{E}_c \left[\mathcal{D}_{\mathrm{KL}} \left(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||p(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_C, \boldsymbol{\theta}) \right) \right], \tag{2.2}$$

where the approximate posteriors $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ belong to a distribution family $\mathcal{Q}$ parametrized by the set of parameters $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_C\}$, 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. (2.2) allows to minimize the discrepancy between the variational approximations and the target posterior. In §2.2.1 we show that the optimization (2.2) is equivalent to the optimization of the following evidence lower bound $\mathcal{L}(\mathcal{D})$:

$$\mathcal{L}(\mathcal{D}) = \mathbb{E}_c \left[L_c - \mathcal{D}_{\mathrm{KL}} \left(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||p(\mathbf{z}) \right) \right], \tag{2.3}$$

where $L_c = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)} \sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\theta}_i)$ 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 $\mathcal{Q}$.

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 (2.2) states that variational inference is carried out by introducing a set of probability density functions $q(\mathbf{z}|\mathbf{x}_c)$, belonging to a distribution family $\mathcal{Q}$, that are as close as possible to the target posterior over the latent variable $p(\mathbf{z}|\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_C\})$. Given the intractability of $p(\mathbf{z}|\mathbf{x})$ for most complex models, we cannot solve directly

this optimization problem. We look then for an equivalent problem, by rearranging the objective:

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

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

$$
\begin{aligned}
&\ln p\left(\mathbf{x}\right) - \underbrace{\mathbb{E}_c \left[\mathcal{D}_{\mathrm{KL}} \left(q\left(\mathbf{z}|\mathbf{x}_c\right) || p\left(\mathbf{z}|\mathbf{x}\right) \right) \right]}_{\geq 0} = \\
&= \underbrace{\mathbb{E}_c \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)} \left[\ln p\left(\mathbf{x}|\mathbf{z}\right) \right] - \mathcal{D}_{\mathrm{KL}} \left(q\left(\mathbf{z}|\mathbf{x}_c\right) || p\left(\mathbf{z}\right) \right) \right]}_{\text{lower bound } \mathcal{L}}.
\end{aligned}
\tag{2.4}
$$

Since the $\mathcal{D}_{\mathrm{KL}}$ 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 (2.2).

We note that the lower bound (2.4) is composed by a regularization term and a data matching term. The $\mathcal{D}_{\mathrm{KL}}$ 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:

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

Comparison with VAE

Our model extends the VAE: the novelty is in the log-likelihood terms L_c in Eq. (2.3), 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:

$$\hat{\mathbf{x}}_i = \mathbb{E}_j \left[\mathbb{E}_{q(\mathbf{z}|\tilde{\mathbf{x}}_j)} \left[p\left(\mathbf{x}_i|\mathbf{z}\right) \right] \right]. \tag{2.5}$$

An application of Eq. (2.5) is provided in §2.3.4. 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, 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)).

Gaussian linear case

Model (2.1) 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, 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:

$$q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) = \mathcal{N}\left(\mathbf{z}|\mathbf{V}_c^{(\mu)}\mathbf{x}_c, diag(\mathbf{V}_c^{(\sigma)}\mathbf{x}_c)\right), \tag{2.6}$$

$$p\left(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c\right) = \mathcal{N}\left(\mathbf{x}_c|\mathbf{G}_c^{(\mu)}\mathbf{z}, diag(\mathbf{g}_c^{(\sigma)})\right), \tag{2.7}$$

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 (2.3).

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. (2.3) 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 $\mathcal{D}_{\mathrm{KL}}$ term in Eq. (2.3) can be computed analytically (*cf.* [Kingma, 2014b], appendix 2.A). The maximization of $\mathcal{L}\left(\mathcal{D}\right)$ 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, 2014a] we compute adaptive learning rates for the parameters.

2.2.2 Inducing Sparse Latent Representations

In extensive synthetic experiments with the non-sparse version of the multi-channel model, we found that the lower bound (2.3) 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 (2.1), as the number of relationships to take into account decreases.

Regularization via Dropout

Dropout [Srivastava, 2014] and *DropConnect* [Wan, 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:

$$x_i = \sum_k g_{ik}(\xi_k z_k), \qquad \text{(dropout)} \tag{2.8}$$

$$x_i = \sum_k (\xi_{ik} g_{ik}) z_k, \qquad \text{(dropconnect)} \tag{2.9}$$

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, 2013], and their distributions takes the form:

$$x_i \sim \mathcal{N}\left(\textstyle\sum_k \theta_{ik}; \alpha \sum_k \theta_{ik}^2\right), \tag{2.10}$$

where $\alpha = {}^{p}/_{1-p}$ and $\theta_{ik} = g_{ik} z_k (1-p)$. In *Gaussian dropout* [Wang, 2013] the regularization is achieved by sampling directly from (2.10).

Variational Dropout and Sparsity

In the context of the Variational Autoencoder (VAE), posterior distributions on the encoder weights w that take the form $w \sim \mathcal{N}(\mu; \alpha\mu^2)$ are called *dropout posteriors* [Kingma, 2015]. The authors of [Kingma, 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:

$$p(\ln|w|) = \text{const} \Leftrightarrow p(|w|) \propto \frac{1}{|w|}. \tag{2.11}$$

With this prior, the $\mathcal{D}_{\mathrm{KL}}$ of the dropout posterior depends only on α and can be numerically approximated. In [Molchanov, 2017] the authors provide an approximation of $\mathcal{D}_{\mathrm{KL}}$, reported in (2.12), to allow this parameter to be learned through the optimization of the lower bound via gradient-based methods:

$$\begin{gathered} \mathcal{D}_{\mathrm{KL}}\left(\mathcal{N}\left(w; \alpha w^2\right) || p(w)\right) \approx \\ \approx -k_1\sigma(k_2 + k_3 \ln\alpha) + 0.5\ln(1+\alpha^{-1}) + k_1 \\ k_1 = 0.63576 \quad k_2 = 1.87320 \quad k_3 = 1.48695 \\ \sigma(\cdot) \text{ Sigmoid function.} \end{gathered} \tag{2.12}$$

While the optimization of $\mathcal{D}_{\mathrm{KL}}$ 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.

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 §2.2.2. 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:

$$x_i \sim \mathcal{N}\left(\sum_k g_{ik}\mu_k; \alpha \sum_k g_{ik}^2\mu_k^2\right), \tag{2.13}$$

in which the first two moments are as follows:

$$\begin{aligned}
\mathbb{E}\left[x_i\right] &= \mathbb{E}\left[\textstyle\sum_k g_{ik}z_k\right] = \textstyle\sum_k g_{ik}\mu_k, && (2.14)\\
\mathrm{Var}\left[x_i\right] &= \mathrm{Var}\left[\textstyle\sum_k g_{ik}z_k\right] \\
&= \textstyle\sum_k \mathrm{Var}\left[g_{ik}z_k\right] + \sum_{k,\, j\neq k} \mathrm{Cov}\left[(g_{ik}z_k, g_{ij}z_j)\right] \\
&= \textstyle\sum_k g_{ik}^2\alpha\mu_k^2 = \alpha\sum_k g_{ik}^2\mu_k^2, && (2.15)
\end{aligned}$$

with the covariance terms vanishing for the hypothesis of independent elements of $\mathbf{z}$. The analogy with (2.10) 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. (2.3) and parametrize them to be factorized dropout posteriors, that is, for c in $1 \dots C$:

$$q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) = \mathcal{N}\left(\boldsymbol{\mu}_c; \mathrm{diag}(\sqrt{\boldsymbol{\alpha}} \odot \boldsymbol{\mu}_c)^2\right), \tag{2.16}$$

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

$$p\left(\mathbf{z}\right) = \prod p\left(|z_i|\right), \quad \text{such that} \quad p\left(\ln|z_i|\right) \propto \text{const.} \tag{2.17}$$

Because of these choices, the $\mathcal{D}_{\mathrm{KL}}$ term in Eq. (2.3) can be easily computed by leveraging on Eq. (2.12). 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. 2.1). 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.

2.3 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.

2.3.1 Synthetic Experiments

Table 2.1: Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (2.18).

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

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

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}_l), \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}_{d_c}), \\ \mathbf{G}_c &= diag\left(\mathbf{R}_c \mathbf{R}_c^T\right)^{-1/2} \mathbf{R}_c, \\ \mathbf{x}_c &= \mathbf{G}_c \mathbf{z} + \mathrm{SNR}^{-1/2} \cdot \boldsymbol{\epsilon}, \end{aligned} \tag{2.18}$$

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.,* $diag\left(\mathbb{E}\left[\mathbf{x}_c \mathbf{x}_c^T\right]\right) = (1+\mathrm{SNR}^{-1})\mathbf{I}_{d_c}$. Scenarios where generated by varying one-at-a-time the dataset attributes, as listed in Tab. 3.12.

ELBO in non-sparse Multi-Channel VAE. For each generated scenario, we optimized multiple instances of a Gaussian Linear Multi-Channel model, as defined in §2.2.1. 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.

2.3.2 Sparse Multi-Channel VAE Benchmark

This benchmark is based on the data scenarios illustrated in the previous section (Tab. 3.12). For each generated dataset, we optimized our Multi-Channel VAE with dropout posteriors (eq. 2.16) associated to log-uniform priors as in (eq. 2.17).

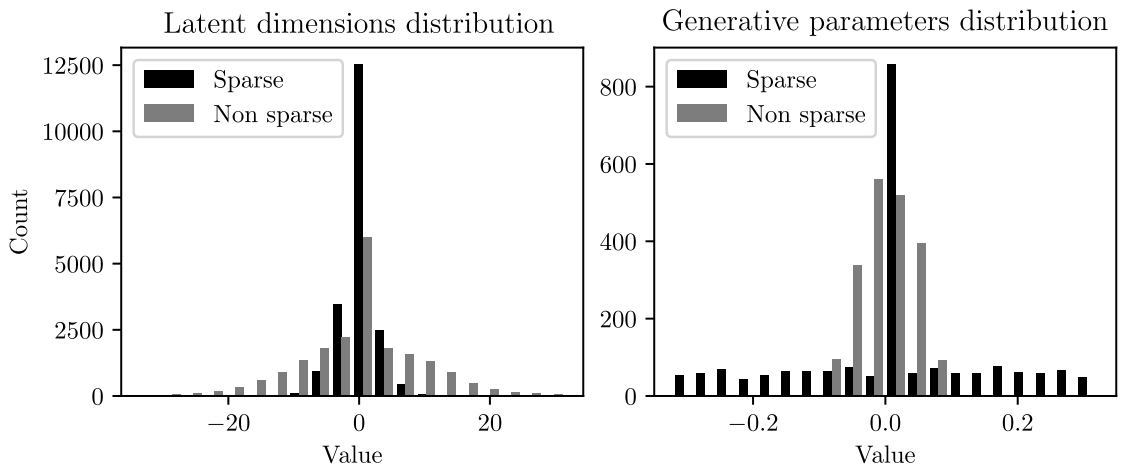


Figure 2.1: Effect of variational dropout on a synthetic experiment modeled with the Multi-Channel VAE. As expected, the minimum amount of non-zero components of $\mathbf{z}$ (left) and generative parameters $\mathbf{G}$ (right) is obtained with the sparse model.

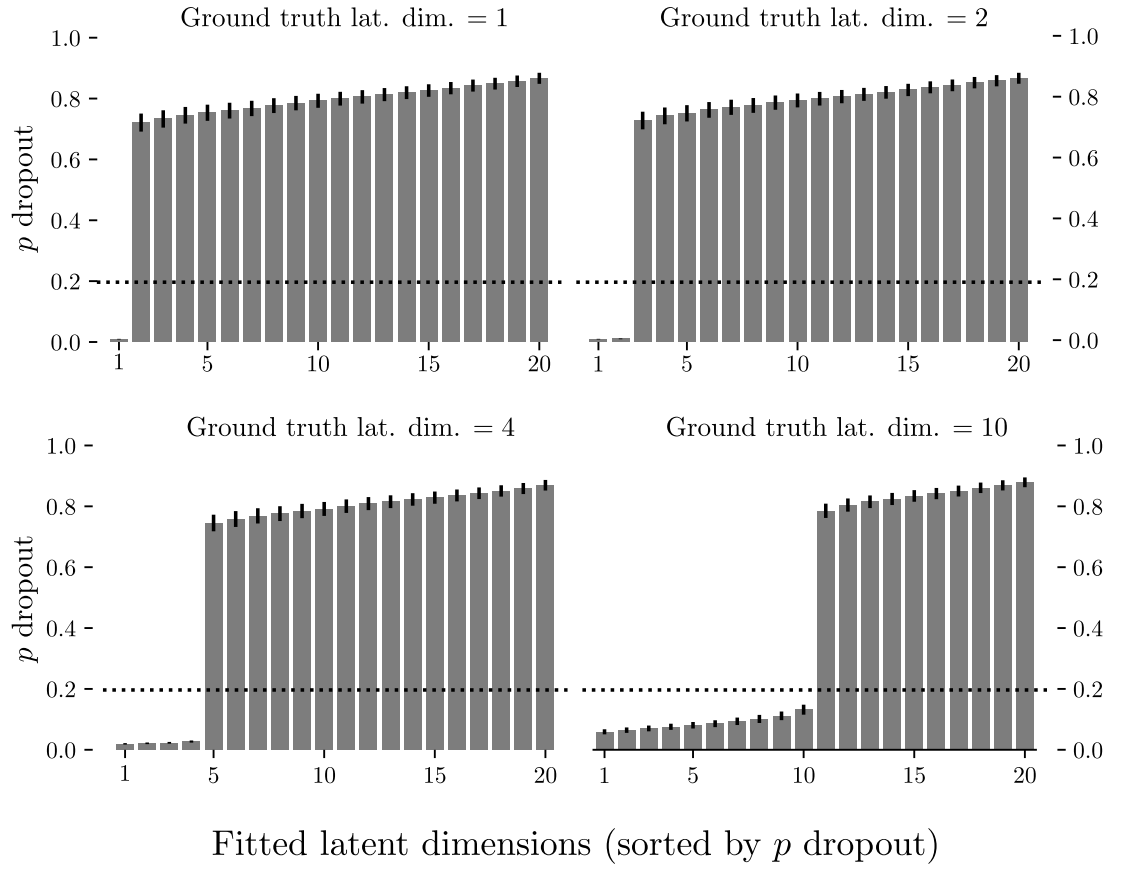


Figure 2.2: Estimated dropout rates for the latent dimensions when the initial latent dimensions of the Sparse Multi-Channel VAE was set to $l_{\text{fit}} = 20$ on data generated with respectively $l = 1, 2, 4,$ and 10 latent dimensions.

Results. In Fig. 2.1 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. 2.2). 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.

2.3.3 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 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.

Table 2.2: 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%	$\mathbf{45}\%$	0%

Results. As depicted in Tab. 2.2, 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. 2.3 we restrict the visualization to the cases where snr $= 10$ and $l_{\text{fit}} = l$ (*cf.* Tab. 3.12). 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. 2.3), 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.

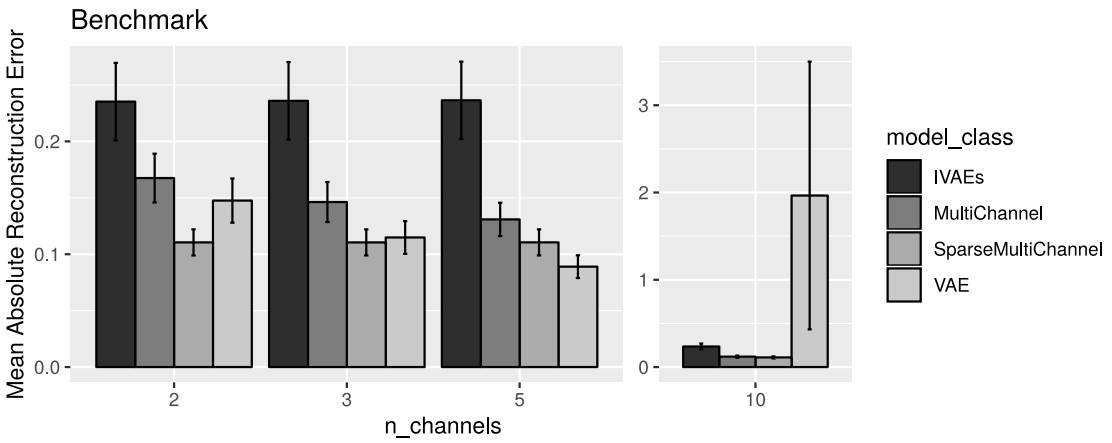


Figure 2.3: Testing benchmark of four variational methods applied to the multi-channel scenarios in Tab. 3.12 (cases snr $= 10$, $l_{\text{fit}} = l$). Sparse Multi-Channel models performs consistently better than non-sparse Multi-Channel ones.

2.3.4 Medical Imaging data

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 the Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog), Clinical Dementia Rating Scale (CDR), Functional Activities Questionnaire (FAQ); scholarity level. The three imaging channels were structural Magnetic Resonance Imaging (MRI) (gray matter only), functional Fluorodeoxyglucose Positron Emission Tomography (FDG-PET), and Amyloid-PET. Raw data from the imaging channels were coregistered in a common geometric space by means of Voxel-based Morphometry (VBM) methods [Ashburner, 2000]. Visual quality check was performed to exclude processing errors. Image intensities were finally averaged over 90 brain regions mapped in the Automated Anatomical Labeling (AAL) atlas [TzourioMazoyer, 2002] to produce 90 features arrays for each image. Lastly, data was centered and standardized across features. Our sparse multi-channel model (§2.2.2) was optimized on the resulting multi-channel dataset, along with MCVAE, IVAEs, and VAE models as described in §2.3.3. 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.

Table 2.3: 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

Results. By applying the dropout threshold of 0.2 as identified in the synthetic experiments (Fig. 2.2), 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. 2.4, where we limited the visualization to the 2D subspace generated by the two most relevant dimensions. This subspace appears stratified by age and disease status, across roughly orthogonal directions. This disentanglement between aging and disease is confirmed also with other modeling approaches [Lorenzi, 2015; Sivera, 2019]. We note however that our 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. 2.3, 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 (LDA) 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. (2.5), 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. 2.4 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. 2.5. Fig. 2.6 shows the generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (2.7)) 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-Cog and CDR), gray matter atrophy emerging from the structural 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, 2014; Jack, 2018].

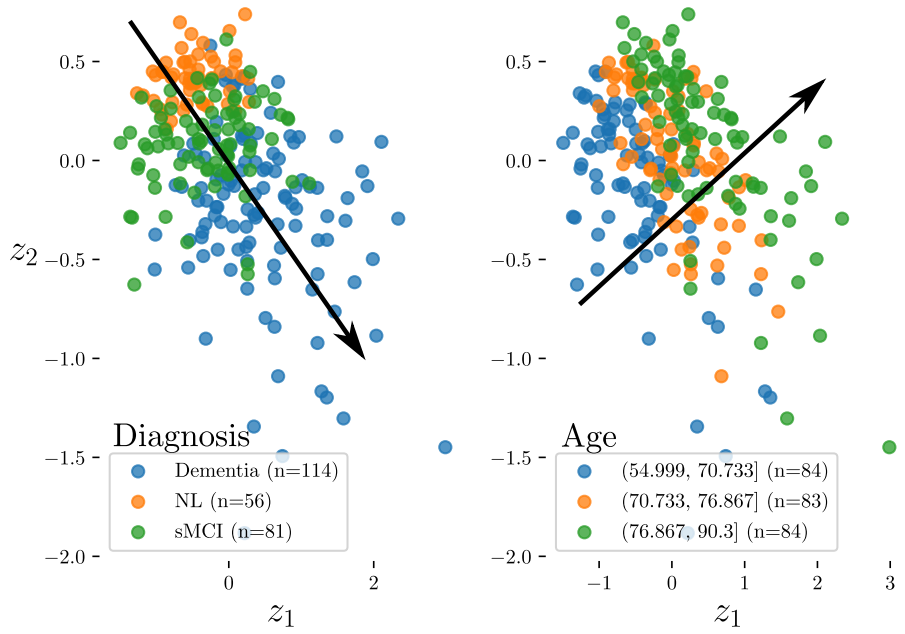
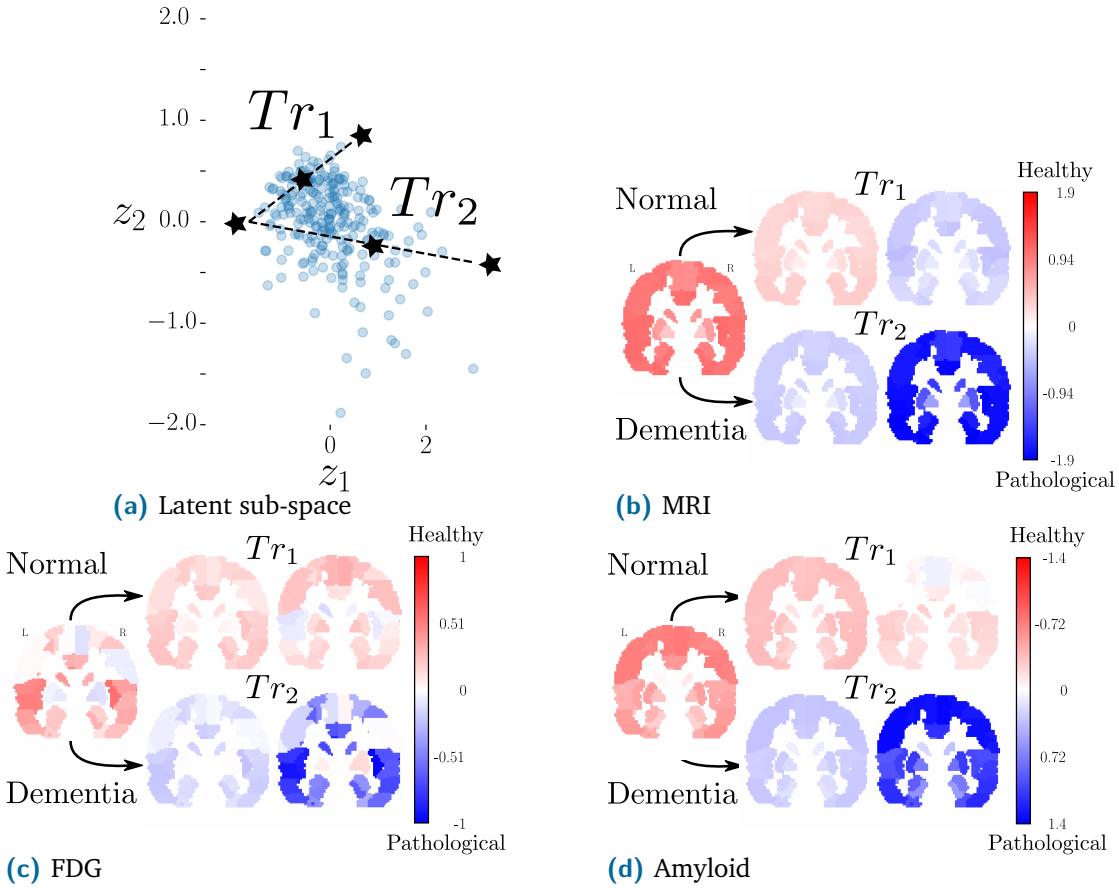


Figure 2.4: Stratification of the ADNI subjects (test data) in the sparse latent subspace inferred from the first two least dropped out dimensions. In the same subspace it is possible to stratify subjects in the test-set by: (left) disease status among Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Normal Cognition (NC), (right) age, in almost orthogonal directions. Classification accuracy for these subjects is given in the fifth numeric column of Tab. 2.3.

2.4 Conclusion

This paper introduces the Sparse Multi-Channel Variational Autoencoder (MCVAE), an extension of the Variational Autoencoder (VAE), 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 gen-



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

Figure 2.5: 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. Stars indicate the sampling points along trajectories. The trajectories share the same origin. (b) MRI data evolution. (c) FDG-PET. (d) Amyloid-PET. All the trajectories show a plausible evolution across disease and healthy conditions.

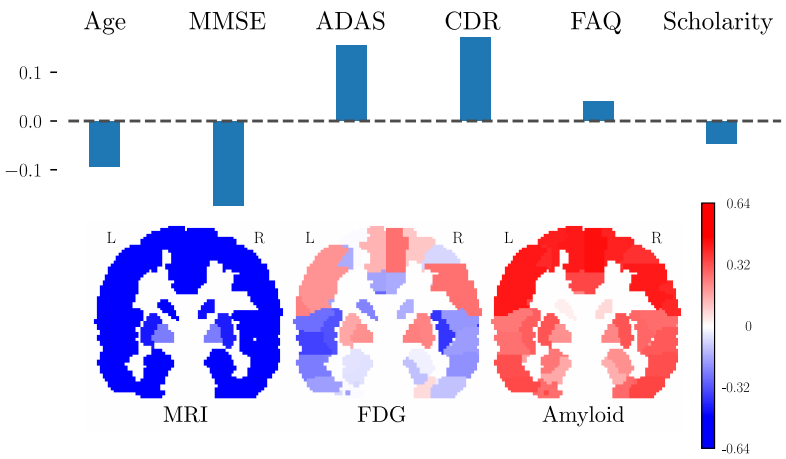


Figure 2.6: Generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (2.7)) of the four channels associated to the least dropout latent dimension in the sparse multi-channel model. (Top) Clinical channel parameters. (Bottom) Imaging ch. parameters.

eral formulation, the proposed method can find various applications as a general data interpretation technique, not limited to the biomedical research area.

2.5 Supplementary Material

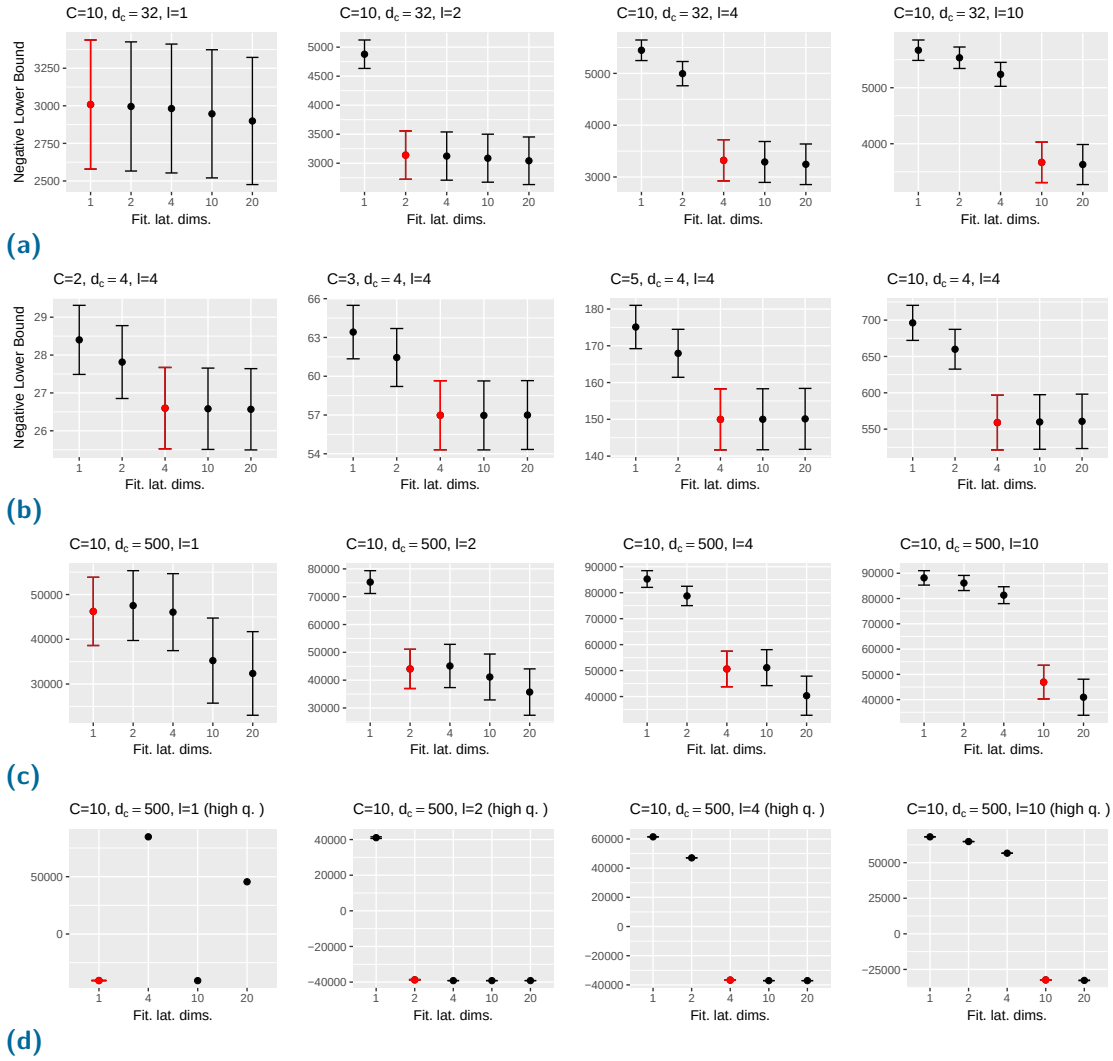
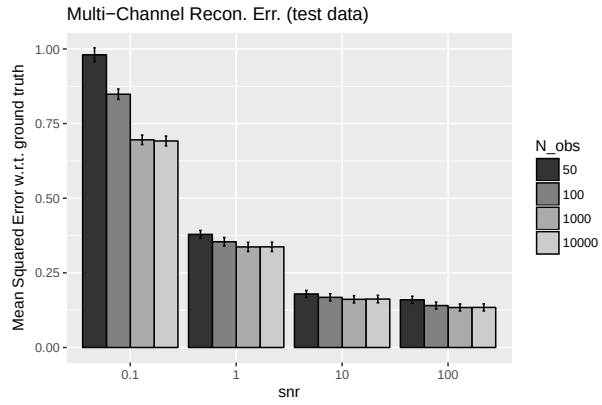
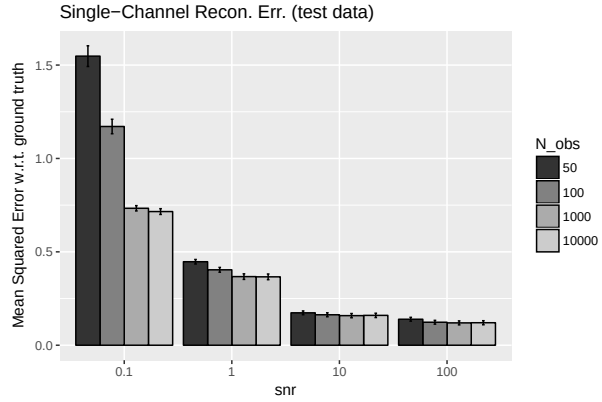


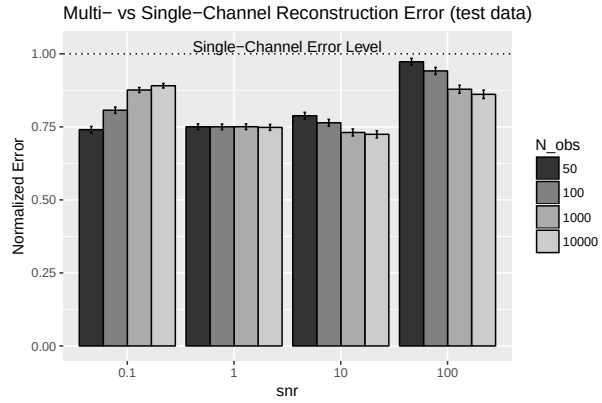
Figure 2.7: Negative lower bound (NLB) on the synthetic training set computed at convergence for all the scenarios. Each bar shows mean $\pm$ std.err. of $N = 80$ total experiments as a function of the number of fitted latent dimensions. Red bars represents experiments where the number of true and fitted latent dimensions coincide. (a) Experimental setup $C = 10$, $d_c = 32$: NLB stops decreasing when the number of fitted latent dimension coincide with the generated ones; notable gap between the under-fitted and over-fitted experiments (elbow effect). (b) Experimental setup $d_c = 4$, $l = 4$: increasing the number of channels C makes the elbow effect more pronounced. (c) Experimental setup $C = 10$, $d_c = 500$: with high dimensional data ($d_c = 500$) using the lower bound as a model selection criteria to assess the true number of latent dimensions may end up in overestimation. (d) Restricted ($N = 5$ total experiments) high quality experimental setup $C = 10$, $d_c = 500$, $S = 10000$, $\mathrm{SNR} = 100$: the risk to overestimate the true number of latent dimensions can be mitigated by increasing the SNR and S of the observations in the dataset.



(a)



(b)



(c)

Figure 2.8: Reconstruction error on synthetic test data reconstructed with the multi-channel model. The reconstruction is better for high SNR and high training data sample size. Scenarios where generated by varying one-at-a-time the dataset attributes listed in Tab. 3.12 for a total of $8\,000$ experiments. (a) Mean squared error from the ground truth test data using the Multi-Channel reconstruction: $\hat{\mathbf{x}}_i = \mathbb{E}_c \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)} \left[p\left(\mathbf{x}_i | \mathbf{z}, \boldsymbol{\theta}_i\right) \right] \right]$. (b) Mean squared error from the ground truth test data using the Single-Channel reconstruction: $\hat{\mathbf{x}}_i = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\phi}_i)} \left[p\left(\mathbf{x}_i | \mathbf{z}, \boldsymbol{\theta}_i\right) \right]$. (c) Ratio between Multi- *vs* Single-Channel reconstruction errors: we notice that the error made in ground truth data recovery with multi-channel information is systematically lower than the one obtained with a single-channel decoder.

3 Multi-Task Multi-Channel Variational Autoencoder

Contents

3.1 Introduction **32**
3.2 Method **35**
3.2.1 Generative Model 36
3.2.2 Inference Model 36
3.2.3 Optimization 37
3.2.4 Comparison with VAE and MCVAE 38
3.2.5 Parameterization 40
3.3 Experiments **41**
3.3.1 Illustration on a simplified brain imaging dataset 41
3.3.2 Synthetic Experiments 43
3.3.3 Experiments on Brain Imaging Data 46
3.4 Discussion **58**
3.5 Conclusions **60**
3.6 Supplementary Material **60**
3.6.1 Derivation of the Lower Bound 61
3.6.2 Data Generation 62

In the previous chapter we presented the Multi-Channel Variational Autoencoder (MCVAE), a latent variable framework allowing to jointly model heterogeneous data. We used the term *channel* as a reference to a group of homogeneous observations. In this chapter, however, we decided to change the nomenclature an use the term *view*, as we found it more appropriate to describe a group of homogeneous features, which indeed gives us a "view", partial and non exhaustive, of the phenomena being measured. In medical imaging data modeling, it is often necessary to increase the sample size by pooling together data from multiple datasets. Training the MCVAE with data coming from multiple datasets is possible with some limitations: 1) after having discarded observations with missing views; 2) when all the datasets have all the views that we want to model. In this chapter we extend the capabilities of the MCVAE framework with a specific optimization scheme that allows the simultaneous learning from multiple datasets, without discarding any observation. This chapter is under review at NeuroImage [Antelmi, 2021].

Abstract

The joint modeling of neuroimaging data across multiple datasets requires to consistently analyze high-dimensional and heterogeneous information in presence of often non-overlapping sets of views across data samples (e.g. imaging data, clinical scores, biological measurements). This analysis is associated with the problem of missing information across datasets, which can take place in two forms: missing at random (MAR), when the absence of a view is unpredictable and does not depend on the dataset (e.g. due to data corruption); missing not at random (MNAR), when a specific view is absent by design for a specific dataset. In order to take advantage of the increased variability and sample size when pooling together observations from many cohorts, and at the same time cope with the ubiquitous problem of missing information, we propose here a multi-task generative latent-variable model where the common variability across datasets stems from the estimation of a shared latent representation across views. Our formulation allows to retrieve a consistent latent representation common to all views and datasets, even in the presence of missing information. Simulations on synthetic data show that our method is able to identify a common latent representation of multi-view datasets, even when the compatibility across datasets is minimal. When jointly analyzing multi-modal neuroimaging and clinical data from real independent dementia studies, our model is able to mitigate the absence of modalities without having to discard any available information. Moreover, the common latent representation inferred with our model can be used to define robust classifiers gathering the combined information across different datasets. To conclude, both on synthetic and real data experiments, our model compared favorably to state of the art benchmark methods, providing a more powerful exploitation of multi-modal observations with missing views.

3.1 Introduction

Because of the inherent complexity of biomedical data and diseases, researchers are required to integrate data across different studies to increase the sample size and obtain better models [Le Sueur, 2020]. When developing integrative models, researchers have to face with multiple concurrent challenges, such as the ones related to datasets interoperability [Tognin, 2020], data heterogeneity [Buch, 2020], and data missingness [Golriz Khatami, 2020]. Emblematic is the case of integrative modeling when datasets come from multi-centric studies in cognitive and neurological disorders, such as in Alzheimer's Disease (AD). Here the datasets interoperability is hampered by the existence of different protocols between studies. Because of this, methods whose modeling task are specifically designed on one dataset cannot be directly applied to another one.

Furthermore, at the level of each single dataset, researchers face the challenge of modeling heterogeneous data, such as multiple imaging modalities, clinical scores and biological measurements. Each of these sources of information represents an important and independent "*view*" on the disease or phenomena under investigation. Efforts to model multi-view data are increasing in the recent biomedical literature [Vieira, 2020; Venugopalan, 2021], where the objective ranges from predicting clinical outcomes [Chen, 2019b; Abi Nader, 2020; Tabarestani, 2020] to synthesizing new modalities [Wei, 2019; Wei, 2020; Zhou, 2020]. The key concept of a shared information space between views is widespread in the literature for the joint model of multi-view data. This is the case for well established multivariate linear methods such as Canonical Correlation Analysis (CCA), Partial Least Squares (PLS), Independent Component Analysis (ICA), which are some of the most popular methods for multivariate analyses on imaging data, as documented in a multitude of works from the state of the art (see [Liu, 2014] for a general review). While these studies essentially focus on the general problem of multivariate association modeling, multi-view methods specifically tailored to medical imaging tasks, such as image registration and segmentation have been proposed in parallel. For example, in [Qin, 2019] the authors propose a registration method for aligning intra-subject multi-view images. Although limited to a two images registration setting, in this work views are projected into a common latent space. The proposed registration approach is then built on the latent code and on an image-to-image translation approach. [Chartsias, 2021] propose a segmentation method based on the learning of information presented jointly in complementary imaging views. From the different inputs views, anatomical factors are encoded into a common latent space and fused to extract more accurate segmentation masks. In [Yang, 2020] a cross-modality segmentation pipeline is built around a similar concept. In all the works cited so far, the problem of missing data, specifically of missing views during training of multi-view methods, is generally not addressed nor considered. Still, this is a very common problem when joint modeling multiple datasets, especially in neuroimaging research. At the level of the single dataset, views can be missing at random (MAR) for some subjects. Typically, as fitting multi-view models requires to establish correspondences between views, observations with at least one missing view are generally discarded, yielding to potentially severe loss of available information. To mitigate this problem, imputation methods can be applied to infer missing views, by modeling the relationship across views from complete observations. The loss of information is exacerbated when considering multiple datasets altogether. Indeed, according to the cohort study design, there may be views which are specifically absent for a given dataset, hence missing not at random (MNAR). This potential mismatch across datasets hampers their interoperability, and prevents the gathering of all the available observations into a single, robust and generalizable joint model accounting for the global data variability. This challenge is typically addressed in machine learning by the field of Multi-Task Learning (MTL). To address this issue, MTL aims at improving the model interoperative capabilities by exploiting the information extracted from multiple datasets. In MTL each task is usually associated to the modeling of a specific dataset and its views

only, when the main idea consists is sharing across datasets the parameters learned through each modeling task [Caruana, 1998; DoradoMoreno, 2020]. As an example of MTL, in model-agnostic meta-learning (MAML) [Finn, 2017] the training of a model on a variety of learning tasks enforces the generalization on new datasets after few fine tuning iterations. In the context of data assimilation, MTL is usually achieved with specific output layers for every task, and by including a shared latent representation for all of them [DoradoMoreno, 2020]. This modeling rationale is at the basis of recent MTL based approaches to heterogeneous data assimilation [Wu, 2018; Shi, 2019], especially in medical imaging approaches. For example, in [Zhou, 2019b], the authors propose a staged deep learning framework for dementia diagnosis classification, able to jointly exploit multi-view data, such as Magnetic Resonance Imaging (MRI), Fluorodeoxyglucose Positron Emission Tomography (FDG-PET), and genetic data. Their approach, where at each stage the model learns feature representations for different combinations of views, solves elegantly the problem of missing data. Although inspiring for their use of the maximum number of available data samples at each stage, the combinatorial nature of their framework makes it in practice applicable only for datasets with very few available views. For example, when considering 3 views, this approach requires to learn 7 networks. With 4 views, the number of networks that need to be trained, considering all the possible couples, triplets and quadruplets of views amounts to 4845; while with 5 views it exceeds 10^{32}. Moreover, this framework is currently designed for classification tasks only, excluding the possibility of modality-to-modality prediction. With the EmbraceNet (EN) of [Choi, 2019] the problem of missing data is managed by zero-filling the missing input views and by the application of a specific dropout technique where multinomial samples are used to assign partitions of the latent space to specific views. As there are latent features that are randomly discarded even when the correspondent view is not missing, this represents still a loss of information. Similarly as for the previous work, the proposed framework is currently applicable in classification tasks only. Dropout is at the basis of the Denoising Autoencoder (DAE), as developed by [Gondara, 2018]. Here an overcomplete deep autoencoder maps input views to a higher dimensional space. The initial dropout layer induces random corruption in the input views, making the model robust to missing data. This framework is currently applicable in feature prediction tasks only.

The common underlying assumption of these approaches consists in the existence of a proper transformation into a common latent code for the solution of multiple tasks, whether classification or feature prediction. Based on this general assumption, the Multi-Channel Variational Autoencoder (MCVAE) [Antelmi, 2019] is a recent analysis method allowing the identification of a common latent representation for different views belonging to a single dataset (Fig. 3.1). MCVAE extends currently available approaches to account for non-linear transformations from the data to the latent space, while it can be adapted to multiple tasks, including data reconstruction and classification. In spite of the high modeling flexibility, the extension of this method to the analysis of multiple datasets

is currently challenging. Training the MCVAE in a multi-dataset context is indeed possible with some limitations: 1) after having discarded observations with missing views; 2) when at train time all the observations are compatible in terms of available views.

To overcome these limitations, in this work we investigate an extension of MCVAE to simultaneously learn from multiple datasets, even in the presence of non compatible views between datasets, and missing views within datasets. While our formulation naturally extends the original MCVAE approach, to the best of our knowledge no systematic investigation of this approach for the modeling of multi-view and multi-dataset neuroimaging data has been proposed so far. Our extension is built upon the following steps: 1) defining tasks across datasets based on the identification of data subsets presenting compatible views, 2) stacking multiple instances of the MCVAE, where each instance models a specific task, 3) sharing the models parameters of common views between modeling tasks. Thanks to these actions, the framework here proposed allows to learn a joint model for all the subjects without discarding any information (Fig. 3.2). The common views between tasks act as a bridge and enable the information to flow through all the other views, while, in the training phase, tasks lacking a particular view will simply not contribute to the learning of those view-specific parameters. All the tasks will nevertheless benefit from the parameters they didn't contribute to learn, for the prediction of their missing views. The proposed variational formulation for computing approximate posterior distributions of the latent variables allows fast and scalable training. Being dataset agnostic, our method allows to integrate all the available data into a joint model, gathering all the available information from multiple datasets at the same time.

The rest of this paper is structured as follow. In § 3.2 we present the mathematical derivation of the classical MCVAE model that will be used to derive the proposed framework. In § 3.3.1 we show an illustrative application for the joint modeling of MRI and FDG-PET images when some modalities are missing in the training phase. In § 3.3.2, experiments on synthetic data show that the prediction error of missing views is competitive with respect to the one obtained with state of the art imputation methods. In § 3.3.3, experiments on real data from independent multi-modal neuroimaging datasets show that our model generalizes better than dataset-specific models, in both the cases of data reconstruction and diagnosis classification. Lastly we discuss our results and conclude our work with summary remarks.

3.2 Method

In this section we recall the theoretical framework of the Multi-Channel Variational Autoencoder (MCVAE) developed in our previous work [Antelmi, 2019], which we now extend to tackle the problem of missing data integration. In § 3.2.1 and § 3.2.2 we

introduce our framework, the Multi-Task Multi-Channel Variational Autoencoder (MT-MCVAE), and derive the model in presence of missing data. In § 3.2.3 we propose the new optimization scheme allowing to account for observations with partially missing views. In § 3.2.4 we emphasize the differences between the MCVAE and our current approach. In § 3.2.5 we briefly recall the main parametric functions adopted later in our experiments with missing data. Code developed in Pytorch [Paszke, 2019] is publicly available at `https://gitlab.inria.fr/epione_ML/mcvae`.

3.2.1 Generative Model

Let $\mathcal{D} = \{D_d\}_{d=1}^{D}$ be a collection of D independent datasets, where each dataset $D_d = \{\mathbf{x}_{d,n}\}_{n=1}^{N_d}$ is composed by N_d independent data-points (*e.g.,* subjects in the case of medical imaging datasets). Every dataset D_d is associated with a total number of V_d available views (*e.g.,* sets of clinical scores and imaging derived phenotypes extracted from multiple imaging modalities), and we assume that each data-point $\mathbf{x}_{d,n} = \{\mathbf{x}_{d,n,v}\}_{v=1}^{V_{d,n}}$ is composed by $V_{d,n}$ views, where $V_{d,n} \leq V_d$. With the latest inequality we account for data-points with an arbitrary number of missing views.

For each view $\mathbf{x}_{d,n,v}$ we rely on the following generative latent variable model:

$$\begin{aligned} &\mathbf{z}_{d,n} \sim p\left(\mathbf{z}\right), \\ &\mathbf{x}_{d,n,v} \sim p\left(\mathbf{x}_{d,n,v}|\mathbf{z}_{d,n}, \boldsymbol{\theta}_v\right), \qquad \text{for } v \text{ in } 1 \dots V_{d,n} \leq V_d, \end{aligned} \tag{3.1}$$

where $p\left(\mathbf{z}\right)$ is a prior distribution for the latent variable $\mathbf{z}_{d,n}$ commonly shared by the $V_{d,n}$ views, and where the likelihood functions $p\left(\mathbf{x}_{d,n,v}|\mathbf{z}_{d,n}, \boldsymbol{\theta}_v\right)$ belong to a family of distributions parametrized by $\boldsymbol{\theta}_v$, which represents the view-specific generative parameters shared among all datasets.

3.2.2 Inference Model

The exact solution to the inference problem is given by the posterior $p\left(\mathbf{z}|\left\{\mathbf{x}_{d,n,v}, \boldsymbol{\theta}_v\right\}_{v=1}^{V_{d,n}}\right)$, that is not generally computable analytically. Following [Antelmi, 2019], we can nevertheless look for its approximation through Variational Inference (VI) [Blei, 2017], applied in our specific context of missing data.

The variational approximations $q\left(\mathbf{z}|\mathbf{x}_{d,n,w}, \boldsymbol{\phi}_w\right)$, where $\boldsymbol{\phi}_w$ represents the view-specific variational parameters shared among all datasets, are such that:

$$\ln p\left(\mathbf{x}_{d,n,v}|\boldsymbol{\theta}_v\right) \geq \mathcal{L}_v^{(\mathbf{x}_{d,n})} = \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathcal{L}_{w \to v}^{(\mathbf{x}_{d,n})}, \tag{3.2}$$

where:

$$\mathcal{L}_{w\to v}^{(\mathbf{x}_{d,n})} = \mathbb{E}_{q_{d,n,w}(\mathbf{z})}\left[\ln p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right)\right] - \mathcal{D}_{\mathrm{KL}}\left(q_{d,n,w}(\mathbf{z})||p\left(\mathbf{z}\right)\right) \tag{3.3}$$

is the lower bound associated to the data-point $\mathbf{x}_{d,n}$ when its view v is predicted from its view w. In Fig. 3.1 we sketch the model structure induced by Eq. (3.3). The complete derivation of Eq. (3.2) is detailed in the *Supplementary Material* section of this work.

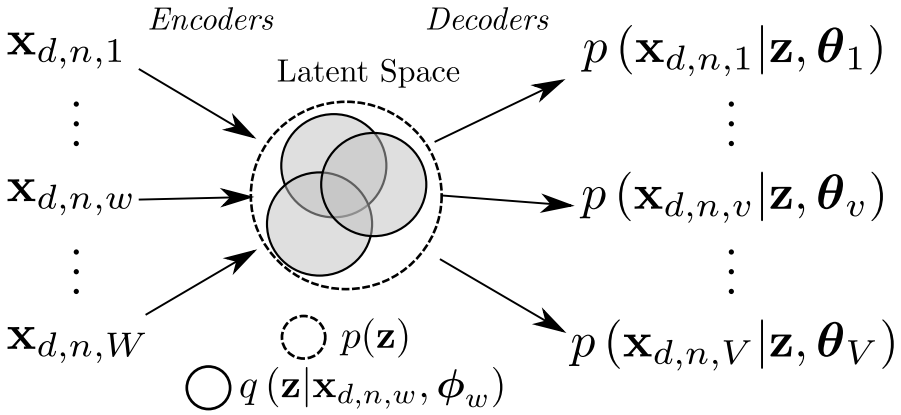


Figure 3.1: General variational framework for our multi-view and multi-dataset model. Compatibly with the MCVAE formulation, for every pair of views w and v there is a prediction path $w \to v$ composed by two learnable functions: the encoding distribution $q\left(\mathbf{z}|\mathbf{x}_{d,n,w},\boldsymbol{\phi}_w\right)$ and the decoding likelihood $p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right)$. Parameters $\boldsymbol{\phi}_w$ and $\boldsymbol{\theta}_v$ are optimized through Eq. (3.4) to maximize the likelihood of our generative model under the encoding distributions, and at the same time minimize the Kullback-Leibler distance between every encoding distribution and the prior $p\left(\mathbf{z}\right)$.

3.2.3 Optimization

Assuming independent observations, the marginal log-likelihood in the left hand side of Eq. (3.2) can be summed up over all the datasets, data-points, and views. As a consequence, inference on the model generative parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_v\}$ and variational parameters $\boldsymbol{\phi} = \{\boldsymbol{\phi}_w\}$ can be achieved by solving the maximization problem:

$$\begin{aligned}\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}} &= \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{d,n,v} \mathcal{L}_v^{(\mathbf{x}_{d,n})} \\ &= \arg\max_{\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{d,n,v} \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathcal{L}_{w\to v}^{(\mathbf{x}_{d,n})}.\end{aligned} \tag{3.4}$$

We implemented Algorithm 1 to solve Eq. (3.4). The summation in Eq. (3.4) is done for every dataset d along all the available data-points n and their specific views v. If missing, a particular view v will be simply not accounted for that specific observation, without having to discard all the other views that can still contribute to optimize Eq. (3.4). We note that batching data-points with common views can speed up the computation by reducing the number of second level *for* loop iterations in Algorithm 1. The presence of at least one common view among datasets acts as a link across datasets and allows the information to flow through all the datasets to the other views. In Fig. 3.2 the learning

Algorithm 1 Multi-view model optimization.

Require:
Set the dimensionality of $\mathbf{z}$.
Initialize the model parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$.
Set the optimizer learning rate.

while $\boldsymbol{\phi}, \boldsymbol{\theta}$ not converged **do**
  Initialize the total cost:
  $\mathcal{L} \leftarrow 0$
  for every dataset $d \in D$ **do**
    for every datapoint $\mathbf{x}_{d,n}, n \in N_d$ **do**
      for every view $v \in V_{d,n}$ **do**
        Accumulate the cost of predicting v from w:
        $\mathcal{L}_v \leftarrow 0$
        for every view $w \in V_{d,n}$ **do**
          $\mathcal{L}_v \leftarrow \mathcal{L}_v + \mathcal{L}_{w \rightarrow v}^{(\mathbf{x}_{d,n})}$. See Eq. (3.3).
        end for
        Accumulate the average $\mathcal{L}_v$ in the total cost:
        $\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{V_{d,n}} \mathcal{L}_v$.
      end for
    end for
  end for
  $\boldsymbol{\theta}, \boldsymbol{\phi} = \mathrm{Optim}(\boldsymbol{\phi}, \boldsymbol{\theta}, \nabla_{\boldsymbol{\phi}} \mathcal{L}, \nabla_{\boldsymbol{\theta}} \mathcal{L})$. *Adam* optimizer used to maximize $\mathcal{L}$.
end while

scheme of our model in a simple case with four views and one common view between batches.

i

3.2.4 Comparison with VAE and MCVAE

In Tab. 3.1 we show how the Multi-Task framework detailed in Algorithm 1 extends the capabilities of the Multi-Channel VAE (MCVAE, [Antelmi, 2019]), which is itself a multi-view extension of the VAE [Kingma, 2014b; Rezende, 2014]. In our former work we proposed a multi-view generative model trainable only with observation in the training set have all the available views, limited to model one dataset at a time (in the case of datasets with multiple views), after having discarded incomplete observations in that dataset. We address this limitation by allowing missing views in the training set for some observations, thanks to the adapted optimization scheme in Eq. (3.4). This aspect naturally extends the training paradigm of MCVAE to the more challenging scenario of multi-dataset analysis. As in the MCVAE, at test time, the trained MT-MCVAE model can

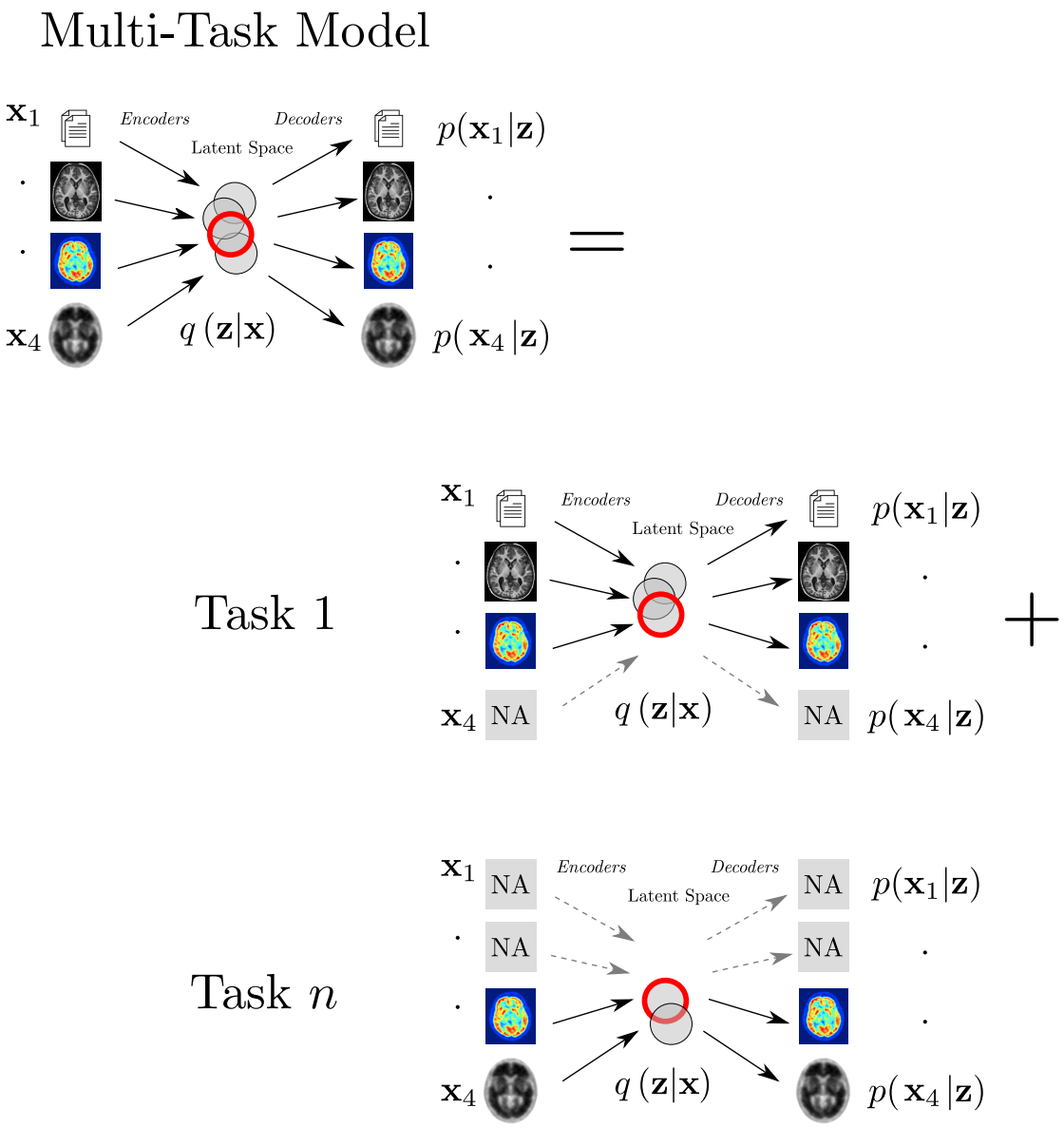


Figure 3.2: Simple example of a Multi-Task Model learning scheme in the presence of missing not available (NA) views. Arrows represent learnable functions used as network encoders and decoders, transforming respectively input views (*e.g.,* clinical scores, imaging derived phenotypes, ...) from the observation space to the representation space (circles) and from the representation space back to the observation space. The separability of the loss function $\mathcal{L}_v^{(\mathbf{x}_{d,n})}$ in Eq. (3.2) allows to group together observations into homogeneous learning tasks. For every task, functions associated to missing views (dashed gray arrows) are locally not updated by the learning algorithm. Globally, common latent representations (red circles) across pairs of tasks act as a link allowing the information to flow throughout the views.

Table 3.1: The Multi-Task Multi-Channel VAE (MT-MCVAE) extends the MCVAE, which is itself an extension of the VAE.

Method	Train with missing data	Test with missing data	# views modeled
VAE	no	no	1
MCVAE	no	yes	> 1
MT-MCVAE	yes	yes	> 1

estimate missing views $\hat{\mathbf{x}}_{d,n,v}$ from the available ones through the formula:

$$\hat{\mathbf{x}}_{d,n,v} = \frac{1}{V_{d,n} - 1} \sum_{w=1,\, w \neq v}^{V_{d,n}} \mathbb{E}_{q_{d,n,w}(\mathbf{z})} \left[p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v \right) \right], \tag{3.5}$$

where the available views $\mathbf{x}_{d,n,w}$ are encoded into the distributions $q_{d,n,w}$, which are then used to predict the missing view through its decoding distribution $p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v \right)$.

3.2.5 Parameterization

With the right choice of the functional form of $q\left(\mathbf{z} | \mathbf{x}_{d,n,w}, \boldsymbol{\phi}_w \right)$, $p\left(\mathbf{z} \right)$, and $p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v \right)$, the right hand side of Eq. (3.2) becomes amenable to computation and optimization, yielding to the maximization of the left hand side, quantity also known as the model evidence. Of course, the choice for the likelihood function $p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v \right)$ depends on the nature of the view $\mathbf{x}_{d,n,v}$. For example it can be parametrized as a multivariate Gaussian in the case of continuous data (*i.e.,*imaging derived phenotypes), as a Bernoulli likelihood for dichotomic data, and as a Categorical likelihood for categorical data.

Linear parameterization

In general, the prior distribution $p\left(\mathbf{z} \right)$ is the multivariate Gaussian distribution $\mathcal{N}\left(\mathbf{0}; \mathbf{I} \right)$. The same family of distributions is also commonly used for the variational and likelihood functions, such that respectively:

$$q\left(\mathbf{z} | \mathbf{x}_{d,n,w}, \boldsymbol{\phi}_w \right) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{V}_w^{(\mu)} \mathbf{x}_{d,n,w} ; \boldsymbol{\Sigma} = \mathrm{diag}\left(\mathbf{V}_w^{(\sigma)} \mathbf{x}_{d,n,w} \right) \right), \tag{3.6}$$

$$p\left(\mathbf{x}_{d,n,v} | \mathbf{z}, \boldsymbol{\theta}_v \right) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{G}_v^{(\mu)} \mathbf{z} \quad ; \boldsymbol{\Sigma} = \mathrm{diag}\left(\mathbf{g}_v^{(\sigma)} \right) \right), \tag{3.7}$$

where the moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are obtained from linear transformations of the conditioning variables. Here, $\boldsymbol{\theta}_v = \{\mathbf{G}_v^{(\mu)}, \mathbf{g}_v^{(\sigma)}\}$ and $\boldsymbol{\phi}_w = \{\mathbf{V}_w^{(\mu)}, \mathbf{V}_w^{(\sigma)}\}$ are the parameters to be optimized. A non-linear parameterization can be used as well, for example in the form of deep neural networks.

In [Antelmi, 2019] we also introduced the following alternative parameterization for the posterior distribution:

$$q_{d,n,w}(\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu} = \mathbf{V}_w^{(\mu)}\mathbf{x}_{d,n,w}; \boldsymbol{\Sigma} = \mathrm{diag}\left(\sqrt{\boldsymbol{\alpha}} \odot \boldsymbol{\mu}^2\right)\right), \tag{3.8}$$

which is known as *dropout posterior* [Kingma, 2015]. The dropout parameter $\boldsymbol{\alpha}$ has components $\alpha_i = {}^{p_i}/_{1-p_i}$ linked to the probability p_i of dropping out the i-th latent variable component [Wang, 2013]. It has been shown that the association of this dropout posterior with a log-uniform prior distribution $p(\mathbf{z})$ leads to sparse and interpretable models [Molchanov, 2017; Garbarino, 2021].

Thanks to the flexibility of modern neural network frameworks, it is straightforward to implement non linear parametrizations $\boldsymbol{\mu} = \mathbf{f}^{(\mu)}(\mathbf{x})$ and $\boldsymbol{\Sigma} = \mathbf{f}^{(\sigma)}(\mathbf{x})$ for the mean and covariance functions in the variational and likelihood distributions. Typically it is done by stacking linear or convolution layers, interleaved with non-linear activation functions such as sigmoid and hyperbolic tangent. This modeling is in general highly task-dependent.

3.3 Experiments

3.3.1 Illustration on a simplified brain imaging dataset

In this section we describe a simple experiment where we use MT-MCVAE to model the joint relationship between Magnetic Resonance Imaging (MRI) and Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) images when there are missing data at training time. The trained model will be then applied on test data to the cross-modality reconstruction problem (MRI to FDG-PET and *vice versa*). Data comes from the 'Adni2' [1] dataset (see details in § 3.3.3, Tab. 3.3), from which we took the MRI and FDG-PET brain imaging modalities. In what follows, each one of this two modalities corresponds to a specific data view. For each subject ($n = 424$, with both MRI and FDG-PET) we extracted 3 brain slices for each one of the sagittal, coronal, and axial plane. The resulting 3816 slices were randomly allocated to a training and testing set with respectively sizes of 3500 and 316 samples. We downsampled the slices to dimension 28×28 (784 pixels). To simulate a datasets with missing views, we controlled for the fraction of observations with complete views (f) in the training set: this procedure is depicted in Fig. 3.3 where we show an example of training dataset created with $f = 1/3$. For our experiments we took all the 3500 training images and we randomly removed MRI and FDG-PET views to obtain different training sets for which $f \in \{0, 0.25, 0.5, 0.75, 1\}$. In the case $f = 0$,

[1] 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.

from each subjects we kept only its MRI or FDG-PET slices, representing the limit case where no direct relationship between views is observable. In the limit case $f = 1$, all MRI and FDG-PET are paired, representing the ideal case of no missing views, that is the working case of the MCVAE [Antelmi, 2019]. We adopted a deep architecture with 4 layers for both encoders and decoders, having ReLU activation functions and layer dimensions of $784 - 1024 - 1024 - 16$ in the encoding and $16 - 1024 - 1024 - 784$ in the decoding path, an architecture inspired from those used by [Andrew, 2013b] and [Wang, 2015] for a similar task on the MNIST dataset [LeCun, 2010]. We adopted a Gaussian likelihood for the decoders, with independent diagonal covariance parameters, and we trained our model with mini-batches of size 500 for 3000 epochs, after setting up the Adam optimizer with a learning rate of 0.001. Training was repeated 5 times, by changing the initialization random seed of the model parameters. In Tab. 3.2 we show

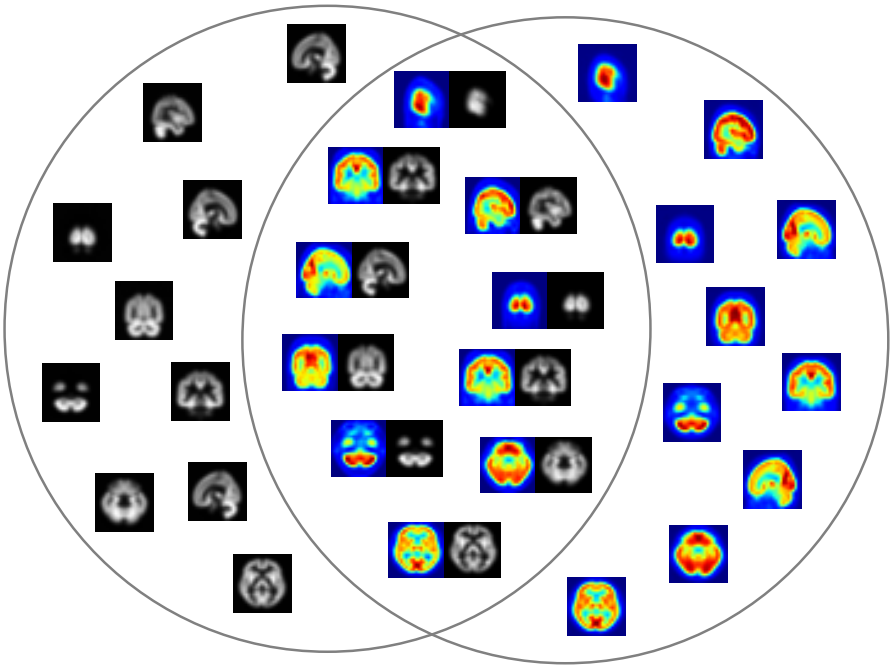


Figure 3.3: Pictorial example of training an imaging dataset with two views: MRI (left side, in gray scale) and FDG-PET (right side, in color scale). In this case we have data from 30 independent observations: 10 with left-views only; 10 with right-views only; 10 with complete views. The fraction of observations with complete views amounts to: $f = 1/3$.

Table 3.2: Mean squared error (MSE) and negative log-likelihood (NLL) - the lower the better - measured as mean $_{\text{st.dev.}}$ on the reconstructed brain images of the test-set. The MRI were used to infer the FDG-PET slices in the same subject, and *vice versa*. Results stratified by f, the fraction of observations with no missing views in the training set. Notice the immediate drop in the error metrics as soon as f increases.

f	0.00	0.25	0.50	0.75	1.00
MSE	$40.72_{4.31}$	$1.77_{0.04}$	$1.63_{0.06}$	$1.54_{0.03}$	$1.51_{0.03}$
NLL	$96.44_{10.33}$	$0.53_{0.09}$	$0.16_{0.12}$	$-0.07_{0.07}$	$-2.63_{0.03}$

the Mean Squared Error (MSE) and Negative Log-Likelihood (NLL) when predicting MRI from the FDG-PET slices and *vice versa* in the testing set. We notice the immediate drop in the error metrics as soon as the parameter f increases, which means that as the model is fed with an increasing proportion of multi-view data points in the training set, its predictions on the testing set become more precise.

3.3.2 Synthetic Experiments

In this section we describe our results on extensive synthetic experiments performed with our model and different benchmark methods in two conditions: 1) missing at random views for each dataset, and 2) datasets with systematically missing views (missing not at random).

Data preparation

To simulate multi dataset observations, we sample the latent variable $\mathbf{z}_{d,n}$ from a multivariate Gaussian with zero-mean and identity covariance matrix, and subsequently we transform each sample with random linear mapping towards the observation space to obtain $\mathbf{x}_{d,n,v}$. The detailed procedure is described in *Sup. Mat.*. We then corrupt the observations with increasing levels of noise and we finally remove views in the context of the missing at random (MAR) and missing not at random (MNAR) experiments.

In the MAR experiments views were randomly removed according to a parameter $0 \leq f \leq 1$, which controls the fraction of data-points with complete views. In the limit case $f = 1$, each data-point has all the views, representing the ideal case of no missing views, that is the working case of the Multi-Channel Variational Autoencoder [Antelmi, 2019]. In the case $f = 0$, each data-point has one and only one randomly assigned view, representing the extreme case where no direct relationship between views is observable. Here our multi-view model collapses into a disjoint series of independent Variational Autoencoders [Kingma, 2014b; Rezende, 2014]. In the general case, each data-point has probability f to have all the views, and probability $1 - f$ to have a randomly assigned view out of the total available views. The general case represents the case where the relationship between views can be established only through a fraction f of the total available data-points.

In the MNAR experiments we removed specific views for each simulated dataset, ensuring at the same time the absence of at least one view for a datasets, and the presence of at least one view in common between pairs of datasets. As an example, in the case with

three datasets and three views, the association view-dataset can be expressed through the following association matrix A:

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \tag{3.9}$$

where $A(v, d) = 1$ indicates the presence of view v in dataset d. For experimental purposes we limited our MNAR simulations to cases that can be defined with square association matrices having a dimensionality not greater than 5×5.

Model Fitting and Evaluation

In both MAR and MNAR experiments we fit the synthetic scenarios with our model, where we choose a linear Gaussian parametrization for variational and likelihood distributions, made explicit respectively in Eq. (3.6) and Eq. (3.7), with a latent dimension matched to the one used to generate the data. We trained our model for 10000 epochs which ensured convergence, after setting up the Adam optimizer with a learning rate of 0.001. For each simulated scenario we predicted the missing views according to Eq. (3.5) on testing hold-out datasets.

Results, cross-validated on 5 folds, are summarized with the Mean Squared Error (MSE) metric on testing hold-out datasets for every simulated scenario. We applied the same evaluation procedure for the benchmark methods.

Benchmark Methods

Among state of the art multivariate linear and non linear imputation methods, we selected the following benchmark approaches: 1) k-Nearest Neighbors (KNN) with $k = \{1, 5\}$; 2) Denoising Autoencoder (DAE) [Gondara, 2018]; 3) Multivariate Imputation by Chained Equations (MICE) [Buuren, 2000].

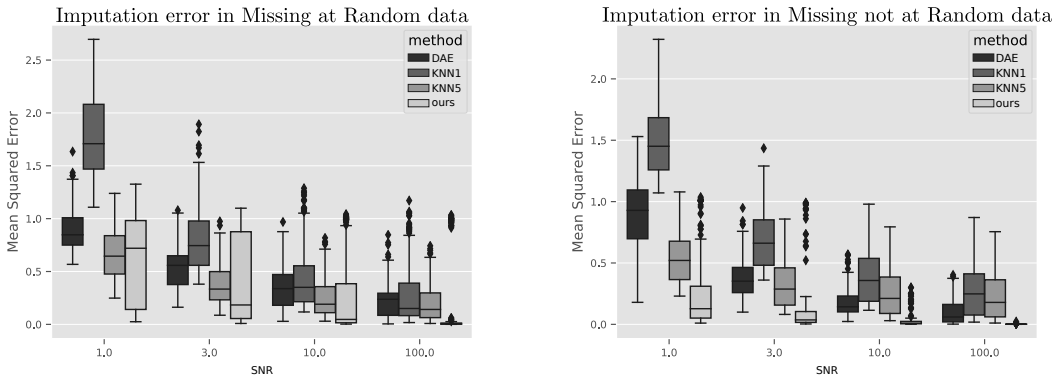
For the KNN approach we used the *KNNImputer* method as implemented in the *Scikit-Learn* library [Pedregosa, 2011]. Here each sample's missing values are imputed using the mean value from k nearest neighbors found in the training set, according to their Euclidean distance.

The Denoising Autoencoder, initially developed by [Vincent, 2008], is based on an overcomplete deep autoencoder. It maps input data to a higher dimensional space which, in combination with an initial dropout layer inducing corruption, makes the model robust

to missing data. We used the same architecture proposed by [Gondara, 2018], that is three hidden layers for encoder and decoder networks, Tanh activation functions, hyperparameter $\Theta = 7$, and dropout $p = 0.5$, as they proved to provide consistently better results.

In MICE, as implemented in [van Buuren, 2011], missing values are modeled as a multivariate linear combination of the available features. This methodology is attractive if the multivariate distribution is a reasonable description of the data, which in our case it is by construction. MICE specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, MICE draws imputations by iterating over the conditional densities.

Results



(a) (b)

Figure 3.4: Mean Squared Error (MSE) of imputation in synthetic held-out datasets (5-folds cross-validation). Compared to the best competing methods among k-Nearest Neighbor ($k = \{1, 5\}$) and Denoise Autoencoder (DAE), our model comes out as the best performer, with a mean MSE improvement of 17% in MAR cases (a) and 71% in MNAR cases (b). Stratification by signal-to-noise ratio (SNR) is shown.

In the synthetic tests our model provides the best performances overall, with a mean MSE improvement compared to the best competing method of 17% in MAR cases and 71% in MNAR cases (Fig. 3.4). We notice that DAE is not always better than KNN ($k = 5$), especially in low Signal-to-Noise Ratio (SNR) cases. We were able to fit the MICE model only on MNAR cases with high SNR, where it performed poorly (boxplot not shown), while in all the other cases, including all MAR cases, the model did not converge.

In Fig. 3.5 we show MAR experiments results stratified by SNR and by the fraction f of data-points with complete views. Here we notice how with already $f = 0.25$ we can significantly reduce the prediction error on testing data-points compared to the case $f = 0$, where no relationship between views can be established. Moreover, reaching the

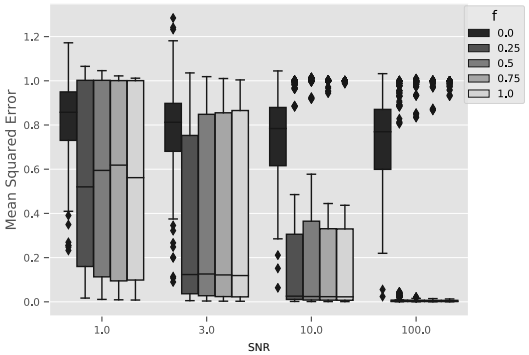


Figure 3.5: Mean Squared Error of test sets predictions in synthetic held-out datasets in MAR scenarios. Stratification by SNR and by the fraction f of data-points with complete views is shown. A value of $f = 0.25$ is enough to reduce the prediction error on testing data-points at the level of the ideal case ($f = 1$).

ideal case of $f = 1$, that is when there are no missing views in the dataset, does not improve significantly the prediction performance of our model compared to the case $f = 0.25$.

3.3.3 Experiments on Brain Imaging Data

In this section we describe our results on jointly modeling real medical imaging datasets, independently acquired in the context of dementia studies.

We executed three kinds of experiments: 1) benchmark evaluation of our model against the best competing methods from the previous section; 2) multi-view feature prediction with our model on all the available datasets in multiple training and testing conditions. 3) diagnosis classification from multi-view heterogeneous data in different training and testing conditions.

Data Sources

Data used in the preparation of this section were obtained from the following sources.

1. The Alzheimer's Disease Neuroimaging Initiative (ADNI), a database of brain imaging and related clinical data of cognitively normal subjects, and on patients presenting various degrees of cognitive decline.

2. Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) dataset, a database of brain imaging and related clinical data of cognitively normal subjects and patients affected by Alzheimer's disease [Malone, 2013].

3. Open Access Series of Imaging Studies 3 (OASIS-3) dataset, a database of brain imaging and related clinical data of cognitively normal subjects and subjects at various stages of cognitive decline [LaMontagne, 2019].

4. A local cohort collected at the Geneva University Hospitals, with brain imaging and related clinical data of patients with various cognitive disorders.

Subjects enrollment, data collection, and data sharing were approved by the ethic committees associated to each study dataset in accordance with the principles of the Declaration of Helsinki.

The available imaging modalities comes from the following acquisitions:

1. structural Magnetic Resonance Imaging (MRI) to measure anatomical volumes in the brain;

2. Positron Emission Tomography (PET) with ^{18}F-Fluorodeoxyglucose (FDG) tracer, to measure the glucose uptake, which reflects the functional status of the brain;

3. PET with the AV45 tracer, to measure the amyloid deposits in the brain;

4. PET with the AV1451 tracer, to measure the tau protein aggregates in the brain.

We divided the ADNI dataset into two complementary datasets: 'Adni1', composed by subjects recruited in the initial ADNI1 study (2004-2009), and 'Adni2' composed by those subjects subsequently recruited in ADNI-GO, ADNI2, and ADNI3 (2010-ongoing). Data modalities and acquisition protocols of 'Adni1' present differences from those of 'Adni2'. Specifically, in 'Adni1' and 'Adni2' the MRI imaging was performed respectively on 1.5T and 3T scanners. The two cohorts differs also for the presence of PET imaging data. Therefore we consider these two cohorts as separated datasets.

To summarize, we grouped our data into five distinct datasets which we named as follows: 'Adni1', 'Adni2', 'Miriad', 'Oasis3', 'Geneva'.

Table 3.3: Number of subjects per view available in each dataset. The last columns provide the size of the intersection (∩) and union (∪) of subjects with available views. Notice how in the jont set no subject has all the modalities.

View $_{\text{\# features}}$:	clin $_{2}$	MRI $_{99}$	FDG $_{94}$	AV45 $_{94}$	AV1451 $_{94}$	∩	∪
Dataset							
Adni1	740	730	-	-	-	730	740
Adni2	1324	710	424	417	61	53	1324
Miriad	67	67	-	-	-	67	67
Oasis3	529	489	-	148	-	147	529
Geneva	999	-	65	120	54	15	999
Tot. subjects	3659	1996	489	685	115	0	3659
Tot. datasets	5	4	2	3	2		

Imaging Processing

The brain scans were processed in order to have measurements on regions defined in the Desikan-Killiany atlas [Desikan, 2006]. Brain MRI scans were processed with FreeSurfer [2] [Reuter, 2012] to measure brain cortical and sub-cortical volumes, and volumes occupied by the cerebrospinal fluid (CSF), for a total of 99 regions of interest. Relative Standardized Uptake Value (SUVR) was computed voxel-wise for the PET scans (FDG, AV45, AV1451), processed with SPM [Ashburner, 2000]. SUVRs were computed using the cerebellum as reference region, and averaged in the same regions used for the MRI, except those containing the CSF, for a total of 94 regions of interest.

Gathering Observations into Views

Observations from the five available datasets (§ 3.3.3) were grouped into the following views.

1. clin: grouping age and Mini-Mental State Examination (MMSE).

2. MRI: grouping brain volumes computed with FreeSurfer.

3. FDG: average brain glucose uptake measured through the analysis of FDG-PET scans.

4. AV45: average brain amyloid deposits measured through the analysis of AV45-PET scans.

[2] surfer.nmr.mgh.harvard.edu

Table 3.4: Mean Squared Error (MSE) of test data from Adni2. All models were trained on all the available datasets by holding-out data from the Adni2 test dataset. 5-folds cross validation of MSE is shown as $\text{mean}_{\text{st.dev.}}$. Best results in boldface are significant with an α level of 0.01 with respect to both competing methods.

View	model		
	DAE	KNN5	ours
clin	$0.73_{0.14}$	$0.44_{0.05}$	$0.45_{0.07}$
MRI	$1.23_{0.31}$	$0.88_{0.15}$	$\mathbf{0.70}_{0.13}$
FDG	$4.20_{0.56}$	$4.15_{0.59}$	$\mathbf{1.09}_{0.15}$
AV45	$1.45_{0.35}$	$1.20_{0.25}$	$\mathbf{0.89}_{0.15}$
AV1451	$1.54_{0.82}$	$1.44_{0.83}$	$\mathbf{1.05}_{0.45}$

5. AV1451: average brain tau protein aggregates measured through the analysis of AV1451-PET scans.

For each subject belonging to the 'Adni1', 'Adni2', 'Miriad' and 'Geneva' datasets, we choose the first available time-point, or baseline. In 'oasis3', since measurements were mostly acquired in different days, we choose to pair nearby time points across modalities into a single one. Time interval between views within one subject was minimal (AV45 vs MRI: ≤ 90 days, MRI vs clin: ≤ 90 days).

In Tab. 3.3 we show the number of observations stratified by dataset and view. Size of the intersection ($\cap$) and union ($\cup$) of subjects with available views is also provided. Please note that the only view in common across datasets is the clinical one, composed by MMSE and age features only.

We adjusted all the views feature-wise with *ComBat* [Johnson, 2007], a normalization method originally develop in genomics, which was adopted in neuroimaging studies to reduce unwanted sources of variation in the data due to the differences in acquisition protocols among datasets [Fortin, 2017; Fortin, 2018; Orlhac, 2020]. In ComBat, we set the variable 'age' as main regressor, and 'Adni2' as reference dataset for the training set. The ComBat reference dataset for testing was the whole training split.

A final feature-wise standardization step was applied by zero centering the data and by rescaling them to have a unity variance. Standardization parameters were computed on the training sets and applied to training and testing sets.

Experiment 1: Benchmark Validation

The purpose of this experiment is to validate on real data the benchmarked results obtained with the synthetic experiments (§ 3.3.2).

As benchmark methods, we choose the best performers on the synthetic experiments, namely KNN5 and DAE. We choose for our MT-MCVAE model a linear Gaussian parameterization for the likelihood and sparse variational distributions of Eq. (3.7) and Eq. (3.8) respectively, the latter with a latent dimension of 32. We trained it for 20000 epochs which ensured convergence, after setting up the Adam optimizer with a learning rate of 0.001. In testing, we set up a dropout threshold for the latent space of 0.5.

We trained all the models (KNN5, DAE, ours) with data coming from all the datasets except from 'Adni2', left out for testing purposes. We choose the 'Adni2' dataset as testing dataset since it provides all the views, and the highest number of observations per view (Tab. 3.3).

Prediction performances were evaluated with the Mean Squared Error (MSE) metric, measured on the available views in the testing dataset, reconstructed with Eq. (3.5). All results were validated by means of 5-folds cross-validation.

Results In Tab. 3.4, we show the MSE metric on predicting missing views in the testing dataset with our model and with the benckmark ones. Best results are in boldface, which show a clear advantage of using our model and confirm our findings in the synthetic experiments.

Table 3.5: Mean Squared Reconstruction Error (the lower the better) measured on test dataset views (clinical scores and imaging derived phenotypes) predicted with the Multi-Channel VAE (MCVAE) and the Multi Task MCVAE (MT-MCVAE). 5-folds cross-validation results shown as $\text{mean}_{\text{st.dev.}}$. Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Experiments were run in two different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; In all cases the MT-MCVAE performs either similarly or statistically better than the MCVAE, with alpha levels at 0.05 (*), 0.01 (**), and 0.001 (***).

		view	clin		MRI		FDG		AV45		AV1451	
		model	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE
test on	condition	train on										
Adni1	STIB	Adni1	$0.90_{0.12}$	$0.89_{0.13}$	$0.85_{0.11}$	$0.83_{0.12}^{*}$	-	-	-	-	-	-
	STEB	Adni2	$0.91_{0.17}$	$0.77_{0.13}^{*}$	$1.02_{0.23}$	$0.85_{0.11}^{***}$	-	-	-	-	-	-
		Miriad	$0.96_{0.17}$	$1.14_{0.27}$	$0.80_{0.14}$	$0.82_{0.13}^{*}$	-	-	-	-	-	-
		Geneva	-	-	-	-	-	-	-	-	-	-
		Oasis3	$0.83_{0.30}$	$0.54_{0.10}^{*}$	$0.80_{0.15}$	$0.76_{0.11}^{*}$	-	-	-	-	-	-
Adni2	STIB	Adni2	$0.83_{0.11}$	$0.73_{0.15}$	$0.74_{0.13}$	$0.70_{0.11}^{**}$	$0.73_{0.14}$	$0.59_{0.10}^{***}$	$1.03_{0.19}$	$0.80_{0.10}^{***}$	$1.33_{0.5}$	$1.18_{0.52}^{*}$
	STEB	Adni1	$0.77_{0.18}$	$0.80_{0.14}$	$0.74_{0.11}$	$0.75_{0.12}$	-	-	-	-	-	-
		Miriad	$0.73_{0.20}$	$0.71_{0.18}$	$0.78_{0.13}$	$0.77_{0.13}$	-	-	-	-	-	-
		Geneva	$0.47_{0.06}$	$0.48_{0.09}$	-	-	$1.40_{0.21}$	$1.09_{0.15}^{***}$	$1.10_{0.21}$	$0.91_{0.15}^{**}$	$1.34_{0.52}$	$1.05_{0.45}^{***}$
		Oasis3	$0.76_{0.23}$	$0.61_{0.13}$	$0.68_{0.12}$	$0.68_{0.11}$	-	-	$1.32_{0.29}$	$1.13_{0.26}^{***}$	-	-
Geneva	STIB	Geneva	$0.79_{0.34}$	$0.98_{0.52}$	-	-	$3.63_{1.35}$	$3.18_{1.04}^{*}$	$1.82_{0.57}$	$1.76_{0.47}^{*}$	$1.27_{0.82}$	$1.19_{0.67}^{*}$
	STEB	Adni1	-	-	-	-	-	-	-	-	-	-
		Adni2	$2.57_{1.09}$	$2.07_{1.05}$	-	-	$3.01_{1.05}$	$2.69_{0.77}^{*}$	$1.92_{0.90}$	$1.41_{0.39}$	$1.81_{0.81}$	$1.42_{0.66}^{***}$
		Miriad	-	-	-	-	-	-	-	-	-	-
		Oasis3	$1.93_{0.66}$	$2.28_{0.89}$	-	-	-	-	$1.70_{0.51}$	$1.63_{0.55}^{*}$	-	-
Miriad	STIB	Miriad	$3.21_{1.07}$	$3.23_{2.55}$	$6.39_{1.57}$	$6.38_{1.52}$	-	-	-	-	-	-
	STEB	Adni1	$6.90_{3.33}$	$6.49_{3.42}$	$6.60_{1.61}$	$6.73_{1.55}$	-	-	-	-	-	-
		Adni2	$5.60_{2.76}$	$3.97_{3.14}$	$5.93_{1.90}$	$6.59_{1.64}$	-	-	-	-	-	-
		Geneva	-	-	-	-	-	-	-	-	-	-
		Oasis3	$6.80_{6.52}$	$6.24_{4.62}$	$6.29_{1.68}$	$6.23_{1.40}$	-	-	-	-	-	-
Oasis3	STIB	Oasis3	$0.83_{0.33}$	$0.68_{0.28}$	$0.68_{0.13}$	$0.66_{0.12}^{*}$	-	-	$1.58_{0.63}$	$1.22_{0.26}^{***}$	-	-
	STEB	Adni1	$1.20_{0.25}$	$1.23_{0.28}$	$0.78_{0.14}$	$0.79_{0.14}$	-	-	-	-	-	-
		Adni2	$1.11_{0.33}$	$1.09_{0.24}$	$0.89_{0.18}$	$0.76_{0.15}^{***}$	-	-	$0.94_{0.22}$	$1.02_{0.26}^{*}$	-	-
		Miriad	$0.98_{0.21}$	$1.02_{0.20}$	$0.83_{0.18}$	$0.83_{0.18}$	-	-	-	-	-	-
		Geneva	$0.55_{0.28}$	$0.49_{0.26}$	-	-	-	-	$1.23_{0.61}$	$1.11_{0.26}^{*}$	-	-

Experiment 2: Feature Prediction

The purpose of this experiment is to compare, in features prediction experiments, the generalization performance the MCVAE model with respect to our new Multi Task extension (MT-MCVE). This experiment was run in three different conditions:

1. Single Task with Internal Benchmark (STIB): when training and testing data are chosen from the same dataset;

2. Single Task with External Benchmark (STEB): when models trained on one dataset are tested on another one;

3. Multi-Task Learning (MTL): when models are trained on all the available datasets except the testing one.

In STIB and STEB experiments, both MCVAE and MT-MCVAE models are trained on the same views, but while in MCVAE we need to discard observations with missing views from the training set, with MT-MCVAE we can include them by grouping together observations with common views into homogeneous tasks. In MTL experiments, MCVAE models cannot be trained because no observation has simultaneously all the views.

We choose for both MCVAE and MT-MCVAE a linear Gaussian parameterization for the likelihood and variational distributions as in Eq. (3.7) and Eq. (3.8) respectively. Models were trained on all the available views in the training dataset. We trained them for 20000 epochs which ensured convergence, after setting up the Adam optimizer with a learning rate of 0.001. Prediction performances were evaluated with the Mean Squared Error (MSE) metric, measured on the available views in the testing dataset, reconstructed with Eq. (3.5).

Non-linear experiments were also made on the MTL scenarios with our MT-MCVAE model, where the encoding and decoding distributions were parametrized with neural networks with up to 4 layers and LeakyReLU activation functions. In this case we choose hidden dimension as the mean value between the input features and latent dimension (32 features), rounded towards the nearest integer (*e.g.,* for the MRI views and a depth of 3 layers we used a symmetric encoding-decoding architecture with dimensions: $99 - 66 - 66 - 32 - 66 - 66 - 99$). Training for 20000 epochs with Adam and a learning rate of 0.001 ensured convergence.

All results were validated by means of 5-folds cross-validation.

Table 3.6: Mean Squared Reconstruction Error (the lower the better) measured on test dataset views (clinical scores and imaging derived phenotypes) predicted with our model. 5-folds cross-validation results shown as as mean$_{\text{st.dev.}}$. Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Experiments were run in three different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; 3) when models are trained on all the available datasets except the testing one, or Multi Task Learning (MTL). We measure a better performance in the MTL condition with respect to the STIB (§) in $7/12$ of cases, and in $10/12$ of cases with respect to the average STEB (†) experiments.

	view	clin	MRI	AV45
test dataset	condition			
Adni1	STIB	$0.89_{0.13}$	$0.83_{0.12}$	-
	STEB (avg)	$0.82_{0.17}$	$0.81_{0.12}$	-
	MTL	$0.45_{0.07}$§†	$0.77_{0.10}$§†	-
Adni2	STIB	$0.73_{0.15}$	$0.70_{0.11}$	$0.80_{0.10}$
	STEB (avg)	$0.65_{0.14}$	$0.73_{0.12}$	$1.02_{0.21}$
	MTL	$0.45_{0.07}$§†	$0.70_{0.13}$†	$0.89_{0.15}$†
Geneva	STIB	$0.98_{0.52}$	-	$1.76_{0.47}$
	STEB (avg)	$2.18_{0.97}$	-	$1.52_{0.47}$
	MTL	$1.80_{1.16}$†	-	$1.35_{0.37}$§†
Miriad	STIB	$3.23_{2.55}$	$6.38_{1.52}$	-
	STEB (avg)	$5.57_{3.73}$	$6.52_{1.53}$	-
	MTL	$2.31_{1.65}$§†	$6.17_{1.37}$§†	-
Oasis3	STIB	$0.68_{0.28}$	$0.66_{0.12}$	$1.22_{0.26}$
	STEB (avg)	$0.96_{0.25}$	$0.79_{0.16}$	$1.07_{0.26}$
	MTL	$0.72_{0.09}$†	$0.81_{0.15}$	$1.09_{0.30}$§

Table 3.7: Mean Squared Reconstruction Error (mean (st.dev.), the lower the better) measured on clinical scores and imaging derived phenotypes predicted with our MT-MCVAE model in MTL experiments. Results stratified by the number of layers in the encoder-decoder architecture. We measure no significant differences among architectures (anova statistical test at an alpha level of 0.05). Best overall results in boldface.

#layers	1	2	3	4
clin	$\mathbf{0.97}_{0.49}$	$1.05_{0.65}$	$1.04_{0.60}$	$1.02_{0.50}$
MRI	$\mathbf{2.09}_{0.92}$	$2.14_{0.69}$	$2.13_{0.68}$	$2.13_{0.68}$
AV45	$\mathbf{1.09}_{0.29}$	$1.16_{0.25}$	$1.15_{0.26}$	$1.15_{0.25}$

Table 3.8: Number of subjects stratified by dataset and diagnosis: Alzheimer's Disease (AD); Mild Cognitive Impairment (MCI); Normal Cognition (NC).

	AD	MCI	NC	other	total
Adni1	403	172	165	-	740
Adni2	328	455	541	-	1324
Geneva	147	405	-	447	999
Miriad	44	-	23	-	67
Oasis3	149	-	380	-	529

Results In Tab. 3.5 and Tab. 3.6 we show the prediction error in terms of MSE for each test dataset and view, on the three experimental conditions described earlier.

In STIB and STEB cases (Tab. 3.5), the MT-MCVAE model performs either similarly or statistically better than the MCVAE, especially in cases where the difference between the union and intersection set of observations is higher (cfr. Tab. 3.3).

In the MTL scenario (Tab. 3.6) there are 12 cases that could be fitted with MT-MCVAE only. We measure an overall better performance of MTL with respect to STIB ($7/12$ of cases) and with respect to STEB ($10/12$ of cases).

In Tab. 3.7, the results on a non linear application of our method in MTL cases show that no improvement is gained when changing the architecture depth (anova test, alpha level 0.05).

Table 3.9: Experiment of diagnosis classification run with the Multi-Channel VAE (MCVAE) and the Multi Task MCVAE (MT-MCVAE). 5-folds classification accuracy in % is shown as mean (standard deviation). Since there are no MCI in miriad and oasis3 datasets, the classification tests 'AD vs MCI' and 'MCI vs NC' are meaningless and not reported. Since there are no NC in the geneva dataset, the classification tests 'AD vs NC' and 'MCI vs NC' are meaningless and not reported. Experiments were run in two different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case. In all cases the MT-MCVAE model performs either similarly or statistically better than the MCVAE, with alpha levels at 0.05 (*), 0.01 (**), and 0.001 (***).

		classification task	AD vs MCI		AD vs NC		MCI vs NC	
		model	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE	MCVAE	MT-MCVAE
test dataset	condition	train dataset						
Adni1	STIB	Adni1	$72.70_{3.72}$	$72.87_{4.37}$	$81.69_{2.97}$	$81.51_{3.14}$	$62.00_{8.91}$	$62.90_{8.72}$
	STEB	Adni2	$47.48_{3.56}$	$58.96_{3.55}^{***}$	$68.50_{4.86}$	$73.77_{2.80}^{*}$	$53.12_{6.42}$	$59.65_{2.76}^{*}$
		Miriad	-	-	$82.58_{4.75}$	$80.82_{3.16}$	-	-
		Oasis3	-	-	$48.57_{6.48}$	$62.31_{6.43}^{**}$	-	-
		Geneva	$36.52_{5.29}$	$46.61_{8.03}^{*}$			-	-
Adni2	STIB	Adni2	$50.58_{3.90}$	$80.07_{2.53}^{***}$	$82.86_{3.28}$	$87.92_{3.46}^{*}$	$58.63_{4.27}$	$65.56_{1.11}^{**}$
	STEB	Adni1	$57.59_{2.61}$	$58.23_{2.87}$	$64.21_{3.36}$	$64.21_{3.52}$	$63.05_{2.00}$	$62.75_{1.80}$
		Miriad	-	-	$70.32_{7.29}$	$70.20_{7.17}$	-	-
		Oasis3	-	-	$68.24_{2.97}$	$75.72_{1.90}^{**}$	-	-
		Geneva	$64.49_{2.98}$	$63.98_{3.30}$	-	-	-	-
Geneva	STIB	Geneva	$65.76_{3.62}$	$77.70_{8.12}^{*}$	-	-	-	-
	STEB	Adni1	$29.17_{5.87}$	$30.08_{5.49}$	-	-	-	-
		Adni2	$38.61_{15.08}$	$70.11_{2.90}^{**}$	-	-	-	-
Miriad	STIB	Miriad	-	-	$83.85_{13.84}$	$86.70_{15.68}$	-	-
	STEB	Adni1	-	-	$74.18_{14.37}$	$74.18_{14.37}$	-	-
		Adni2	-	-	$74.95_{11.58}$	$78.90_{11.54}^{*}$	-	-
		Oasis3	-	-	$45.71_{18.08}$	$66.04_{19.35}$	-	-
Oasis3	STIB	Oasis3	-	-	$74.47_{2.49}$	$80.35_{3.59}^{*}$	-	-
	STEB	Adni1	-	-	$49.16_{6.34}$	$48.22_{5.78}$	-	-
		Adni2	-	-	$67.86_{3.80}$	$75.42_{4.68}^{*}$	-	-
		Miriad	-	-	$64.48_{8.65}$	$62.02_{9.74}$	-	-

Experiment 3: Diagnosis Prediction

The purpose of this experiment is to compare, in diagnosis prediction experiments, the generalization performance of the MCVAE model with respect to the MT-MCVAE, in the three experimental conditions described earlier: STIB, STEB, and MTL. Diagnostic classes are: Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Normal Cognition (NC).

For both MCVAE and MT-MCVAE we choose a linear Gaussian parameterization for the variational distributions as in Eq. (3.8). To adapt the models to this new classification experiment, we adopt as decoding function for the latent variable $\mathbf{z}$, the following Categorical likelihood:

$$p\left(y_{d,n}|\mathbf{z},\boldsymbol{\theta}\right) = \mathrm{Cat}\left(\pi = \boldsymbol{\theta}\mathbf{z}\right), \tag{3.10}$$

where $y_{d,n}$ is the diagnosis associated to the data-point n in the dataset d. The probability vector π is a two dimensional vector representing the class probability for each of the three binary comparisons across the three diagnostic classes, namely AD *vs* MCI, AD *vs* NC, MCI *vs* NC ,and is parametrized with a linear transformation of the latent $\mathbf{z}$ by the matrix $\boldsymbol{\theta}$.

Non-linear experiments were also made on the MTL scenarios with our MT-MCVAE model, benchmarked against the EmbraceNet (EN) method [Choi, 2019], where the encoding distributions were parametrized with neural networks with up to 4 layers and LeakyReLU activation functions. Training for 20000 epochs with the Adam optimizer and a learning rate of 0.001 ensured convergence.

Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Classes probabilities were inferred from the all the available views in the testing dataset with the following equation:

$$\hat{y}_{d,n} = \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathbb{E}_{q_{d,n,w}(\mathbf{z})}\left[p\left(y_{d,n}|\mathbf{z},\boldsymbol{\theta}\right)\right]. \tag{3.11}$$

We attributed to each subject the diagnostic class with the highest inferred probability.

The performance on test datasets was evaluated by measuring the classification accuracy (%). All results were validated by means of 5-folds cross-validation.

Results In Tab. 3.9 we show the classification accuracy of MCVAE and MT-MCVAE. In STIB and STEB cases, the MT-MCVAE model performs either similarly or statistically better than the MCVAE. There are 7 cases in the MTL condition (Tab. 3.10) that could be

Table 3.10: Experiment of diagnosis classification run with our model. 5-folds classification accuracy in % is shown as mean (standard deviation). Experiments were run in three different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; 3) when models are trained on all the available datasets except the testing one, or Multi Task Learning (MTL). In all cases we measure a better performance in the MTL condition with respect to the average STEB one (†).

	classification task	AD vs MCI	AD vs NC
test dataset	condition		
Adni1	STIB	$72.87_{4.37}$	$81.51_{3.14}$
	STEB (avg)	$52.79_{5.79}$	$72.30_{4.13}$
	MTL	$59.30_{2.08}$†	$81.86_{3.26}$†
Adni2	STIB	$80.07_{2.53}$	$87.92_{3.46}$
	STEB (avg)	$61.11_{3.09}$	$70.04_{4.20}$
	MTL	$67.82_{1.91}$†	$85.16_{2.13}$†
Geneva	STIB	$77.70_{8.12}$	-
	STEB (avg)	$50.10_{4.20}$	-
	MTL	$52.54_{4.82}$†	-
Miriad	STIB	-	$86.70_{15.68}$
	STEB (avg)	-	$73.04_{15.09}$
	MTL	-	$98.46_{3.44}$†
Oasis3	STIB	-	$80.35_{3.59}$
	STEB (avg)	-	$61.89_{6.73}$
	MTL	-	$77.70_{4.22}$†

Table 3.11: Diagnosis classification with our model and the EmbraceNet (EN, [Choi, 2019]). Accuracy in % as mean$_{\text{st.dev.}}$ over 5-folds. Results are stratified by the classification task and by the number of layers in the encoder-decoder architecture. We measure no significant difference among architectures depth (anova test, alpha level 0.05) and between models (t-test, alpha level 0.05).

Condition: MTL (avg)	AD vs NC		AD vs MCI	
	ours	EN	ours	EN
# layers				
1	$85.79_{3.26}$	$85.34_{2.30}$	$59.89_{2.94}$	$61.02_{3.47}$
2	$79.04_{5.56}$	$77.68_{4.86}$	$56.93_{5.02}$	$61.38_{3.21}$
3	$79.78_{5.92}$	$78.60_{5.34}$	$56.55_{5.33}$	$62.07_{3.59}$
4	$82.47_{4.11}$	$77.12_{7.22}$	$57.49_{6.03}$	$61.29_{5.04}$

fitted with the MT-MCVAE model only. In all of them we measure a better performance with respect to the best STEB cases.

In Tab. 3.11, the results on a non linear application of our method in MTL cases show that no improvement is gained when changing the architecture depth (anova test, alpha level 0.05) for both the EmbraceNet and MT-MCVAE models. No significant differences (t-test, alpha level 0.05) are detectable between the EmbraceNet and MT-MCVAE models for any given architecture depth level. This result show that on the classification task the MT-MCVAE is equivalent to advanced MTL approaches from the state of the art.

3.4 Discussion

In both the experiments on synthetic and real data, our model compared favorably with respect to state of the art benchmark methods.

An interesting result is the one presented in Fig. 3.5, suggesting that collecting a minimum amount of data-points with complete views is enough for our model to capture the joint relationship among views. The empirical bound on this minimum level of data-points with all available views amounts to 25%. In fact, in our synthetic tests, training on scenarios with completeness level above this bound does not seem to improve significantly the testing results. This condition may be explained by the high collinearity between features due to the linear mappings used to generate the multi-view data. The same bound may be noticed also in our showcase experiment (§ 3.3.1) where we jointly modeled MRI and FDG-PET brain images. This results suggest that our model can reach its highest prediction power also when data collection resources are scarce, such as in studies were the acquisition of complete observations is hampered by economical reasons or subject dropout.

As a secondary result, we report the positive performance of knn (k = 5) in synthetic scenarios, especially in low snr cases, and on real data experiment, were it is most of the time superior to the DAE. This finding is corroborated by [Platias, 2020] were knn is found to be superior to methods based on autoencoders.

The experimental results on real medical imaging datasets (Tab. 3.5, Tab. 3.9) show on the horizontal axes the clear improvement of our MT-MCVAE method with respect to the MCVAE, that inspired our work, given the very same training and testing conditions for both of models. The features and diagnosis prediction clearly improves when using our method, that allows to not discard observations with missing views in the training phase. On the same tables, when read on the vertical axes, we note that models trained and tested on the same single dataset (STIB cases) tend to be more accurate than those

trained on multiple other datasets (STEB cases). This is an expected result pointing to the issue of "domain shift", i.e. when observations coming from different datasets are not identically distributed, leading to generally high "within task" accuracy, and low generalization ability in the "between task" setting. We want to emphasize that we mitigated this problem with a data harmonization step based on ComBat [Johnson, 2007], one of the state-of-the-art normalization method in biomedical applications [Fortin, 2017; Fortin, 2018; Orlhac, 2020]. For this reason, we believe that the domain shift has a marginal impact for the application proposed in our study, and that those differences on the vertical axes are most likely due to the large variety of number of observations, available views, and differences in stratification by diseases in the datasets (*cf.* Tab. 3.3, Tab. 3.8).

In feature prediction experiments (Tab. 3.6) we showed that MT-MCVAE models trained jointly on multiple neuroimaging datasets (ADNI, MIRIAD, OASIS-3, Geneva cohort) perform generally better than the ones trained on a single dataset. We suspect that there are two reasons explaining these results. The first is that modeling simultaneously multiple datasets with our method brings more variability and information at play, making the generalization to unseen data less prone to prediction errors. The second reason maybe that every decoder, associated to its specific view, acts, through the shared latent space, as a regularizer for all the other decoders.

In experiments where we seek to classify subjects to predict their cognitive status (Tab. 3.10), the MT-MCVAE generalizes better to new unseen datasets when trained jointly on multiple datasets (MTL cases) with respect to cases where the training happens on a single dataset. We notice that the best results happen in cases where testing data and training data come from the same dataset (ST cases), that is when the testing dataset is not anymore unseen to our model. This is a different result than the analogous one in the feature prediction experiments, and we argue that the reason may be due to the lack of the regularization mechanism induced by having concurring decoders. Indeed, the MT-MCVAE classifier is composed by a single decoder only, which can become highly specialized in decoding testing data coming from the same dataset of the training data.

In our non linear experiments we did not capture any improvement by using deep architectures with respect to simple linear mappings, in both feature prediction (Tab. 3.7) and classification tasks (Tab. 3.11) on real neuroimaging datasets. These results are in line with our previous work [Antelmi, 2019], were we benchmarked other auto-encoding based methods on observations coming from the ADNI dataset. We suspect that this result is due to the general high heterogeneity and relatively small sample size of typical neuroimaging data. Our results on the classification task in multi-view and multi-dataset problem also showed that our approach is equivalent to the EmbraceNet [Choi, 2019] recently proposed in the literature (§ 3.3.3). While this finding indicates the ability of MT-MCVAE to provide results compatible with the state of the art in MTL

classification problems, we note that the architecture of our framework enables a much larger set of applications than the one tackled by the EmbraceNet, such as cross-modality reconstruction and cross-dataset dimensionality reduction.

In our work we have thoroughly investigated architectures with a one-to-one correspondence between encoding and decoding views. This makes our model part of the family of the auto-encoders, where the model acts as identity transformation between the input and the output. Other architectures are nevertheless possible, such as the classifier described in § 3.3.3. In general, there may be an m-to-n relationship, with partially overlapping views among m input views and n output views. Investigating the properties of all the possible architectures is beyond the scope of this work.

As final remark, we want to stress that our model is based on the assumption of independent and identical distributed observations. This assumption may be limiting in healthcare datasets, such as the ones used in this work. In our work we mitigated these biases by harmonizing the datasets before applying our model, and we leave the extension and development of a bias-transparent multi-view models to future works.

3.5 Conclusions

We proposed a new multi-task latent variable generative model able to learn simultaneously from multiple datasets, even in the presence of non-overlapping views among all the datasets. The available overlap between pairs of datasets allows the information to flow through all the views in the dataset pool. Since the learned view-specific parameters are shared among datasets, missing views can be automatically imputed for every dataset. The method proposed in this work is a coherent extension of classical variational generative models, making the training fast and scalable. Being dataset agnostic, our method allows to integrate all the available data into a joint model, gathering all the available information from multiple datasets at the same time. We conducted extensive tests for the joint modeling of synthetically generated data and of multi-modal neuroimaging datasets from independent dementia studies and associated clinical data, showing the competitiveness of our method with respect to the state of the art. Thanks to its general formulation, the proposed method can find applications beyond the neuroimaging research field.

3.6 Supplementary Material

3.6.1 Derivation of the Lower Bound

The exact solution to the inference problem induced by Eq. (3.1) is given by the posterior $p\left(\mathbf{z}|\left\{\mathbf{x}_{d,n,v},\boldsymbol{\theta}_v\right\}_{v=1}^{V_{d,n}}\right)$, that is not generally computable analytically. Following [Antelmi, 2019], we can nevertheless look for its approximation $q\left(\mathbf{z}\right)$ through *Variational Inference* [Blei, 2017]. By introducing the latent variational approximation $q\left(\mathbf{z}\right)$, we can derive the lower bound on the marginal log-likelihood for a single data-point as follows:

$$\begin{aligned}\ln p\left(\mathbf{x}_{d,n,v}|\boldsymbol{\theta}_v\right) &= \ln \int p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right) p\left(\mathbf{z}\right)\ d\mathbf{z} \\ &= \ln \int \frac{q\left(\mathbf{z}\right)}{q\left(\mathbf{z}\right)} p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right) p\left(\mathbf{z}\right)\ d\mathbf{z} \\ &= \ln \mathbb{E}_{q(\mathbf{z})}\left[\frac{p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right) p\left(\mathbf{z}\right)}{q\left(\mathbf{z}\right)}\right] \\ &\geq \mathbb{E}_{q(\mathbf{z})}\left[\ln p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right)\right] - \mathcal{D}_{\mathrm{KL}}\left(q\left(\mathbf{z}\right)||p\left(\mathbf{z}\right)\right).\end{aligned} \tag{3.12}$$

To derive the last line of Eq. (3.12) we leverage on the *Jensen's inequality* and collect the result into a new expectation term and in the Kullback-Leibler divergence term ($\mathcal{D}_{\mathrm{KL}}$).

We define the distribution function $q\left(\mathbf{z}\right)$ to depend on a specific dataset d, data-point n, and view w, such that:

$$q\left(\mathbf{z}\right) = q_{d,n,w}(\mathbf{z}) = q\left(\mathbf{z}|\mathbf{x}_{d,n,w},\boldsymbol{\phi}_w\right), \tag{3.13}$$

where $\boldsymbol{\phi}_w$ represents the view-specific variational parameters shared among all datasets. To force a link among views, we impose the inequality Eq. (3.12) to hold for any w in $1\ldots V_{d,n}$. To do so, we average the right hand side of Eq. (3.12) across the $V_{d,n}$ views and rewrite Eq. (3.12) as follows:

$$\ln p\left(\mathbf{x}_{d,n,v}|\boldsymbol{\theta}_v\right) \geq \mathcal{L}_v^{(\mathbf{x}_{d,n})} = \frac{1}{V_{d,n}} \sum_{w=1}^{V_{d,n}} \mathcal{L}_{w\to v}^{(\mathbf{x}_{d,n})}, \tag{3.14}$$

where

$$\mathcal{L}_{w\to v}^{(\mathbf{x}_{d,n})} = \mathbb{E}_{q_{d,n,w}(\mathbf{z})}\left[\ln p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right)\right] - \mathcal{D}_{\mathrm{KL}}\left(q_{d,n,w}(\mathbf{z})||p\left(\mathbf{z}\right)\right) \tag{3.15}$$

is the lower bound associated to the data-point $\mathbf{x}_{d,n}$ when its view v is predicted from its view w.

3.6.2 Data Generation

Data points with V views $\mathbf{x}_{d,n} = \{\mathbf{x}_{d,n,v}\}_{v=1}^{V}$ with $\mathbf{x}_{d,n,v} \in \mathbb{R}^{f_v}$ where created from a common latent code $\mathbf{z}_{d,n} \in \mathbb{R}^{l}$ with l latent dimensions according to the following model:

$$\begin{aligned}
&\mathbf{z}_{d,n} \sim \mathcal{N}\left(\mathbf{0}; \mathbf{I}_l\right), \\
&\boldsymbol{\epsilon}_v \sim \mathcal{N}\left(\mathbf{0}; \mathbf{I}_{f_v}\right), \\
&\mathbf{G}_v = diag\left(\mathbf{R}_v \mathbf{R}_v^T\right)^{-1/2} \mathbf{R}_v, \\
&\mathbf{x}_{d,n,v} = \mathbf{G}_v \mathbf{z}_{d,n} + \mathrm{SNR}^{-1/2} \cdot \boldsymbol{\epsilon}_v,
\end{aligned} \tag{3.16}$$

where for every view v, $\mathbf{R}_v \in \mathbb{R}^{f_v \times l}$ is a random matrix with l orthonormal columns (*i.e.,* $\mathbf{R}_v^T \mathbf{R}_v = \mathbf{I}_l$), $\mathbf{G}_v$ is the linear generative law, and SNR is the signal-to-noise ratio. With this choice, the diagonal elements of the covariance matrix of $\mathbf{x}_{d,n,v}$ are inversely proportional to SNR, *i.e.,* $diag\left(\mathbb{E}\left[\mathbf{x}_{d,n,v}\mathbf{x}_{d,n,v}^T\right]\right) = (1 + \mathrm{SNR}^{-1})\mathbf{I}_{f_v}$. This generative Scenarios where generated by varying one-at-a-time the dataset attributes, as listed in Tab. 3.12.

Table 3.12: Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (3.16).

Attribute description	Iteration list
Total views (V)	3 4 5
Features per view (f_v)	5 10 100
Latent space dimension (l)	2 4 8
Training Samples	100 500 1000
Testing Samples	1000
Signal-to-noise ratio (SNR)	1 3 10 100
Seed (re-initialize $\mathbf{R}_v$)	1 2 3 4 5

4 Benchmark of Harmonization Methods

Contents

4.1 Introduction **64**
4.2 Methods **66**
4.2.1 ComBat 66
4.2.2 CovBat 67
4.2.3 Domain Invariant Variational Autoencoder 67
4.2.4 Qualitative benchmark 68
4.3 Experiments **71**
4.3.1 Synthetic data generation procedure 71
4.3.2 Quantitative benchmark 73
4.3.3 Results 73
4.4 Discussion **74**
4.5 Conclusion **77**

In the previous chapter we developed a Multi-Task (MT) optimization scheme allowing to train the Multi-Channel Variational Autoencoder (MCVAE) without completely discarding data-points if one or more views are missing. This allows to increase the data sample size by gathering observations from multiple datasets. Besides the problem of missing views, another complication when increasing the sample size is due to the bias induced by the *domain shift*. Indeed, when observations come from different non-harmonized datasets, they are not identically distributed, which is the general working hypothesis of many modeling frameworks, such as the MCVAE and MT-MCVAE. For the same reason, models trained on observations coming from one dataset, will perform poorly in an "out-of-sample" observation setting. In the previous chapter we mitigated this problem with a data harmonization step based on ComBat [Johnson, 2007], one of the state-of-the-art normalization methods in biomedical applications [Fortin, 2017; Fortin, 2018; Orlhac, 2020]. Since the introduction of ComBat, data harmonization has been proposed through the use of more complex multivariate and non-linear approaches. This chapter proposes a thorough investigation of the modeling capabilities of some of the most popular harmonization methods from the state of the art, to determine their applicability in analysis scenarios as the ones proposed in Chapter 2 and Chapter 3.

Abstract
The bias induced by the *domain shift*, that is the existence of different protocols between studies, multiple imaging machine manufacturers, image reconstruction softwares, and preprocessing algorithms, creates barriers to the integration of multi-centric datasets. As early *a priori* correction methods at acquisition time requires high-level technical expertise and can be used for prospectively acquired data only, to make the most out of existing data late *a posteriori* correction methods such as ComBat [Johnson, 2007] are becoming more and more of interest for the neuroimaging community. Since in the recent literature more advanced multivariate linear (*e.g.,* ComBat [Chen, 2019a]) and non-linear (*e.g.,* Domain Invariant Variational Autoencoder [Ilse, 2020]) bias correction methods have been also proposed, in this chapter we develop a quantitative framework to benchmark these approaches on extensive data simulation scenarios. Results show that although more advanced methods can indeed perform better in cases where the number of data features is in the order of thousands, ComBat is nevertheless legit for harmonizing datasets with less features. Assuming that the modeling assumptions we made in this chapter are valid for the data we observe, our results *a posteriori* justifies the use of ComBat in the previous Chapter 3, where we harmonized observations with features in the order of hundreds.

4.1 Introduction

One of the major challenges for understanding Alzheimer's Disease (AD) is that the pathology evolves unnoticed for a long period (up to 20 years) before the manifestation of clinical, cognitive, and behavioral recognizable symptoms [Frisoni, 2003; Solomon, 2011; Scarmeas, 2007; Fostinelli, 2020]. Therefore, efforts have focused on finding a set of quantitative biomarkers (*e.g.,* volume, shape, intensity, texture, *etc.*) that would allow an early detection and follow-up monitoring of the AD hallmarks along the disease progression.

Radiomics is a procedure that relies on the quantification of vast amounts of biomarkers using high-throughput computing from medical images, such as for example Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET) [DaAno, 2020]. To demonstrate the potential value of radiomic procedures as a valuable tool for tracking the dementia stages in AD, researchers are required to integrate data across different studies and datasets to increase the sample size, which leads to potentially more robust disease models and statistical inference [Le Sueur, 2020]. However, bias induced by the *domain shift*, that is the existence of different protocols between studies, multiple imaging machine manufacturers, image reconstruction softwares, and

preprocessing algorithms, creates barriers to the integration of multi-centric datasets [Jovicich, 2019]. Indeed, when observations come from different non-harmonized datasets they are not identically distributed, which is the general working hypothesis of many modeling frameworks: aggregating them can lead to poor analysis results, for example due to false discoveries or to miss existing correlations.

There are generally two non exclusive ways to address this issue [DaAno, 2020]: 1) early *a priori* correction at acquisition time, by eliminating or reducing the differences across images that have been acquired on different machines; 2) late *a posteriori* correction by eliminating or reducing the differences across the features with statistical corrective tools. The early correction requires high-level technical expertise [Palesi, 2019] to lower the barriers to participate in multi-centric neuroimaging studies [Jovicich, 2019], and is feasible for prospectively acquired data only. To make the most out of existing data and past investments, most radiomics studies are generally retrospective, that is conducted by collecting data already acquired. This is why late correction methods are becoming more and more of interest for the neuroimaging community.

Among late correction methods, ComBat [Johnson, 2007] has been shown to be superior to the other existing methods (see [DaAno, 2020] for a comprehensive review) in controlling the variation related to the domain shift, increasing the correlation among test re-test replicates, and producing the highest of overall performances. The goal of ComBat is to transform the data from each domain, so they have similar mean and variance for each feature, while retaining the biological information of the data. It can also robustly manage high-dimensional data when sample sizes are small, as it uses an empirical Bayes framework to improve the variance of the parameter estimates, which is important for experiments with limited sample size, meta-analyses and clinical diagnosis.

In the recent work of [Chen, 2019a], the authors studied the cortical thickness measurements derived from MRI images in the Alzheimer's Disease Neuroimaging Initiative (ADNI), showing the existence of scanner effects in the covariance of structural imaging measures that are not harmonized by ComBat. To reduce this bias they proposed CovBat [Chen, 2019a].

In the domain of the machine learning, too, methods to deal with the problem of domain shift, such as the Domain Invariant Variational Autoencoder (DIVA), are being proposed to increase the inference power and generalizability of deep learning based methods. The interest in these methods arise from their non-linear modeling capabilities, whereas CovBat and CovBat, in their original formulation, can only model linear relationships.

As it is important to determine to what extent recently proposed harmonization methods are effective in reducing the domain shift bias, in this work we benchmark, on extensive synthetically generated scenarios, the most promising ones: ComBat, CovBat, DIVA.

4.2 Methods

4.2.1 ComBat

The acronym ComBat stands for "Combining Batches", and it was proposed by [Johnson, 2007] to combine multiple batches from gene expression microarray experiments, to the increase statistical power in detecting biological variations. Since non-biological variations, or "batch effects", are frequently observed across batches, it is inappropriate, in general, to combine datasets without adjusting for batch effects.

ComBat has been recently adopted in the neuroimaging community [Fortin, 2017; Fortin, 2018; Orlhac, 2020] to combine observations coming from different centers, where the "batch effect" is generally induced by the diversity of existing imaging scanners. Here it is assumed that the observations come from m centers, containing each n_i subjects for $i = 1, 2, \ldots, m$. For feature $v = 1, 2, \ldots, p$, let y_{ijv} represent the feature (e.g. volume of a cortical brain region) measure for the subject j at center i. Given these premises, ComBat posits the following generative model:

$$y_{ijv} = \alpha_v + \mathbf{X}_{ij}\boldsymbol{\beta}_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv} \tag{4.1}$$

where α_v is the overall measure for feature v, $\mathbf{X}$ is a design matrix for the covariates of interest (e.g. gender, age), and $\boldsymbol{\beta}_v$ is the feature-specific vector of regression coefficients corresponding to $\mathbf{X}$. The terms γ_{iv} and δ_{iv} represent the additive and multiplicative center effects. The error terms ϵ_{ijv} are assumed to follow a normal distribution with mean zero and variance σ_v^2. ComBat uses an empirical Bayes framework to improve the variance of the parameter estimates of γ_{iv} and δ_{iv}. It estimates an empirical statistical distribution for each of those parameters by assuming that all features v share the same common distribution. After estimating the model parameters, the ComBat residuals and harmonized values are respectively defined as:

$$\begin{aligned} \epsilon_{ijv}^{\text{ComBat}} &= \frac{y_{ijv} - \hat{\alpha}_v - \mathbf{X}_{ij}\hat{\boldsymbol{\beta}}_v - \hat{\gamma}_{iv}}{\hat{\delta}_{iv}} \\ y_{ijv}^{\text{ComBat}} &= \epsilon_{ijv}^{\text{ComBat}} + \hat{\alpha}_v + \mathbf{X}_{ij}\hat{\boldsymbol{\beta}}_v. \end{aligned} \tag{4.2}$$

4.2.2 CovBat

In [Chen, 2019a], the authors show that considerable differences in covariance exist across sites, and that the state-of-the-art harmonization techniques do not address this issue. In particular, in ComBat the error term ϵ_{ijv} in Eq. (4.1) is assumed to be identical distributed across sites. Its covariance matrix, however, may differ across sites. This is why to further improve the harmonization results, they generalize ComBat for the estimation and correction for the residual covariance differences, renaming the harmonization method in CovBat. The detailed procedure is described in [Chen, 2019a], and is based on the Principal Component Analysis (PCA) decomposition of the full data to obtain the full data covariance matrix, assuming that the site-specific covariances can be reconstructed with site-specific eigen-values, and that the residuals ϵ are linear combination of the global eigen-vectors. A tuning parameter of the CovBat model is the desired proportion of variance explained, used to threshold the number of principal components for reconstruction.

4.2.3 Domain Invariant Variational Autoencoder

In [Ilse, 2020], the authors consider the problem of domain shift and how to learn unbiased representations given data from a set of domains $\{d\}$. To do so, they propose the Domain Invariant Variational Autoencoder (DIVA), a Variational Autoencoder (VAE) [Kingma, 2014b; Rezende, 2014] with two output networks: 1) the classic decoding network of the VAE; 2) an auxiliary domain classifier. The latent space $\mathbf{z} = [\mathbf{z}_x, \mathbf{z}_d]$ is composed by two independently encoded sub-spaces: $\mathbf{z}_d$, regularized by its associated learnable domain-conditioned prior $p_{\psi_d}(\mathbf{z}_d|d)$, and $\mathbf{z}_x$, regularized by the Gaussian prior $p(\mathbf{z}_x) = \mathcal{N}(\mathbf{0}; \mathbf{I})$. The domain classifier $p_{\boldsymbol{\theta}_d}(d|\mathbf{z}_d)$ is conditioned only on the sub-space $\mathbf{z}_d$, while the decoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_x, \mathbf{z}_d)$ is also conditioned on the sub-space $\mathbf{z}_x$. With this setup the latent $\mathbf{z}_x$ should retain only data information unrelated to the domain. The cost function to be maximized to train a DIVA model is hence built as follows:

$$\begin{aligned}
\mathcal{L} = & \mathbb{E}_{q_{\boldsymbol{\phi}_x}(\mathbf{z}_x|\mathbf{x})q_{\boldsymbol{\phi}_d}(\mathbf{z}_d|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_x, \mathbf{z}_d)\right] && \text{VAE} \\
& + \mathbb{E}_{q_{\boldsymbol{\phi}_d}(\mathbf{z}_d|\mathbf{x})}\left[\ln p_{\boldsymbol{\theta}_d}(d|\mathbf{z}_d)\right] && \text{Domain Classifier} \\
& - \mathcal{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}_x}(\mathbf{z}_x|\mathbf{x}) \,||\, p(\mathbf{z}_x)\right) && \mathbf{z}_x \text{ regularizer} \\
& - \mathcal{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}_d}(\mathbf{z}_d|\mathbf{x}) \,||\, p_{\boldsymbol{\psi}_d}(\mathbf{z}_d|d)\right) && \mathbf{z}_d \text{ regularizer}
\end{aligned} \tag{4.3}$$

where $\mathbf{x}$ represents the biased observation, d and $\mathbf{z}_d$ are respectively the domain label and the portion of the latent space associated to the domain, $\mathbf{z}_x$ represents the unbiased portion of the latent space, and $\boldsymbol{\phi} = \{\boldsymbol{\phi}_x, \boldsymbol{\phi}_d\}, \boldsymbol{\theta}, \boldsymbol{\psi}_d$ are respectively the encoding, decoding, and prior network parameters. The posterior encoding distributions $q_{\boldsymbol{\phi}_x}(\mathbf{z}_x|\mathbf{x})$ and $q_{\boldsymbol{\phi}_d}(\mathbf{z}_d|\mathbf{x})$ are regularized by minimizing the Kullback-Leibler divergence ($\mathcal{D}_{\mathrm{KL}}$) from their respective prior distributions.

At test time, the unbiased estimation of the data is decoded after setting to zero the portion of the latent space $\mathbf{z}_d$ associated to the classifier:

$$\mathbf{x}^{\text{DIVA}} = \mathbb{E}_{q_{\hat{\phi}_x}(\mathbf{z}_x|\mathbf{x})}\left[\ln p_{\hat{\boldsymbol{\theta}}}\left(\mathbf{x}|\mathbf{z}_x, \mathbf{z}_d = \mathbf{0}\right)\right]. \tag{4.4}$$

By choosing the appropriate architecture for the encoding and decoding networks, the DIVA can model linear and non linear relationships in the data. In this work we choose a linear architecture (DIVA-1), with one fully connected (FC) layer per network, and a non linear one (DIVA-4), with 4 FC layers interleaved with LeakyReLU activation functions.

4.2.4 Qualitative benchmark

For a better understanding on how the data are harmonized, we now apply the methods described in the previous sections to MRI derived brain volumetric features coming from real research datasets, and measure the change in the feature correlation matrix of each dataset before and after harmonization, and between datasets.

To do so, we gathered together MRI cortical volumes from the following three datasets: Alzheimer's Disease Neuroimaging Initiative (ADNI) -Study number 1 (adni1) and -Study number 2 (adni2), Open Access Series of Imaging Studies (OASIS) -Study number 3 (oasis3). For an extended description of the datasets and the feature extraction procedure, see Chapter 3 (§ 3.3.3).

In Tab. 4.1 we show the stratification of all the available data by dataset and diagnosis, along with age statistics.

Table 4.1: MRI Observations stratified by dataset and diagnosis.

	diagnosis				Age
	AD	MCI	NC	Total	$\text{mean}_{\text{st.dev.}}$ [min, max]
Datasets					
adni1	399	168	163	730	$75.19_{6.94}$ [54, 91]
adni2	222	268	220	710	$71.97_{6.96}$ [55, 90]
oasis3	121	1	367	489	$70.62_{9.32}$ [42, 97]

We harmonized the MRI features with ComBat, CovBat, and DIVA, the latter with a linear and non-linear architecture, and measure the change in the feature correlation matrix of data coming from each dataset before and after harmonization, and between datasets.

Results. In Fig. 4.1 we show the correlation matrix of the MRI features before and after harmonization. In Tab. 4.2 and Tab. 4.3 we measure the difference between all the pairs of correlation matrices with the Frobenius norm. We can see that only with

CovBat the correlation matrices are actually harmonized and very similar between centers (Tab. 4.2).

With DIVA, in both linear and non-linear cases, the data seems to be highly corrupted after harmonization (Fig. 4.1), with the harmonized data structure far from the original data structure in terms of Frobenius norm (Tab. 4.3).

Table 4.2: Pairwise Frobenius norms between dataset-specific correlation matrices for every harmonization method. We find that ComBat adjustment does not harmonize the correlation matrices whereas CovBat adjustment shows large reductions in the between-datasets distances. The almost perfect reduction of the Frobenius distances in the DIVA-1 linear cases is spurious, as the original correlation structure is very different from the harmonized ones (see Tab. 4.3, Fig. 4.1). In the DIVA-4 non-linear case Frobenius distances generally increases, which is not ideal for an harmonization method.

	Raw	ComBat	CovBat	DIVA-1 (linear)	DIVA-4 (non-linear)
$\|\|\text{adni1} - \text{adni2}\|\|_F$	9.05	9.05	7.65	0.0004	41.88
$\|\|\text{adni1} - \text{oasis3}\|\|_F$	17.04	17.04	8.97	0.0007	38.30
$\|\|\text{adni2} - \text{oasis3}\|\|_F$	13.02	13.02	8.12	0.0012	10.14

Table 4.3: Pairwise Frobenius norms between between correlation matrices of harmonized and raw data for every dataset. With ComBat the original covariance matrices are unchanged ($||\cdot||_F = 0$). With CovBat the harmonization tends to slightly change the original covariance matrices ($||\cdot||_F < 10$) to harmonize them, while with both DIVA methods the original covariance structures become very different from the original ones ($||\cdot||_F > 40$).

	adni1	adni2	oasis3
$\|\|\text{ComBat} - \text{Raw}\|\|_F$	0.00	0.00	0.00
$\|\|\text{CovBat} - \text{Raw}\|\|_F$	5.53	1.77	7.72
$\|\|\text{DIVA-1} - \text{Raw}\|\|_F$	98.93	100.04	101.75
$\|\|\text{DIVA-4} - \text{Raw}\|\|_F$	54.45	46.27	41.81

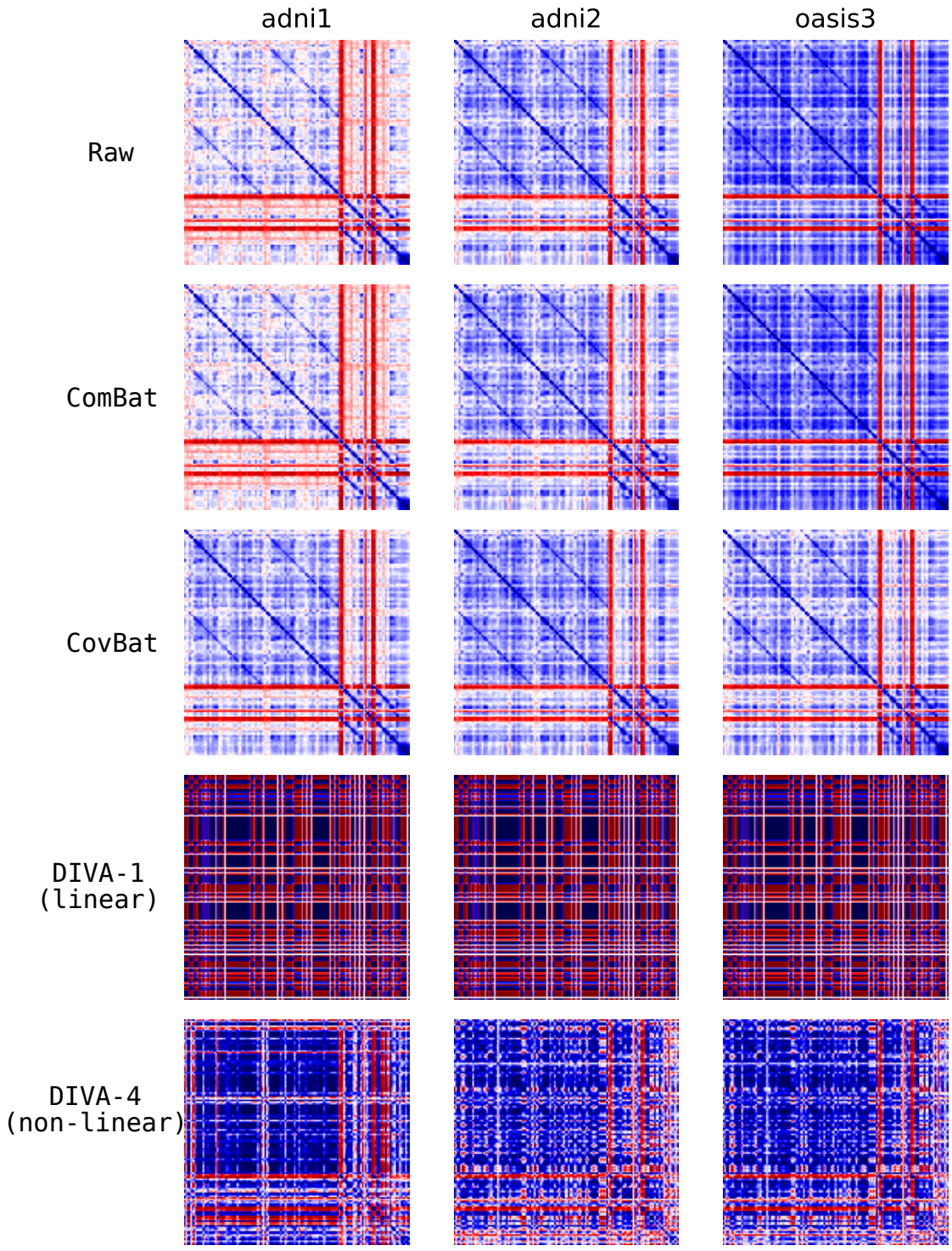


Figure 4.1: Correlation matrices (in range $[-1, 1]$) before and after MRI gray matter volume harmonization. ComBat-adjusted matrices are visually indistinguishable from Raw, where each center is characterized by its own data covariance, whereas between-center differences are still conspicuous, meaning that these covariances are not harmonized. With CovBat covariances are harmonized because between-center differences are less noticeable. With DIVA, both linear an non-linear (4 layers architecture), the original data correlations are lost. All these visual clues are quantified with the Frobenius norm in Tab. 4.2 and Tab. 4.3.

4.3 Experiments

We consider a harmonization method to be successful if it removes the batch effect induced by the domain, and if it preserves biological variability [Fortin, 2017]. Both conditions must be simultaneously tested on the same set of images. This is why we approach the benchmark by simulating datasets presenting features variability induced by a class label y, which may be thought as a diagnosis label, and at the same time being biased with a domain dependent transformation of the ground truth features value.

4.3.1 Synthetic data generation procedure

To generate synthetic data we rely on the `make_classification` [1] dataset generator function from the `sklearn` python library [Pedregosa, 2011]: this function generates a random n-class classification problem by creating clusters of points normally distributed about vertices of an n-informative-dimensional hypercube and assigns an equal number of clusters to each class. This approach introduces interdependence between these features and adds various types of further noise to the data. Redundant features can be added by random linear combination of the informative ones. The `make_classification` function is adapted from [Guyon, 2003] and was designed to generate the "Madelon" dataset, an artificial dataset originally developed for benchmarking two-class classification methods with continuous input variables.

Furthermore, for every dataset d, we can simulate the domain bias with the following procedure:

$$\begin{aligned} \mathbf{x}, y &\sim \texttt{make\_classification(...)} \\ \mathbf{x}_d &= \mathbf{L}_d\mathbf{x} + \boldsymbol{\alpha}_d, \end{aligned} \tag{4.5}$$

where $\mathbf{x}$ is the bias-free observation and y the associated label; the bias inducing $\mathbf{L}_d$ is the Cholesky decomposition of a symmetric positive definite matrix $\boldsymbol{\Sigma}_d = \mathbf{L}_d\mathbf{L}_d^T$. One way to create $\boldsymbol{\Sigma}_d$ is to randomly generate a matrix of k d-dimensional loadings $\mathbf{W} \in \mathbb{R}^{d\times k}$ with $k < d$, then form covariance matrix $\mathbf{W}\mathbf{W}^T$ and add to it a random diagonal matrix $\mathbf{D}$ with positive elements to make $\mathbf{W}\mathbf{W}^T + \mathbf{D}$ full rank. The resulting covariance matrix can be normalized to have ones on its diagonal.

In Fig. 4.2 we can see an example of such matrices generated with k number of loadings. The parameter k can be interpreted as the level of data covariance complexity introduced by biasing the observations with the linear transformation $\mathbf{L}$.

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html`

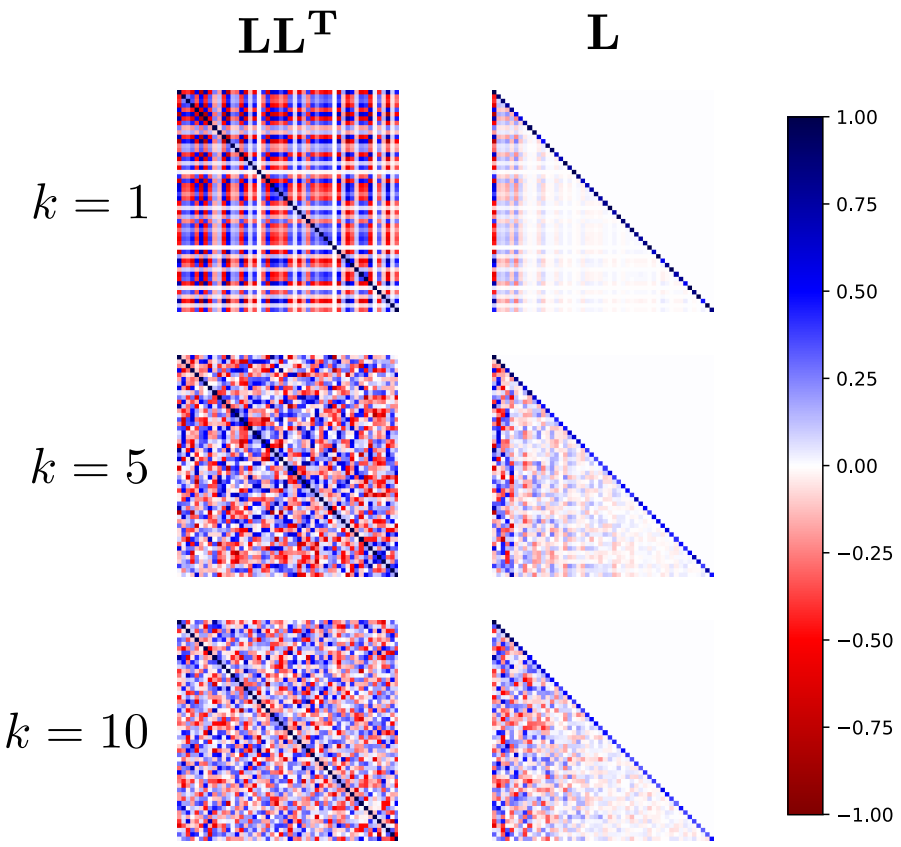


Figure 4.2: The bias inducing matrix $\mathbf{L}$ (see Eq. (4.5)) is the Cholesky decomposition of a symmetric positive definite matrix $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T$. One way to create $\mathbf{\Sigma}$ is to randomly generate a matrix of k d-dimensional loadings $\mathbf{W} \in \mathbb{R}^{d\times k}$ with $k < d$, then form covariance matrix $\mathbf{W}\mathbf{W}^T$ and add to it a random diagonal matrix $\mathbf{D}$ with positive elements to make $\mathbf{W}\mathbf{W}^T + \mathbf{D}$ full rank. The resulting covariance matrix can be normalized to have ones on its diagonal. Here we see examples generated with $k \in \{1, 5, 10\}$ and $d = 50$ features.

In Tab. 4.4 we show the parameters, varied one-at-a-time in the prescribed ranges, used to generate synthetic scenarios for the experimental campaign. The ground truth data $\mathbf{x}$ generated with `make_classification(...)` is composed by an equal number of informative, redundant, and non-informative features.

Table 4.4: Parameters, varied one-at-a-time in the prescribed ranges, used to generate synthetic scenarios for the experimental campaign.

Parameter description	Iteration list
total features	50 100 1000
informative features	$1/3$ of total
redundant features	$1/3$ of total
non-informative features	$1/3$ of total
observations per features	1 2 5 10
complexity of $\mathbf{L}_d$ (k)	1 3 5 7 9
additive bias ($\boldsymbol{\alpha}_d$)	+10

4.3.2 Quantitative benchmark

To assess the performance of an harmonization method we trained, for each synthetic scenario and for each harmonization method, two Linear Discriminant Analysis (LDA) classifiers: one to classify the dataset from the harmonized features; the other one to classify the label y from the same features. As baseline we considered the performances of the same classifiers trained on non-harmonized data ($\mathbf{x}_d$).

For DIVA we choose a linear and non-linear architecture: the linear one (DIVA-1), with encoder and decoders consisting in one linear transformation layers; the non-linear one (DIVA-4) with encoder and decoders consisting in a stack of 4 linear layers, interleaved with LeakyReLU activation functions. The dimension of the latent space is choosen to match the number of informative features used to generate the synthetic scenarios. The DIVA models are trained with Adam [Kingma, 2014a], with learning rate $= 0.001$ for $30k$ epochs, which ensured convergence.

4.3.3 Results

Label classification

In Fig. 4.3 we plot the LDA classifiers accuracy on the label y classification task for different harmonization methods.

As general pattern, we notice that data harmonized with ComBat and CovBat are always easier to classify than the non-harmonized ones. The accuracy increase with a higher

number of training examples given the same number of features in the datasets. We also notice that when the bias complexity "k" induced by the linear transformation $\mathbf{L}_d$ increases, CovBat seems to better harmonize the data, especially in cases with high number of features.

For DIVA, we are not able to identify a clear pattern. In general we notice that the classification performance on DIVA harmonized data are low and in line with those on non-harmonized data. The linear version (DIVA-1) seems to perform generally better than the non-linear one (DIVA-4), although there are very few cases where DIVA-4 is the best overall performer.

Dataset classification

In Fig. 4.4 we plot the LDA classifiers accuracy on the dataset classification task for different harmonization methods. As a good harmonization method would ideally remove the dataset induced bias, the best methods are the ones where the classifier performs badly.

As general pattern, we notice that ComBat and CovBat gives similar results in removing the dataset information, although to reach the random chance level, meaning that the whole information has been removed, we need a high number of data-points.

With the DIVA methods we are not able to identify a clear pattern, although the linear version (DIVA-1) seems to perform generally better than the non-linear one (DIVA-4).

4.4 Discussion

We expect, for a good harmonization method, to improve the classification task for out-of-dataset samples. At the same time we expect the dataset induced bias to be reduced after harmonization, hence reducing a classifier ability to discern between datasets.

As for the DIVA, we notice already in Fig. 4.1 that, although the dataset covariances seem to be harmonized (DIVA-1), they look very different from the original ones (Raw). Indeed, when we measure the distance between datasets (Tab. 4.2, column DIVA-1), this is very low in all cases. In the synthetic experiments results, we also notice that apparently the dataset information is absent from the DIVA-1 harmonized data, as generally the domain classifier has performances around the random-chance level (Fig. 4.4). This is likely to be a spurious result, as the distance between original and DIVA-harmonized data (Tab. 4.3) is the highest among all the harmonization methods. Since there are still cases where

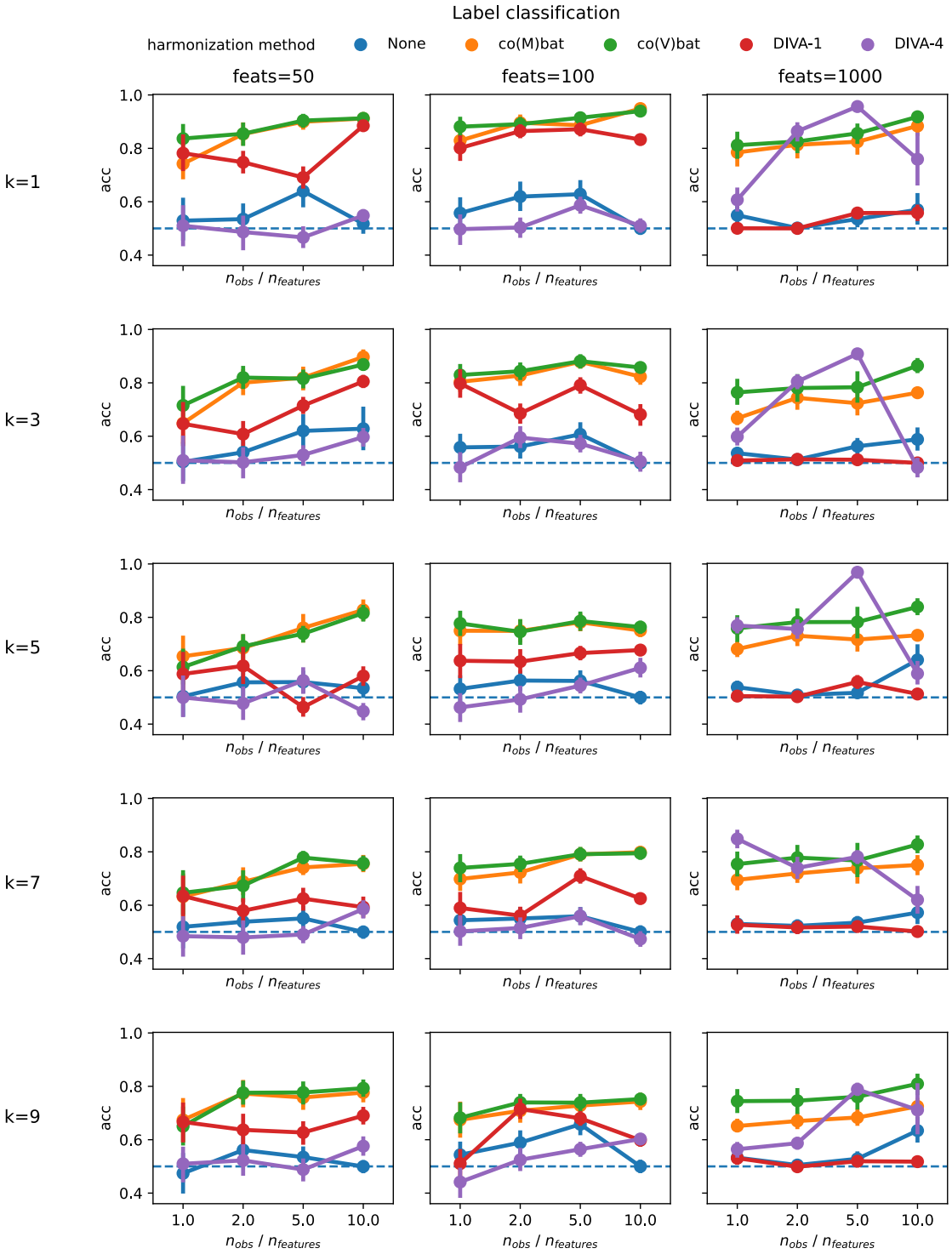


Figure 4.3: Label classification accuracy (the higher the better) of a LDA classifier on harmonized synthetic data created by varying the parameters in Tab. 4.4. Random chance level (acc$=0.5$) rendered as a dashed line. The accuracy is measured on test observations coming from the dataset not used to train the classifier.

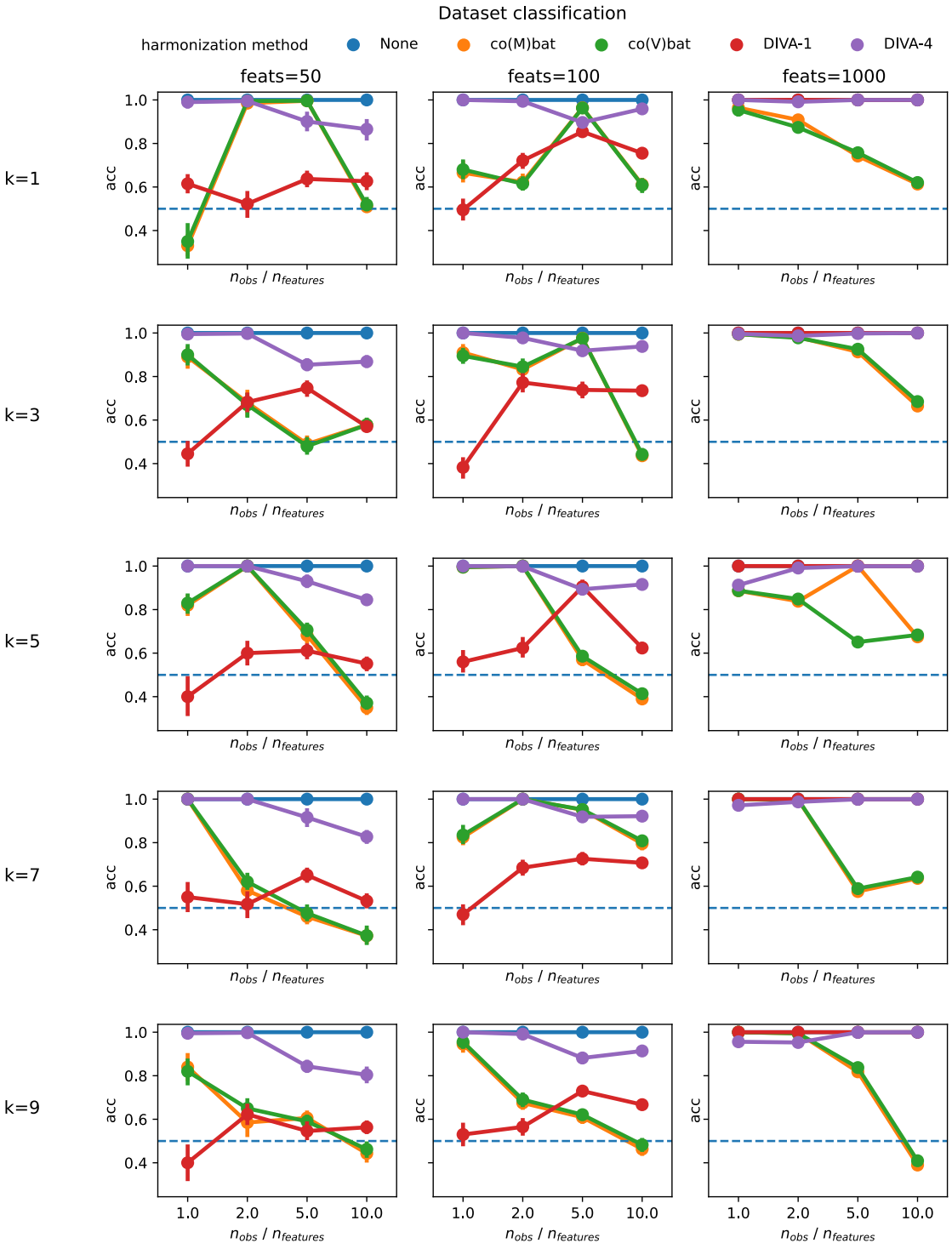


Figure 4.4: Dataset classification accuracy (the lower the better) of a LDA classifier on harmonized synthetic data created by varying the parameters in Tab. 4.4. Random chance level (acc$= 0.5$) rendered as a dashed line.

the DIVA-4 outperforms all the other methods (Fig. 4.3, column feats$= 1000$, top three rows), we argue that, although generally we would not advise to adopt it as a robust harmonization method, DIVA has the potential to compete with the state of the art. For this reason, further research is needed.

The performances of ComBat and CovBat seems to be generally comparable. With CovBat the correlation matrices of the datasets are actually harmonized, as we can see it already in Fig. 4.1. This is confirmed numerically in Tab. 4.2, where we notice a clear reduction in the dataset differences after CovBat harmonization, while the original covariance structure is preserved (Tab. 4.2). In extensive experiments, the performances of the two methods are generally aligned, both in the label classification (Fig. 4.3) and in the datset classification experiments (Fig. 4.4), although we notice a significant difference favoring CovBat in label classification tasks with a high number of features ($n_{\text{feats}} = 1000$).

Lastly, we highlight that our extensive experimental setup is based on the linear generative model of Eq. (4.5), and that supplementary experiments with non-linear generated datasets are necessary to further verify the harmonization performances of the methods here discussed in more complex modeling scenarios.

4.5 Conclusion

In this chapter we benchmarked state of the art harmonization methods, needed to compensate the bias induced by the domain shift, that is the existence of different acquisition protocols between centers, that creates barriers to the integration of multi-centric datasets in dementia studies.

We tested the performances of ComBat and CovBat linear harmonization methods, as well as the Domain Invariant Variational Autoencoder (DIVA) based linear and non-linear ones, on extensive synthetically generated scenarios.

We conclude that the performances of both ComBat and CovBat are robust, and generally aligned between them, and that the latter should be preferred when harmonizing data with a high number of features ($n_{\text{feats}} \geq 1000$). Keeping in mind the limits of the current work discussed earlier, this result can justify *a posteriori* the use of ComBat in Chapter 3, where we harmonized neuroimaging derived features in the order of hundreds (*cf.* Tab. 3.3).

As for the DIVA based methods, we suggest to adopt them cautiously, as in our experiments we noticed data corruption and generally bad performances on label and dataset classification tasks after harmonization. The existence of experimental cases where the

DIVA outperformed all the others suggests that more research is needed to clarify how the DIVA based methods can be improved and made reliable and competitive with the state of the art.

5

`mcvae`: an Open Source Python Toolbox

In this chapter we introduce and document the open-source Python toolbox `mcvae`, for jointly model low- and high-dimensional heterogeneous data. We provide an object-oriented implementation, extensible with custom modules, such as new encoder-decoder architectures. Statistics and learning algorithms provide methods for feature estimation, imputation, and dimension reduction on the latent space. All associated operations are vectorized for batch computation and provide support for the PyTorch backend [Paszke, 2019], enabling GPU acceleration. This chapter presents the package and provides relevant code examples. We show that Mcvae provides reliable building blocks to foster research in joint modeling of heterogeneous data. The source code is freely available online.

Main repository: `https://gitlab.inria.fr/epione_ML/mcvae`.
Mirror: `https://github.com/ggbioing/mcvae`.

5.1 Introduction

High-dimensional heterogeneous data naturally arises in many fields of modern scientific research, such as genomics [Uppu, 2018], biomedical imaging [Miotto, 2018], finance [Sezer, 2020] and so on. Understanding the relationship among heterogeneous data is essential: for example in medical applications, where performing a diagnosis, or understanding the dynamics of a pathology require to jointly analyze multiple data modalities, such as demographic data, medical imaging data, and psychological tests.

In Chapter 2 of this Thesis we developed the Multi-Channel Variational Autoencoder (MCVAE), a method for the joint analysis of heterogeneous observations, which is scalable to high-dimensional observations and high sample sizes. To do so, we generalized the Variational Autoencoder (VAE), a state-of-the-art single modality latent variable model (Fig. 5.1a), by assuming a single low-dimensional latent variable as the common source of the multi-modal observations (Fig. 5.1b). This modeling rationale is well suited to model biomedical data, as the patient can be considered as the common source of all the data collected through imaging and medical examinations. For each data modality, independent encoders (from the observation space to the latent space) and decoders

(from the latent space to the observation space) can be used to track and interpret the informative path from one modality to another, through the common latent space bottleneck.

In Chapter 3 we further proposed a Multi-Task (MT) extension for of the MCVAE to make it robust to missing data, especially when gathering observations from different acquisition centers (Fig. 5.1c). This was done by introducing an optimization scheme specifically designed for maximum data usage, based on the identification of subgroups of observations with common modalities. The common modalities are used to train the encoders and decoders parameters associated to those subgroups, while holding the learning of out-of-group parameters.

The inheritance relationships between the VAE, MCVAE, and MT-MCVAE, can be naturally and elegantly implemented with the Object-oriented programming (OOP) paradigm. To do so we choose the Python programming language and the PyTorch library [Paszke, 2019], which makes the OOP implementation of custom machine learning methods straightforward.

In this chapter we provide an implementation of the MCVAE and MT-MCVAE by presenting the open-source `mcvae` package to 1) reduce duplication of efforts in research; and 2) facilitate research in joint modeling of heterogeneous data. The `mcvae` package comprises the core classes for multi-channel latent variable modeling, with and without missing data, and synthetic datasets generator utilities to simulate modeling scenarios. It was recently presented at the AI4hHealth winter school [1], during the practical session on "Handling heterogeneity in the analysis of biomedical information" [2].

5.2 Implementation overview

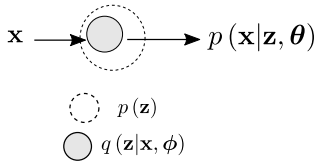
The package `mcvae` is based on the PyTorch deep learning library [Paszke, 2019] and is organized in modules. The module `mcvae.models` contains the latent variable models VAE, MCVAE, and MT-MCVAE. They are implemented as Python classes whose constructors create and initialize their parameters, Since each model extends the previous one, an inheritance scheme is adopted to ease the code readability and the debugging process in development. In Fig. 5.2 we show the inheritance design adopted in our package.

Since the complexity of the models is highly dependent on the modeling task at hand, the user is encouraged to define her/his own VAE module by inheriting from the `mcvae.models.Vae` class, and redefine solely the `init_encoder()` and `init_decoder()`

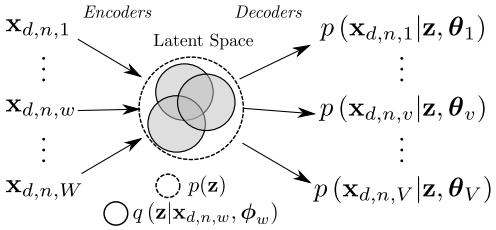
[1] `https://ai4healthschool.org/`

[2] `https://epione.gitlabpages.inria.fr/flhd/heterogeneous_data/heterogeneous_data.html#multi-channel-variational-autoencoder`

(a) Variational Auto Encoder (VAE) model



(b) Multi-Channel VAE model



(c) Multi-Task Multi-Channel VAE model

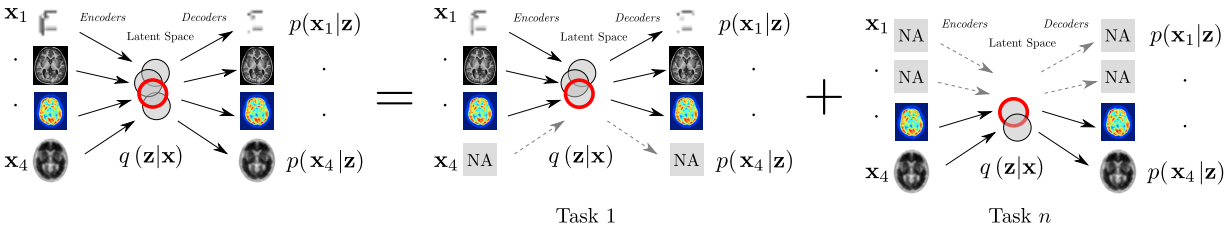


Figure 5.1: Generalization of the (a) Variational Autoencoder (VAE) latent variable model to the (b) multi-channel (or multi-view) case, where multiple related views are encoded into and decoded from the same latent space. In (c) further extension to the multi-task case, where a specific optimization scheme allow missing non available (NA) data in the training phase, to jointly model observations from multiple datasets. Arrows represent learnable functions used as network encoders and decoders, transforming respectively input views (*e.g.,* clinical scores, imaging derived phenotypes, ...) from the observation space to the representation space (circles) and from the representation space back to the observation space. Globally, common latent representations (red circles) across pairs of tasks act as a link allowing the information to flow throughout the views.

methods to output the desired PyTorch distributions (*e.g.,* Normal, Categorical, Bernoulli, *etc.*). The `Mcvae` class builds the MCVAE model based on the input data shape and user defined architecture. The `MtMcvae` class builds the MT-MCVAE model, where the optimization is guided by the observation identifiers (`ids`) needed to correctly pair observations between channels.

The sparsity flag is used to model the encoding distributions with *Dropout posteriors*, that with posterior distributions on the latent variable $\mathbf{z} = \{z_i\}_{i=1}^{L}$ that take the form $z_i \sim \mathcal{N}(\mu_i; \alpha_i \mu_i^2)$ [Kingma, 2015]. The regularization of a dropout posterior depends only on α_i [Molchanov, 2017]. The dropout parameter $\alpha_i = {}^{p_i}/_{1-p_i}$ is linked to the probability p_i of dropping out the i-th latent variable component [Wang, 2013]. It has been shown that the association of this dropout posterior with a log-uniform prior distribution $p(\mathbf{z})$ leads to sparse and interpretable models [Molchanov, 2017; Antelmi, 2019; Garbarino, 2021]. While the regularization promotes $\alpha_i \to \infty$, the implicit drop rate p_i tends to 1, meaning that the associated latent z_i can be discarded. Sparsity arises naturally: large values of z_i correspond to even larger uncertainty $\alpha_i z_i^2$ because of the quadratic relationship and the tendency of the optimization objective to favors $\alpha_i \to \infty$; therefore, unless that latent z_i is beneficial for the optimization objective, that is to maximize the data log-likelihood, it will be set to zero.

We also provide sub-modules and utilities to generate synthetic datasets, with and without missing data, for simulation and benchmark purposes. Scenarios with complete data can be simulated with our `mcvae.datasets.synthetic.py` submodule. The data missingness is defined through utility routines (`mcvae.utilities`) that simulate data missing at random (MAR) and data missing not at random (MNAR). Working examples for fitting MCVAE and MT-MCVAE models are provided within the released python package.

5.3 Usage: example of multi-modal learning

To fit a MCVAE model is just needed to provide multi-modal data formatted as a list of PyTorch tensors $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C]$, where each element of the list corresponds to each data modality. For each data modality, the first dimension of each $\mathbf{x}_c$ must be equal to the number of observations (or subjects) in the dataset, that is $n_c = N$, for $c = 1, \ldots, C$. See Chapter 2 for a complete theoretical background and use cases of the MCVAE model.

If there are missing data in the dataset, that is if $n_c \leq N$, for $c = 1, \ldots, C$, a list of identifiers should be provided for each data modality, to allow the identification of subgroups of observations with common modalities. The theoretical framework for this use case is developed in Chapter 3.

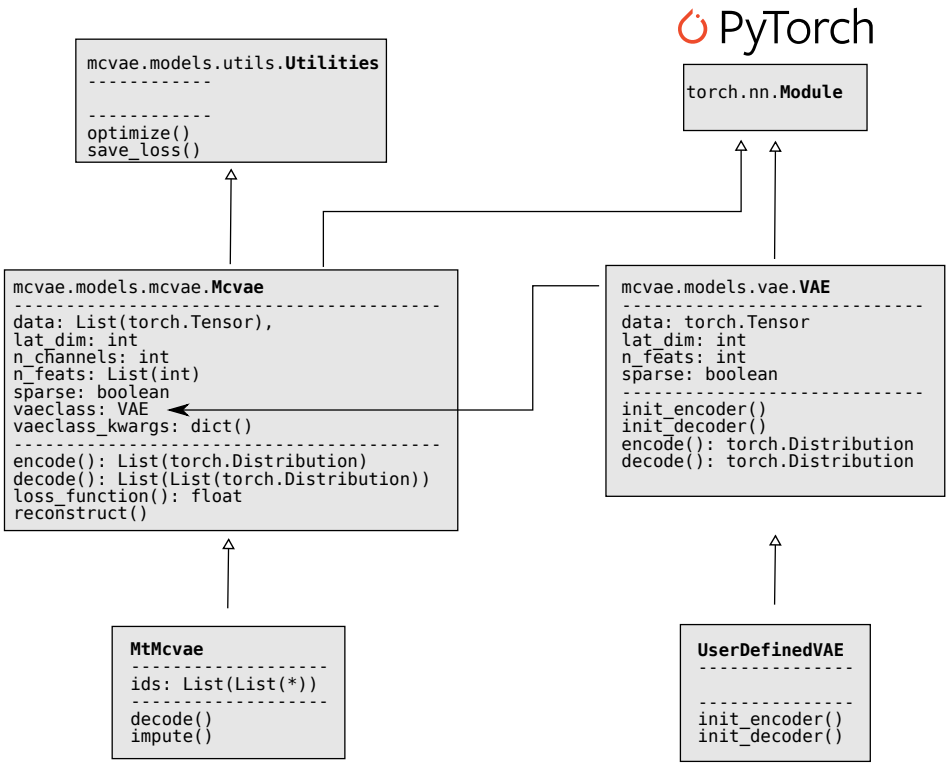


Figure 5.2: Inheritance scheme of the main classes in the `mcvae` package. Class names/attributes/methods are separated by dashed lines. The user can define her/his own VAE module by redefine new `init_encoder()` and `init_decoder()` methods to output the desired distributions (*e.g.,* Normal, Categorical, Bernoulli, *etc.*). The `Mcvae` class builds the MCVAE model based on the input data and user defined architecture. The `MtMcvae` class builds the MT-MCVAE model, where the optimization is guided by the observation identifiers (`ids`) needed to correctly pair observations between channels.

In Listing 5.1 we can see a bare minimum script to fit a MCVAE model. The basic VAE class is used to build the finally architecture of the MCVAE.

The user can also define new VAE based building blocks depending on his/her modeling purposes. In Listing 5.2 we show the prototype of a simple VAE block.

Working examples for fitting MCVAE and MT-MCVAE models are provided within the released python package. It includes examples of modeling synthetic datasets with and without missing data.

```
#!/usr/bin/env python
from torch.optim import Adam
from mcvae.models import Mcvae, VAE
from mcvae.models.utils import load_or_fit
from mcvae.diagnostics import plot_loss

# X must be a list of C tensors,
# corresponding to the channels / views you want to model jointly.
X = torch.load('my_data_file.pt')

# check that there are no missing data
n = len(X[0])
for x in X:
  assert len(x) == n

# Instantiate an empty model
# by choosing the number of latent dimensions
# and set up the sparsity flag
model = Mcvae(
  data=X, lat_dim=15,
  vaeclass=VAE,
  sparse=True,
)

# Choose an optimizer and a learning rate
model.optimizer = Adam(params=model.parameters(), lr=1e-3)

# Load the model from 'ptfile' if exists
# otherwise fit it and save it to 'ptfile'.
load_or_fit(model, data=X, epochs=10000, ptfile='my_model.pt')
```

Listing 5.1: Code to fit a MCVAE model. The user defined VAE class is used as a prototype to build the MCVAE.

```
import torch
from torch.distributions import Normal, kl_divergence


class ConditionalDistributionNet(torch.nn.Module):

```

```
  def __init__(self, in_features, out_features):

    super().__init__()

    self.mu = torch.nn.Linear(in_features, out_features)
    self.logvar = torch.nn.Linear(in_features, out_features)

  def forward(self, x):

    loc = self.mu(x)
    scale = self.logvar(x).exp().pow(0.5)

    return Normal(loc=loc, scale=scale)


class MyVAE(torch.nn.Module):

  def __init__(self, n_feats, lat_dim, *args, **kwargs):

    super().__init__()

    self.encode = ConditionalDistributionNet(
      in_features=n_feats,
      out_features=lat_dim,
    )
    self.decode = ConditionalDistributionNet(
      in_features=lat_dim,
      out_features=n_feats,
    )

  def forward(self, x):

    q = self.encode(x)
    z = q.rsample()
    p = self.decode(z)

    return x, q, p

  def loss_function(self, x, q, p):

    kl = kl_divergence(q, Normal(0, 1)).sum(1).mean(0)
    ll = p.log_prob(x).sum(1).mean(0)
    total = kl - ll

    return total
```

Listing 5.2: PyTorch code that produces a simple linear Variational AutoEncoder. The user can easily extend it with multiple layers, convolution operations, non-linearities, *etc.*, depending on the complexity of the modeling task.

5.4 Supplementary documentation

5.4.1 Main model classes (`mcvae/models`)

- **Mcvae**: main class used to build a Multi-Channel Variational AutoEncoder as in [Antelmi, 2018a; Antelmi, 2019] (usage example in Fig. **??**). Arguments:
  - **data**: example of a multi-channel dataset from which infer the model architecture (n_channels, n_feats)
  - **lat_dim**: number of latent dimension
  - **n_channels**: number of channels. Can be inferred from "data".
  - **n_feats**: number of features for each channel. Can be inferred from "data".
  - **beta**: scaling factor for Kullback-Leibler distance.
  - **enc_channels**: specify the channels to encode from.
  - **dec_channels**: specify the channels to decode.
  - **sparse**: True for a sparse model (default False).
  - **vaeclass**: basic class for building the Mcvae model.
  - **vaeclass_kwargs**: dictionary of arguments for "vaeclass".
- **MtMcvae**: extension of Mcvae to allow training of multi-channel data with missing observations, as in [Antelmi, 2021]. This class inherits all the properties of the Mcvae class and extend it with the following argument:
  - **ids**: a list of observation identifiers for each channel. It is used internally to batch together observations with the same id, necessary to properly update the model parameters during training.
- **VAE**: basic class for building the Mcvae model (Fig. **??**).
- **ConditionalDistributionNet**: basic class for building the VAE model (Fig. **??**).

5.4.2 Utilities

The MCVAE package comes with utilities that help the user in managing his/her models.

Fit, load, save (`mcvae/models/utils`)

- **update_model**: to load/copy the model parameters from disk.
- **save_model**: to save a trained model to disk.
- **load_or_fit**: equivalent to **update_model** if the model has been saved in a previous training session. If not, this utility trains the model and saves it to disk. It uses a context manager [3] that prevents the model to be trained if there is a job already in place to fit the model. This is useful to avoid conflicts between jobs when heavy experimental campaigns need to be run. Arguments:
  - **model**: model to optimize.
  - **data**: training data. It can be also a PyTorch DataLoader for mini-batch training.
  - **epochs**: number of training epochs.
  - **ptfile**: path to *.pt file where to save the trained model.
  - **minibatch**: True if training with mini-batches (default False).
  - **force_fit**: force the training even if the model is already trained.
- **load_data_from_spreadsheet**: utility to load multi-channel observations, with or without missing data, from a spreadsheet. The spreadsheet should contain one sheet per channel. The observation identifier is assumed to be in the first column of every sheet.

Diagnostic utilities (`mcvae/diagnistics`)

- **plot_loss**: to check the convergence of the training operation (Fig. 5.3).

[3] `https://book.pythontips.com/en/latest/context_managers.html`

- **plot_latent_space**: to visualize the projection of multi-channel data into the latent space (Fig. 5.4).

- **plot_dropout**: to easily check which latent dimensions have been dropped out in the sparse MCVAE model (Fig. 5.5).

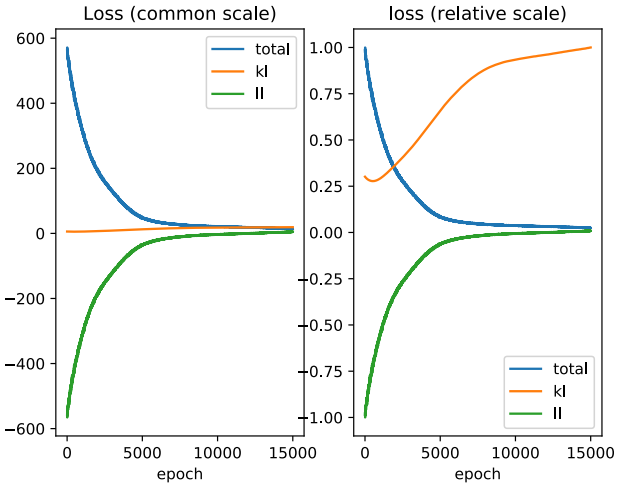


Figure 5.3: Losses plotted in absolute and relative scale with the `plot_loss` utility (ll: log-likelihood, kl: Kullback-Leibler divergence; total: kl - ll).

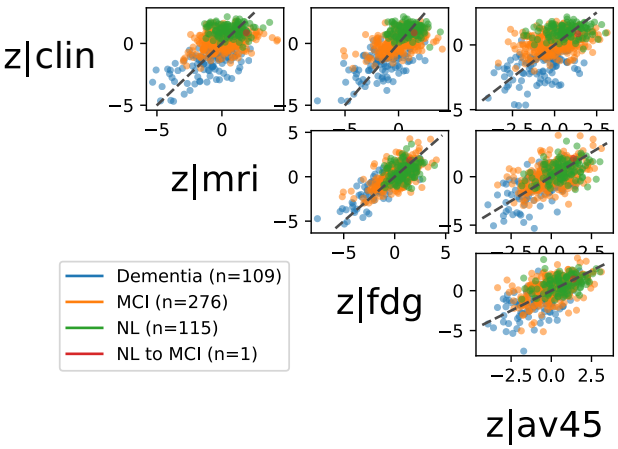


Figure 5.4: Multi-channel observations projected in one selected latent dimension z_i with the `plot_latent_space` utility. The utility can optionally take a grouping variable to highlight clusters of points (diagnosis in this figure).

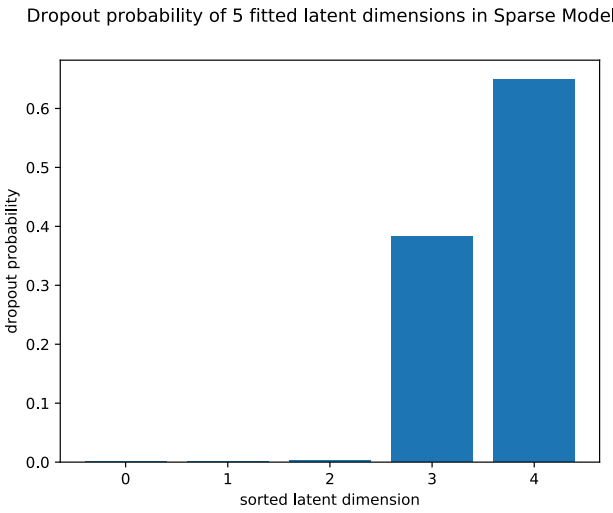


Figure 5.5: Dropout probability of a "sparse" model plotted with the `plot_dropout` utility.

6 Conclusion

The works illustrated in this Thesis show how the combination of: 1) generative modeling, 2) variational inference, and 3) state of the art machine learning methods, can have a positive impact in solving complex data modeling scenarios, such as when dealing with heterogeneous, high-dimensional observations for multi-modal feature prediction or classification tasks.

Generative modeling is a powerful designing procedure that we adopted it in Chapter 2 and Ch. 3 to posit a latent variable $\mathbf{z}$ as the single source of all the heterogeneous observations collected in neuroimaging datasets for dementia studies. The parallelism with respect to what happen in reality is clear, as usually there is a patient undergoing various medical exams, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) Imaging, who can be considered as the single source of the medical results.

In the following sections, we summarize the methodological contributions of this Thesis as well as the obtained results. We also propose new applications for our methods and build upon their limitations to propose research perspectives for the field of the joint modeling of heterogeneous data, and we conclude the Thesis with final remarks.

6.1 Summary of the main contributions

Multi-Channel Variational Autoencoder (MCVAE)

In Chapter 2 we introduced a novel latent variable framework to jointly model complex heterogeneous observations. To do so we posit a single latent variable as the source of the variability observed in the data, and we applied Variational Inference and modern learning algorithm to infer that latent source. The joint modeling is promoted in the latent space by constraining the inferred distribution of each data modality to a common target prior. We argue that our framework can be of interest for the neuroimaging community as it can be adopted to model the joint relationship between multi-modal neuroimaging data, such as those ones regularly collected in research dementia studies. Indeed, in this context of high heterogeneity due to the presence of, among many others, Magnetic Resonance Imaging (MRI) data and Positron Emission Tomography (PET) imaging data,

that is channels with their own informative content, there is a rational need for methods to establish relationships between observations.

Main contributions

- We proposed a novel methodology for the modeling of multi-modal heterogeneous data, when these data share a common origin. It is particularly suited for medical application as in this domain the patient can be considered as the common origin of all the data collected to infer his/her unknown diagnostic state.

- Interpretable latent representations are enforced by variational dropout, leveraging on sparsity to provide an effective mean to model selection in the latent space. 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.

- Thanks to its general formulation, the proposed method can be applied as a general data interpretation technique, not limited to the biomedical research area.

- The implementation of the method is open source and freely available online.

Multi-Task Multi-Channel Variational Autoencoder (MT-MCVAE)

In medical imaging data modeling, it is often necessary to increase the sample size by pooling together data from multiple datasets, which often introduce the problem of missing data and incompatibilities between datasets. In Chapter 3 we extended the capabilities of the MCVAE framework with a specific optimization scheme that allows the simultaneous learning from multiple datasets, without discarding any observation nor data modality. When a particular data modality is missing from a particular observation, for example if we miss the PET imaging data in the observations set of a patients that includes MRI and other clinical data, it will be simply not contribute to the leaning of the associated data modality parameters, without discarding all the other modalities which can still contribute to the learning of their associated parameters. The presence of at least one common data modality among datasets acts as a link across datasets and allows the information needed for the joint modeling of an MCVAE to flow through all the datasets to the other data modalities.

Main contributions

- We introduced a specific optimization scheme allowing the MCVAE to learn simultaneously from multiple datasets, even in the presence of missing data and non-compatible data modalities among all the datasets.

- The robustness to missing data allowed to reach the same joint modeling performances of the MCVAE with only the 25% of observations with no missing data.

- Extensive tests for the joint modeling of synthetically generated data and of real multi-modal neuroimaging datasets from independent dementia studies, showed the competitiveness of our method in classification and feature prediction tasks with respect to the state of the art methods.

- The implementation of the method is open source and freely available online.

6.2 Future developments

6.2.1 Disjoint representations

In the work developed in this Thesis, the concept of a shared latent space $\mathbf{z}$ between multi-modal observations plays a key role. Although very useful and powerful, it may still be not enough to capture the complexity of real world data.

An interesting perspective to pursue is to understand what are the peculiarities of each single modality, and how this information, that is not shared with other modalities, is structured. In a more practical sense, the question we may want to answer can be: given a specific medical context, what is the information content in a PET that cannot be predicted from an MRI? In order to answer this question we need to introduce new latent variables and new generative models. In Fig. 6.1 we compare the current generative model of the MCVAE and MT-MCVAE and possible extensions to take into account modality-specific variables.

More precisely, in Fig. 6.1b we propose a simple generative model with one extra latent $\mathbf{z}_c$ for each modality. A model taking into account the specific contributions of each modality, would probably have a better prediction performance as the latent space would host more information: the shared and the disjoint ones.

Going one step further the research question may become: what are the latent factors specific to sub-groups of modalities which are disjoint from all the others? In Fig. 6.1c we propose a generative model for answering this question. In this case we expect even better performances, not only because the latent space would be richer and more structured, but also because the prediction of an i-th modality would benefit simultaneously from the common latent $\mathbf{z}$ and the all the other specific and sub-specific latents $\{\mathbf{z}_{i,j}\}_{j=1}^{C}$.

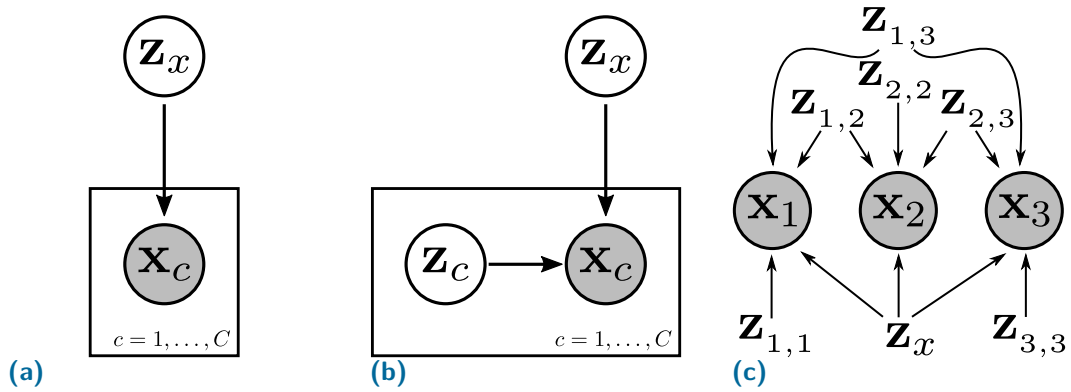


Figure 6.1: (a) Generative model of the Multi-Channel Variational Autoencoder (MCVAE), where a common latent $\mathbf{z}_x$ is the only source for the observations $\mathbf{x}_c$. (b), (c) Possible generative models of a MCVAE with modality specific latent variables, hosting a disentangled, complementary, and richer source of information.

6.2.2 Domain shift compensation

In Chapter 3 we showed that gathering observations from multiple datasets and modeling them with the MT-MCVAE results in better performances with respect to the MCVAE, which is limited to model one dataset at a time. The MT-MCVAE model is based on the assumption of identical distributed observations across datasets. This assumption may not be necessary true, especially in healthcare datasets where usually dataset releted biases exist. In Ch. 3 we mitigated these biases by harmonizing the datasets with ComBat [Johnson, 2007] before applying our model. Given these premises, another possible extension of our work would be to embed a domain-shift compensation mechanism to allow the modeling of multi-dataset observations when the datasets are not harmonized.

In [Ilse, 2020], the authors propose the Domain Invariant Variational Autoencoder (DIVA), a VAE based method to to learn unbiased representations given data from a set of datasets, and show that they are able to capture and correct for the biases of the different domains. In Fig. 6.2a we show the generative model of the DIVA, where samples $\mathbf{z}_d$ come from the prior $p(\mathbf{z}_d|d)$ conditioned on the domain label d. Samples $\mathbf{z}_x$ comes from a Standard Gaussian as in standard VAE and MCVAE. In Fig. 6.2b we propose a multi-channel extension of the DIVA, where we introduce new latent variables $\mathbf{z}_{d,c}$ (one for every channel) coming from the respective prior distributions $p(\mathbf{z}_{d,c}|d)$, conditioned on the domain label d as in DIVA.

Domain invariance and disjoint modeling can be also combined into a more general model, as the one depicted in Fig. 6.3.

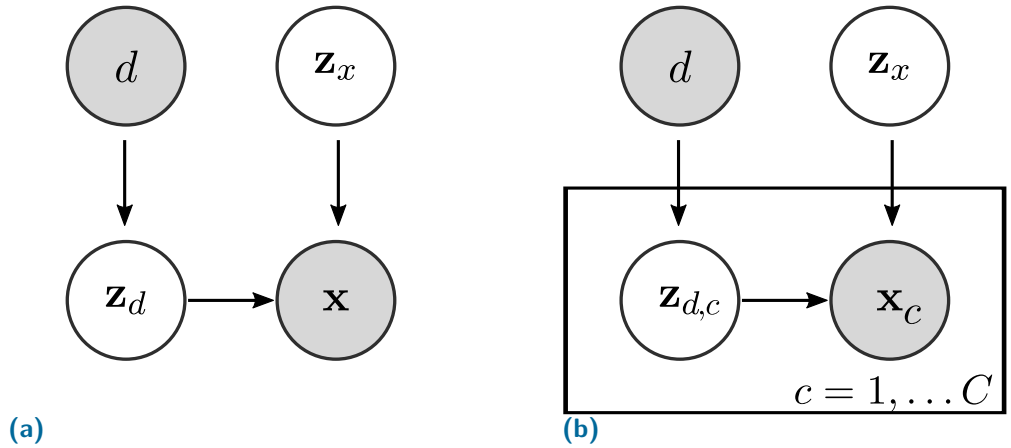


Figure 6.2: Generative models of: (a) Domain Invariant Variational Autoencoder (DIVA); (b) possible Multi-Channel extension of the DIVA. Samples $\mathbf{z}_{d,\cdot}$ come from the prior $p(\mathbf{z}_{d,c}|d)$ conditioned on the domain label d. Samples $\mathbf{z}_x$ comes from a Standard Gaussian as in VAE and MCVAE.

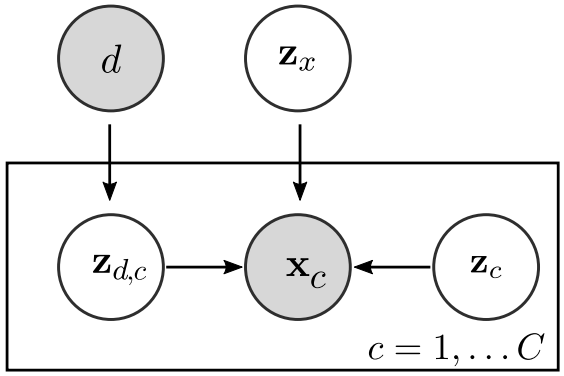


Figure 6.3: Generative model for a Domain Invariant Multi-Channel Variational Autoencoder with joint ($\mathbf{z}_x$) and disjoint ($\mathbf{z}_c$) latent spaces. It comes from the combination of the models proposed in Fig. 6.1b and Fig. 6.2b.

6.2.3 Temporal modeling

In Chapter 2 we showed that in the latent space we can disentangle the disease trajectory from the normal aging trajectory. This result emerged from the joint modeling of cross-sectional observations, without recurring to the concept of time evolution. For a better understanding of Alzheimer's Disease and other dementia related disorders, it is necessary to further disentangle sub-trajectories associated with different pathological subtypes. Indeed, this is a very active area of research [Lorenzi, 2015; Khanal, 2016; Khanal, 2017], By leveraging on the works already developed in the literature, specifically in modeling dynamic phenomena within the Variational Autoencoder framework [Girin, 2020] applied to longitudinal dementia studies [AbiNader, 2021], we propose to generalize our work by embedding it into a longitudinal framework.

6.2.4 Other applications

In our Thesis we focused particularly on the topic of data integration for neuroimaging studies in dementia, although the framework developed here for the joint modeling of heterogeneous data is applicable in other contexts, too.

For example, in the automotive industry cars are becoming more and more equipped with sensors to detect nearby vehicles [Liu, 2021], pedestrians [Held, 2021], road signs [Barodi, 2020], to generally enhance road safety. As these sensors are typically used to gather complementary information about the environment surrounding the car, their integration into a joint model like the one proposed in this Thesis could further increase the overall car safety. Indeed, since a possible sensor failure would result in missing or corrupted data, the compensation coming from all other sensors through the joint modeling would make the inference about the status of the car more robust. Moreover, if temporal dynamics are considered into the modeling, the timely prediction of an imminent danger could be used to adopt preventive safety measurements.

The same concept can be applied for home safety, for example to monitor the activities of people with compromised autonomy. If we assume to solve all the privacy issues beforehand, here the integration of data coming from the environment itself, such as video cameras or wireless location systems, together with wearable sensors, can be used to robustly detect situation of danger such as falls [Zigel, 2009], and as well for assessing the health status through the monitoring of performances during daily activity tasks [Suryadevara, 2012].

6.3 Final remarks

The constant development of mathematical and statistical tools makes it possible to build personalized models for patients, continuously adjustable based on measured health and lifestyle habits assessments. This can ultimately lead to a virtual patient, a digital twin [1], with detailed description of the state of an individual. With this Thesis we introduced new methodologies allowing to integrate and jointly analyze patients' data in a new and more interpretable manner, contributing to the building of a digital twin, for a better diagnosis, intervention simulation, and treatment, by adopting a data-driven and objective approach to healthcare.

[1] https://en.wikipedia.org/wiki/Digital_twin

Acronyms

A/T/N amyloid, tau, and neurodegeneration classification system.

AAL Automated Anatomical Labeling.

AD Alzheimer's Disease.

ADAS-Cog Alzheimer's Disease Assessment Scale - Cognitive Subscale.

ADNI Alzheimer's Disease Neuroimaging Initiative.

ApoE Apolipoprotein-E.

AV1451 ^{18}F-Flortaucipir.

AV45 ^{18}F-Florbetapir.

CCA Canonical Correlation Analysis.

CDR Clinical Dementia Rating Scale.

CSF cerebrospinal fluid.

CT Computed Tomography.

DAE Denoising Autoencoder.

DIVA Domain Invariant Variational Autoencoder.

DTI Diffusion Tensor Imaging.

ELBO Evidence Lower Bound.

EN EmbraceNet.

FAQ Functional Activities Questionnaire.

FDG ^{18}F-Fluorodeoxyglucose.

FDG-PET Fluorodeoxyglucose Positron Emission Tomography.

ICA Independent Component Analysis.

KNN k-Nearest Neighbors.

LDA Linear Discriminant Analysis.

MAE Mean Absolute Error.

MAR missing at random.

MCI Mild Cognitive Impairment.

MCVAE Multi-Channel Variational Autoencoder.

MICE Multivariate Imputation by Chained Equations.

MIRIAD Minimal Interval Resonance Imaging in Alzheimer's Disease.

MMSE Mini-Mental State Examination.

MNAR missing not at random.

MRI Magnetic Resonance Imaging.

MSE Mean Squared Error.

MT Multi-Task.

MT-MCVAE Multi-Task Multi-Channel Variational Autoencoder.

MTL Multi-Task Learning.

NC Normal Cognition.

NIA-AA National Institute on Aging and Alzheimer's Association.

NLL Negative Log-Likelihood.

OASIS Open Access Series of Imaging Studies.

OOP Object-oriented programming.

PCA Principal Component Analysis.

PET Positron Emission Tomography.

PLS Partial Least Squares.

RRR Reduced Rank Regression.

SNR Signal-to-Noise Ratio.

STEB Single Task with External Benchmark.

STIB Single Task with Internal Benchmark.

SUVR Relative Standardized Uptake Value.

VAE Variational Autoencoder.

VBM Voxel-based Morphometry.

VI Variational Inference.

Glossary

Alzheimer's Disease is defined by the presence in the brain of extracellular amyloid-β plaques and aggregates of hyperphosphorylated tau in neurofibrillary tangles, independently of the clinical expression of cognitive symptoms [Jack, 2018] .

Channel see **View**.

ComBat Harmonization method to remove domain bias when pooling data from different datasets [Johnson, 2007] .

CovBat Same as ComBat, with an additional step for covariance harmonization [Chen, 2019a] .

Dementia denotes an acquired, insidious, and progressive cognitive and functional impairment. Alzheimer's Disease (AD) is the most common cause of dementia and accounts for 60% to 80% of the cases [Alzheimer Association Report, 2020] .

Mild Cognitive Impairment refers to a population without, or with subtle, functional disability, but with an acquired objective cognitive impairment. Representing a clinical syndrome, it encompasses cases progressing to AD (about 50%) or non-AD dementia (about $10-15\%$) as well as stable cases (about $35-40\%$). MCI cases positive to AD biomarkers can be defined as prodromal AD or MCI due to AD based on research diagnostic criteria [Dubois, 2014] and consistently also with the 2018 A/T/N framework [Jack, 2018] .

Object-oriented programming is a programming paradigm based on the concept of "objects", which can contain data and code: data in the form of fields (often known as attributes or properties), and code, in the form of procedures (often known as methods) .

View a group of homogeneous features, such as measurements from a specific imaging modality, or clinical scores, or biological measurements, representing an important and independent source of information for the disease or phenomena under investigation .

Bibliography

[Abi Nader, 2020] Clément Abi Nader, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. "Monotonic Gaussian Process for spatio-temporal disease progression modeling in brain imaging data". In: *Neuroimage* 205 (Jan. 2020), p. 116266 (cit. on p. 33).

[AbiNader, 2021] Clément AbiNader, Nicholas Ayache, Giovanni B Frisoni, Philippe Robert, Marco Lorenzi, and Alzheimer's Disease Neuroimaging Initiative. "Simulating the outcome of amyloid treatments in Alzheimer's disease from imaging and clinical data". In: *Brain Commun.* (2021) (cit. on p. 96).

[Alemi, 2017] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. "Fixing a Broken ELBO". In: (Nov. 2017) (cit. on p. 13).

[Alzheimer Association Report, 2020] Alzheimer Association Report. "2020 Alzheimer's disease facts and figures". In: *Alzheimer's Dement.* 16.3 (Mar. 2020), pp. 391–460 (cit. on pp. 1, 103).

[Andrew, 2013a] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. "Deep Canonical Correlation Analysis". In: *Proc. Mach. Learn. Res.* 28.3 (2013), pp. 1247–1255 (cit. on pp. 4, 12).

[Andrew, 2013b] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. "Deep Canonical Correlation Analysis". In: *Proc. Mach. Learn. Res.* 28.3 (2013), pp. 1247–1255 (cit. on p. 42).

[Antelmi, 2018a] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. “Multi-channel Stochastic Variational Inference for the Joint Analysis of Heterogeneous Biomedical Data in Alzheimer’s Disease”. In: *Underst. Interpret. Mach. Learn. Med. Image Comput. Appl.* Ed. by Danail Stoyanov, Zeike Taylor, Seyed Mostafa Kia, Ipek Oguz, Mauricio Reyes, Anne Martel, Lena Maier-Hein, Andre F Marquand, Edouard Duchesnay, Tommy Löfstedt, Bennett Landman, M Jorge Cardoso, Carlos A Silva, Sergio Pereira, and Raphael Meier. Cham: Springer International Publishing, 2018, pp. 15–23 (cit. on pp. 9, 11, 86).

[Antelmi, 2018b] Luigi Antelmi, Marco Lorenzi, Valeria Manera, Philippe Robert, and Nicholas Ayache. *A method for statistical learning in large databases of heterogeneous imaging, cognitive and behavioral data.* EPICLIN 2018 - 12ème Conférence Francophone d’Epidémiologie Clinique / CLCC 2018 - 25èmes Journées des statisticiens des Centre de Lutte Contre le Cancer. Poster. May 2018 (cit. on pp. 9, 11).

[Antelmi, 2019] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. “Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data”. In: *Proc. 36th Int. Conf. Mach. Learn.* Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. PMLR, 2019, pp. 302–311 (cit. on pp. 9, 11, 34–36, 38, 41–43, 59, 61, 82, 86).

[Antelmi, 2021] Luigi Antelmi, Nicholas Ayache, Philippe Robert, Federica Ribaldi, Valentina Garibotto, Giovanni B Frisoni, and Marco Lorenzi. “Combining Multi-Task Learning and Multi-Channel Variational Auto-Encoders to Exploit Datasets with Missing Observations -Application to Multi-Modal Neuroimaging Studies in Dementia”. under review at NeuroImage. Jan. 2021 (cit. on pp. 9, 31, 86).

[Apostolova, 2010] Liana G. Apostolova, Kristy S. Hwang, John P. Andrawis, Amity E. Green, Sona Babakchanian, Jonathan H. Morra, Jeffrey L. Cummings, Arthur W. Toga, John Q. Trojanowski, Leslie M. Shaw, Clifford R. Jack, Ronald C. Petersen, Paul S. Aisen, William J. Jagust, Robert A. Koeppe, Chester A. Mathis, Michael W. Weiner, and Paul M. Thompson. “3D PIB and CSF biomarker associations with hippocampal atrophy in ADNI subjects”. In: *Neurobiol. Aging* 31.8 (Aug. 2010), pp. 1284–1303 (cit. on p. 3).

[Ashburner, 2000] John Ashburner and Karl J. Friston. “Voxel-based morphometry–the methods.” In: *Neuroimage* 11.6 Pt 1 (June 2000), pp. 805–21 (cit. on pp. 24, 48).

[Barodi, 2020] Anass Barodi, Abdrrahim Bajit, Mohammed Benbrahim, and Ahmed Tamtaoui. “An Enhanced Approach in Detecting Object Applied to Automotive Traffic Roads Signs”. In: *2020 IEEE 6th International Conference on Optimization and Applications (ICOA)*. IEEE. 2020, pp. 1–6 (cit. on p. 96).

[Barthel, 2020] Henryk Barthel. “First Tau PET Tracer Approved: Toward Accurate In Vivo Diagnosis of Alzheimer Disease”. In: *J. Nucl. Med.* 61.10 (Oct. 2020), pp. 1409–1410 (cit. on p. 2).

[Blei, 2016] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: (2016). eprint: 1601.00670 (cit. on p. 15).

[Blei, 2017] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: *J. Am. Stat. Assoc.* 112.518 (Apr. 2017), pp. 859–877 (cit. on pp. 36, 61).

[Boccardi, 2021] Marina Boccardi, Alessandra Dodich, Emiliano Albanese, Angèle Gayet-Ageron, Cristina Festari, Nicholas J. Ashton, Gérard N. Bischof, Konstantinos Chiotis, Antoine Leuzy, Emma E. Wolters, Martin A. Walter, Gil D. Rabinovici, Maria Carrillo, Alexander Drzezga, Oskar Hansson, Agneta Nordberg, Rik Ossenkoppele, Victor L. Villemagne, Bengt Winblad, Giovanni B. Frisoni, and Valentina Garibotto. “The strategic biomarker roadmap for the validation of Alzheimer’s diagnostic biomarkers: methodological update”. In: *Eur. J. Nucl. Med. Mol. Imaging* (Mar. 2021) (cit. on p. 2).

[Buch, 2020] Amanda M. Buch and Conor Liston. “Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics”. In: *Neuropsychopharmacology* (Aug. 2020) (cit. on p. 32).

[Burda, 2015] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. “Importance Weighted Autoencoders”. In: (Sept. 2015). arXiv: 1509.00519 (cit. on p. 13).

[Buuren, 2000] S. van Buuren and C. G. M. Groothuis-Oudshoorn. *Multivariate Imputation by Chained Equations: MICE V1.0 User manual*. Vol. PG/VGZ/00.038. Leiden: TNO Prevention and Health, 2000 (cit. on p. 44).

[Caruana, 1998] Rich Caruana. “Multitask Learning”. In: *Learn. to Learn*. Ed. by Sebastian Thrun and Lorien Pratt. Boston, MA: Springer US, 1998, pp. 95–133 (cit. on p. 34).

[Chartsias, 2021] A Chartsias, G Papanastasiou, C Wang, S Semple, D E Newby, R Dharmakumar, and S A Tsaftaris. “Disentangle, Align and Fuse for Multimodal and Semi-Supervised Image Segmentation”. In: *IEEE Trans. Med. Imaging* 40.3 (2021), pp. 781–792 (cit. on p. 33).

[Chen, 2019a] Andrew A Chen, Joanne C Beer, Nicholas J Tustison, Philip A Cook, Russell T Shinohara, and Haochang Shou. “Removal of Scanner Effects in Covariance Improves Multivariate Pattern Analysis in Neuroimaging Data”. In: *bioRxiv* (2019) (cit. on pp. 64, 65, 67, 103).

[Chen, 2019b] Tingting Chen, Xinjun Ma, Xuechen Liu, Wenzhe Wang, Ruiwei Feng, Jintai Chen, Chunnv Yuan, Weiguo Lu, Danny Z. Chen, and Jian Wu. “Multi-view Learning with Feature Level Fusion for Cervical Dysplasia Diagnosis”. In: 2019, pp. 329–338 (cit. on p. 33).

[Chételat, 2021] Gaël Chételat, Javier Arbizu, Henryk Barthel, Valentina Garibotto, Adriaan A. Lammertsma, Ian Law, Silvia Morbelli, Elsmarieke van de Giessen, and Alexander Drzezga. “Finding our way through the labyrinth of dementia biomarkers”. In: *Eur. J. Nucl. Med. Mol. Imaging* (Apr. 2021) (cit. on p. 3).

[Choi, 2019] Jun-Ho Choi and Jong-Seok Lee. “EmbraceNet: A robust deep learning architecture for multimodal classification”. In: *Inf. Fusion* 51 (Nov. 2019), pp. 259–270 (cit. on pp. 6, 34, 56, 57, 59).

[Clark, 2011] Christopher M Clark, Julie A Schneider, Barry J Bedell, Thomas G Beach, Warren B Bilker, Mark A Mintun, Michael J Pontecorvo, Franz Hefti, Alan P Carpenter, Matthew L Flitter, Michael J Krautkramer, Hank F Kung, R Edward Coleman, P Murali Doraiswamy, Adam S Fleisher, Marwan N Sabbagh, Carl H Sadowsky, Eric M Reiman, Simone P Zehntner, Daniel M Skovronsky, and for the AV45-A07 Study Group. “Use of Florbetapir-PET for Imaging β-Amyloid Pathology”. In: *JAMA* 305.3 (2011), pp. 275–283 (cit. on p. 2).

[DaAno, 2020] R Da-Ano, D Visvikis, and M Hatt. “Harmonization strategies for multicenter radiomics investigations”. In: *Phys. Med. Biol.* 65.24 (Dec. 2020), 24TR02 (cit. on pp. 64, 65).

[Desikan, 2006] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest”. In: *Neuroimage* (2006) (cit. on p. 48).

[DoradoMoreno, 2020] M. Dorado-Moreno, N. Navarin, P.A. Gutiérrez, L. Prieto, A. Sperduti, S. Salcedo-Sanz, and C. Hervás-Martínez. "Multi-task learning for the prediction of wind power ramp events with deep neural networks". In: *Neural Networks* 123 (Mar. 2020), pp. 401–411 (cit. on pp. 6, 34).

[Dubois, 2014] Bruno Dubois, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, Stefano Cappa, Sebastian Crutch, Sebastiaan Engelborghs, Giovanni B Frisoni, Nick C Fox, Douglas Galasko, Marie-odile Habert, Gregory A Jicha, Agneta Nordberg, Florence Pasquier, Gil Rabinovici, Philippe Robert, Christopher Rowe, Stephen Salloway, Marie Sarazin, Stéphane Epelbaum, Leonardo C de Souza, Bruno Vellas, Pieter J Visser, Lon Schneider, Yaakov Stern, Philip Scheltens, and Jeffrey L Cummings. "Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria." In: *Lancet. Neurol.* 13.6 (June 2014), pp. 614–29 (cit. on pp. 25, 103).

[Finn, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *PMLR* 70 (Mar. 2017), pp. 1126–1135 (cit. on p. 34).

[Fortin, 2017] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A. Elliott, Kosha Ruparel, David R. Roalf, Theodore D. Satterthwaite, Ruben C. Gur, Raquel E. Gur, Robert T. Schultz, Ragini Verma, and Russell T. Shinohara. "Harmonization of multi-site diffusion tensor imaging data". In: *Neuroimage* 161 (Nov. 2017), pp. 149–170 (cit. on pp. 49, 59, 63, 66, 71).

[Fortin, 2018] Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. "Harmonization of cortical thickness measurements across scanners and sites". In: *Neuroimage* 167 (Feb. 2018), pp. 104–120 (cit. on pp. 49, 59, 63, 66).

[Fostinelli, 2020] Silvia Fostinelli, Ramona De Amicis, Alessandro Leone, Valentina Giustizieri, Giuliano Binetti, Simona Bertoli, Alberto Battezzati, and Stefano F Cappa. "Eating Behavior in Aging and Dementia: The Need for a Comprehensive Assessment". In: *Front. Nutr.* 7 (Dec. 2020) (cit. on pp. 2, 64).

[Fox, 2004] Nick C Fox and Jonathan M Schott. "Imaging cerebral atrophy: normal ageing to Alzheimer's disease". In: *Lancet* 363.9406 (Jan. 2004), pp. 392–394 (cit. on p. 2).

[Frisoni, 2003] Giovanni B. Frisoni, Alessandro Padovani, and Lars-Olof Wahlund. “The Diagnosis of Alzheimer Disease Before It Is Alzheimer Dementia”. In: *Arch. Neurol.* 60.7 (July 2003), p. 1023 (cit. on pp. 2, 64).

[Garbarino, 2021] Sara Garbarino and Marco Lorenzi. “Investigating hypotheses of neurodegeneration by learning dynamical systems of protein propagation in the brain”. In: *Neuroimage* (2021), p. 117980 (cit. on pp. 41, 82).

[Garibotto, 2017] Valentina Garibotto, Karl Herholz, Marina Boccardi, Agnese Picco, Andrea Varrone, Agneta Nordberg, Flavio Nobili, and Osman Ratib. “Clinical validity of brain fluorodeoxyglucose positron emission tomography as a biomarker for Alzheimer’s disease in the context of a structured 5-phase development framework”. In: *Neurobiol. Aging* 52 (Apr. 2017), pp. 183–195 (cit. on p. 2).

[Girin, 2020] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. “Dynamical Variational Autoencoders: A Comprehensive Review”. In: (Aug. 2020). arXiv: 2008.12595 (cit. on p. 96).

[Golriz Khatami, 2020] Sepehr Golriz Khatami, Christine Robinson, Colin Birkenbihl, Daniel Domingo-Fernández, Charles Tapley Hoyt, and Martin Hofmann-Apitius. “Challenges of Integrative Disease Modeling in Alzheimer’s Disease”. In: *Front. Mol. Biosci.* 6 (Jan. 2020) (cit. on p. 32).

[Gondara, 2018] Lovedeep Gondara and Ke Wang. “MIDA: Multiple Imputation Using Denoising Autoencoders”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi. Cham: Springer International Publishing, 2018, pp. 260–272 (cit. on pp. 6, 34, 44, 45).

[Gupta, 2019] Yubraj Gupta, Ramesh Kumar Lama, and Goo-Rak Kwon. “Prediction and Classification of Alzheimer’s Disease Based on Combined Features From Apolipoprotein-E Genotype, Cerebrospinal Fluid, MR, and FDG-PET Imaging Biomarkers”. In: *Front. Comput. Neurosci.* 13 (Oct. 2019) (cit. on p. 3).

[Guyon, 2003] Isabelle Guyon. “Design of experiments for the NIPS 2003 variable selection benchmark”. In: *NIPS 2003 workshop on feature extraction and feature selection* (2003) (cit. on p. 71).

[Haufe, 2014] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. “On the interpretation of weight vectors of linear models in multivariate neuroimaging.” In: *Neuroimage* 87 (2014), pp. 96–110 (cit. on pp. 4, 12).

[Held, 2021] Patrick Held, Dagmar Steinhauser, Andreas Koch, Thomas Brandmeier, and Ulrich T. Schwarz. “A Novel Approach for Model-Based Pedestrian Tracking Using Automotive Radar”. In: *IEEE Transactions on Intelligent Transportation Systems* (2021), pp. 1–14 (cit. on p. 96).

[Herholz, 2012] Karl Herholz. “Use of FDG PET as an imaging biomarker in clinical trials of Alzheimer’s disease”. In: *Biomark. Med.* 6.4 (Aug. 2012), pp. 431–439 (cit. on p. 2).

[Higgins, 2018] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. “Towards a Definition of Disentangled Representations”. In: (Dec. 2018). arXiv: 1812.02230 (cit. on p. 5).

[Hippius, 2003] Hanns Hippius and Gabriele Neundörfer. “The discovery of Alzheimer’s disease.” In: *Dialogues Clin. Neurosci.* 5.1 (Mar. 2003), pp. 101–8 (cit. on p. 1).

[Hotelling, 1936] Harold Hotelling. “Relations Between Two Sets of Variates”. In: *Biometrika* 28.3/4 (Dec. 1936), p. 321 (cit. on p. 12).

[Huang, 2009] Su-Yun Huang, Mei-Hsien Lee, and Chuhsing Kate Hsiao. “Nonlinear measures of association with kernel canonical correlation analysis and applications”. In: *J. Stat. Plan. Inference* 139.7 (2009), pp. 2162–2174 (cit. on pp. 4, 12).

[Ilse, 2020] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. “DIVA: Domain Invariant Variational Autoencoders”. In: ed. by Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal. Vol. 121. Proceedings of Machine Learning Research. Montreal, QC, Canada: PMLR, 2020, pp. 322–348 (cit. on pp. 64, 67, 94).

[Jack, 2018] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. “NIA-AA Research Framework: Toward a biological definition of Alzheimer’s disease”. In: *Alzheimer’s Dement.* 14.4 (2018), pp. 535–562 (cit. on pp. 2, 6, 25, 103).

[Johnson, 2007] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (Jan. 2007), pp. 118–127 (cit. on pp. 49, 59, 63–66, 94, 103).

[Jovicich, 2019] Jorge Jovicich, Frederik Barkhof, Claudio Babiloni, Karl Herholz, Christoph Mulert, Bart N.M. Berckel, and Giovanni B. Frisoni. "Harmonization of neuroimaging biomarkers for neurodegenerative diseases: A survey in the imaging community of perceived barriers and suggested actions". In: *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* 11.1 (Dec. 2019). Ed. by Jorge Jovicich and Giovanni B. Frisoni, pp. 69–73 (cit. on p. 65).

[Kettenring, 1971] J. R. Kettenring. "Canonical analysis of several sets of variables". In: *Biometrika* 58.3 (1971), pp. 433–451 (cit. on pp. 4, 12).

[Khan, 2020] Muhammad Attique Khan, Imran Ashraf, Majed Alhaisoni, Robertas Damaševičius, Rafal Scherer, Amjad Rehman, and Syed Ahmad Chan Bukhari. "Multimodal Brain Tumor Classification Using Deep Learning and Robust Feature Selection: A Machine Learning Application for Radiologists". In: *Diagnostics* 10.8 (Aug. 2020), p. 565 (cit. on p. 7).

[Khanal, 2016] Bishesh Khanal, Marco Lorenzi, Nicholas Ayache, and Xavier Pennec. "A biophysical model of brain deformation to simulate and analyze longitudinal MRIs of patients with Alzheimer's disease". In: *Neuroimage* 134 (July 2016), pp. 35–52 (cit. on p. 96).

[Khanal, 2017] Bishesh Khanal, Nicholas Ayache, and Xavier Pennec. "Simulating Longitudinal Brain MRIs with Known Volume Changes and Realistic Variations in Image Intensity." In: *Front. Neurosci.* 11 (2017), p. 132 (cit. on p. 96).

[Kim, 2018] Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. "Semi-Amortized Variational Autoencoders". In: (Feb. 2018). arXiv: 1802.02550 (cit. on p. 13).

[Kingma, 2014a] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6 (2014). arXiv: 1412.6980 (cit. on pp. 17, 73).

[Kingma, 2014b] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014)*. Dec. 2014. eprint: 1312.6114 (cit. on pp. 4, 13, 17, 38, 43, 67).

[Kingma, 2015] Diederik P Kingma, Tim Salimans, and Max Welling. "Variational Dropout and the Local Reparameterization Trick". In: *Adv. Neural Inf. Process. Syst. 28*. Ed. by C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett. Curran Associates, Inc., 2015, pp. 2575–2583 (cit. on pp. 13, 19, 20, 41, 82).

[Klami, 2007] Arto Klami and Samuel Kaski. “Local Dependent Components”. In: *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*. Ed. by Zoubin Ghahramani. Omnipress, 2007, pp. 425–432 (cit. on p. 13).

[Klami, 2013] Arto Klami, Virtanen Seppo, and Samuel Kaski. “Bayesian Canonical Correlation Analysis”. In: *J. Mach. Learn. Res.* 14 (2013), pp. 965–1003 (cit. on pp. 4, 12, 17).

[König, 2015] Alexandra König, Carlos Fernando Crispim Junior, Alexandre Derreumaux, Gregory Bensadoun, Pierre-David Petit, François Bremond, Renaud David, Frans Verhey, Pauline Aalten, and Philippe Robert. “Validation of an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients.” In: *J. Alzheimers. Dis.* 44.2 (2015), pp. 675–85 (cit. on p. 7).

[LaMontagne, 2019] Pamela J LaMontagne, Tammie L.S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. “OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease”. In: *medRxiv* (2019) (cit. on pp. 3, 47).

[Le Sueur, 2020] Helen Le Sueur, Ian N. Bruce, and Nophar Geifman. “The challenges in data integration – heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus”. In: *BMC Med. Res. Methodol.* 20.1 (Dec. 2020), p. 164 (cit. on pp. 32, 64).

[LeCun, 2010] Yann LeCun, Corinna Cortes, and CJ Burges. “MNIST handwritten digit database”. In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010) (cit. on p. 42).

[Li, 2020] Xiangtao Li, Shaochuan Li, Yunhe Wang, Shixiong Zhang, and Ka-Chun Wong. “Identification of pan-cancer Ras pathway activation with deep learning”. In: *Brief. Bioinform.* (Oct. 2020) (cit. on p. 7).

[Liu, 2014] Jingyu Liu and Vince D. Calhoun. “A review of multivariate analyses in imaging genetics.” In: *Front. Neuroinform.* 8 (Mar. 2014), p. 29 (cit. on pp. 4, 12, 33).

[Liu, 2021] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. “Multimodal Motion Prediction with Stacked Transformers”. In: (Mar. 2021). arXiv: 2103.11624 (cit. on p. 96).

[LopezMartin, 2017] Manuel Lopez-Martin, Belen Carro, Antonio Sanchez-Esguevillas, and Jaime Lloret. “Conditional Variational Autoencoder for Prediction and Feature Recovery Applied to Intrusion Detection in IoT.” In: *Sensors (Basel)*. 17.9 (Aug. 2017) (cit. on p. 5).

[Lorenzi, 2015] Marco Lorenzi, Xavier Pennec, Giovanni B. Frisoni, and Nicholas Ayache. “Disentangling normal aging from Alzheimer’s disease in structural magnetic resonance images”. In: *Neurobiol. Aging* 36 (Jan. 2015), S42–S52 (cit. on pp. 25, 96).

[Lucas, 2019] James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. “Understanding posterior collapse in generative latent variable models”. In: (2019) (cit. on p. 5).

[Luo, 2015] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. “Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction”. In: *IEEE Trans. Knowl. Data Eng.* 27.11 (2015), pp. 3111–3124 (cit. on pp. 4, 12).

[Malone, 2013] Ian B. Malone, David Cash, Gerard R. Ridgway, David G. MacManus, Sebastien Ourselin, Nick C. Fox, and Jonathan M. Schott. “MIRIAD—Public release of a multiple time point Alzheimer’s MR imaging dataset”. In: *Neuroimage* 70 (Apr. 2013), pp. 33–36 (cit. on pp. 3, 47).

[Manera, 2020] Valeria Manera, Sharon Abrahams, Luis Agüera-Ortiz, François Bremond, Renaud David, Kaci Fairchild, Auriane Gros, Cécile Hanon, Masud Husain, Alexandra König, Patricia L. Lockwood, Maribel Pino, Ratko Radakovic, Gabriel Robert, Andrea Slachevsky, Florindo Stella, Anaïs Tribouillard, Pietro Davide Trimarchi, Frans Verhey, Jerome Yesavage, Radia Zeghari, and Philippe Robert. “Recommendations for the Nonpharmacological Treatment of Apathy in Brain Disorders”. In: *Am. J. Geriatr. Psychiatry* 28.4 (Apr. 2020), pp. 410–420 (cit. on p. 7).

[McKhann, 2011] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. “The diagnosis of dementia due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease”. In: *Alzheimer’s Dement.* 7.3 (May 2011), pp. 263–269 (cit. on p. 2).

[Méndez, 2015] C. Andrés Méndez, Paul Summers, and Gloria Menegaz. "Multiview cluster ensembles for multimodal MRI segmentation". In: *Int. J. Imaging Syst. Technol.* 25.1 (Mar. 2015), pp. 56–67 (cit. on p. 7).

[Miotto, 2018] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. "Deep learning for healthcare: review, opportunities and challenges". In: *Brief. Bioinform.* 19.6 (Nov. 2018), pp. 1236–1246 (cit. on p. 79).

[Mita, 2020] Graziano Mita, Maurizio Filippone, and Pietro Michiardi. "An Identifiable Double VAE For Disentangled Representations". In: (Oct. 2020). arXiv: 2010.09360 (cit. on p. 5).

[Molchanov, 2017] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. "Variational Dropout Sparsifies Deep Neural Networks". In: *arXiv* (2017). arXiv: 1701.05369 (cit. on pp. 13, 19, 41, 82).

[Nathoo, 2019] Farouk S. Nathoo, Linglong Kong, and Hongtu Zhu. "A review of statistical methods in imaging genetics". In: *Can. J. Stat.* 47.1 (Mar. 2019), pp. 108–131 (cit. on p. 4).

[Orlhac, 2020] Fanny Orlhac, Augustin Lecler, Julien Savatovski, Jessica Goya-Outi, Christophe Nioche, Frédérique Charbonneau, Nicholas Ayache, Frédérique Frouin, Loïc Duron, and Irène Buvat. "How can we combat multicenter variability in MR radiomics? Validation of a correction procedure." In: *Eur. Radiol.* (Sept. 2020) (cit. on pp. 49, 59, 63, 66).

[Palesi, 2019] Fulvia Palesi, Anna Nigri, Domenico Aquino, Ruben Gianeri, Alice Pirastru, Marcella Laganà, Laura Biagi, Maria Grazia Bruzzone, and Claudia A M Gandini Wheelerkingshott. "MRI quality data assessment in the Italian IRCCS advanced neuroimaging network using ACR phantoms". In: *Proc. ISMRM 27th Annu. Meet. Exhib.* 2019, p. 4514 (cit. on p. 65).

[Paszke, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. "Automatic differentiation in PyTorch". In: (2017) (cit. on p. 14).

[Paszke, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., 2019, pp. 8024–8035 (cit. on pp. 36, 79, 80).

[Pedregosa, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 44, 71).

[Platias, 2020] Christos Platias and Georgios Petasis. "A Comparison of Machine Learning Methods for Data Imputation". In: *11th Hell. Conf. Artif. Intell.* ACM, Sept. 2020, pp. 150–159 (cit. on p. 58).

[Qin, 2019] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. "Unsupervised Deformable Registration for Multi-modal Images via Disentangled Representations". In: *Inf. Process. Med. Imaging*. Ed. by Albert C S Chung, James C Gee, Paul A Yushkevich, and Siqi Bao. Cham: Springer International Publishing, 2019, pp. 249–261 (cit. on p. 33).

[Reuter, 2012] Martin Reuter, Nicholas J. Schmansky, Herminia Diana Rosas, and Bruce Fischl. "Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis". In: *NeuroImage* 61.4 (2012), pp. 1402–1418 (cit. on p. 48).

[Rezende, 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: (Jan. 2014). arXiv: 1401.4082 (cit. on pp. 4, 13, 17, 38, 43, 67).

[Robert, 2014] Philippe H. Robert, Alexandra KÃ¶nig, HÃ©lene Amieva, Sandrine Andrieu, FranÃ§ois Bremond, Roger Bullock, Mathieu Ceccaldi, Bruno Dubois, Serge Gauthier, Paul-Ariel Kenigsberg, StÃ©phane Nave, Jean M. Orgogozo, Julie Piano, Michel Benoit, Jacques Touchon, Bruno Vellas, Jerome Yesavage, and Valeria Manera. "Recommendations for the use of Serious Games in people with Alzheimer's Disease, related disorders and frailty". In: *Front. Aging Neurosci.* 6 (Mar. 2014) (cit. on p. 7).

[Rolinek, 2019] Michal Rolinek, Dominik Zietlow, and Georg Martius. "Variational Autoencoders Pursue PCA Directions (by Accident)". In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* June 2019 (cit. on p. 5).

[Rossi, 2019] Simone Rossi, Pietro Michiardi, and Maurizio Filippone. "Good Initializations of Variational {B}ayes for Deep Models". In: *Proc. 36th Int. Conf. Mach. Learn.* Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5487–5497 (cit. on p. 5).

[Scarmeas, 2007] Nikolaos Scarmeas, Jason Brandt, Deborah Blacker, Marilyn Albert, Georgios Hadjigeorgiou, Bruno Dubois, Davangere Devanand, Lawrence Honig, and Yaakov Stern. "Disruptive Behavior as a Predictor in Alzheimer Disease". In: *Arch. Neurol.* 64.12 (Dec. 2007), p. 1755 (cit. on pp. 2, 64).

[Sezer, 2020] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. "Financial time series forecasting with deep learning : A systematic literature review: 2005–2019". In: *Appl. Soft Comput.* 90 (May 2020), p. 106181 (cit. on p. 79).

[Shaffer, 2013] Jennifer L. Shaffer, Jeffrey R. Petrella, Forrest C. Sheldon, Kingshuk Roy Choudhury, Vince D. Calhoun, R. Edward Coleman, and P. Murali Doraiswamy. "Predicting Cognitive Decline in Subjects at Risk for Alzheimer Disease by Using Combined Cerebrospinal Fluid, MR Imaging, and PET Biomarkers". In: *Radiology* 266.2 (Feb. 2013), pp. 583–591 (cit. on p. 3).

[Shi, 2019] Yuge Shi, Siddharth N, Brooks Paige, and Philip Torr. "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". In: *Adv. Neural Inf. Process. Syst. 32*. Ed. by H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett. Curran Associates, Inc., 2019, pp. 15718–15729 (cit. on p. 34).

[Sivera, 2019] Raphaël Sivera, Hervé Delingette, Marco Lorenzi, Xavier Pennec, and Nicholas Ayache. "A model of brain morphological changes related to aging and Alzheimer's disease from cross-sectional assessments". In: *Neuroimage* 198 (Sept. 2019), pp. 255–270 (cit. on p. 25).

[Solomon, 2011] Alina Solomon, Letitia Dobranici, Ingemar Kåreholt, Cătălina Tudose, and Mircea Lăzărescu. "Comorbidity and the rate of cognitive decline in patients with Alzheimer dementia". In: *Int. J. Geriatr. Psychiatry* 26.12 (Dec. 2011), pp. 1244–1251 (cit. on pp. 2, 64).

[Srivastava, 2014] Nitish Srivastava, Hinton Geoffrey, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *J. Mach. Learn. Res.* 15 (2014), pp. 1929–1958 (cit. on p. 18).

[Suryadevara, 2012] N.K. Suryadevara, A. Gaddam, R.K. Rayudu, and S.C. Mukhopadhyay. “Wireless sensors network based safe home to care elderly people: Behaviour detection”. In: *Sensors Actuators A Phys.* 186 (Oct. 2012), pp. 277–283 (cit. on p. 96).

[Tabarestani, 2020] Solale Tabarestani, Maryamossadat Aghili, Mohammad Eslami, Mercedes Cabrerizo, Armando Barreto, Naphtali Rishe, Rosie E. Curiel, David Loewenstein, Ranjan Duara, and Malek Adjouadi. “A distributed multitask multimodal approach for the prediction of Alzheimer’s disease in a longitudinal study”. In: *Neuroimage* 206 (Feb. 2020), p. 116317 (cit. on p. 33).

[Tognin, 2020] Stefania Tognin, Hendrika H van Hell, Kate Merritt, Inge Winter-van Rossum, Matthijs G Bossong, Matthew J Kempton, Gemma Modinos, Paolo Fusar-Poli, Andrea Mechelli, Paola Dazzan, et al. “Towards Precision Medicine in Psychosis: Benefits and Challenges of Multimodal Multicenter Studies—PSYSCAN: Translating Neuroimaging Findings From Research into Clinical Practice”. In: *Schizophr. Bull.* 46.2 (Feb. 2020), pp. 432–441 (cit. on p. 32).

[TzourioMazoyer, 2002] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. “Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain”. In: *Neuroimage* 15.1 (Jan. 2002), pp. 273–289 (cit. on p. 24).

[Uppu, 2018] Suneetha Uppu, Aneesh Krishna, and Raj P. Gopalan. “A Review on Methods for Detecting SNP Interactions in High-Dimensional Genomic Data”. In: *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 15.2 (Mar. 2018), pp. 599–612 (cit. on p. 79).

[Valieris, 2020] Renan Valieris, Lucas Amaro, Cynthia Aparecida Bueno de Toledo Osório, Adriana Passos Bueno, Rafael Andres Rosales Mitrowsky, Dirce Maria Carraro, Diana Noronha Nunes, Emmanuel Dias-Neto, and Israel Tojal da Silva. “Deep Learning Predicts Underlying Features on Pathology Images with Therapeutic Relevance for Breast and Gastric Cancer”. In: *Cancers* 12.12 (2020) (cit. on p. 7).

[van Buuren, 2011] Stef van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3 (2011), pp. 1–67 (cit. on p. 45).

[Venugopalan, 2021] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D. Wang. “Multimodal deep learning models for early detection of Alzheimer’s disease stage”. In: *Sci. Rep.* 11.1 (Dec. 2021), p. 3254 (cit. on p. 33).

[Vieira, 2020] Sandra Vieira, Walter Hugo Lopez Pinaya, Rafael Garcia-Dias, and Andrea Mechelli. “Multimodal integration”. In: *Mach. Learn. - Methods Appl. to Brain Disord.* Elsevier, 2020. Chap. 16, pp. 283–305 (cit. on p. 33).

[Vincent, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. “Extracting and composing robust features with denoising autoencoders”. In: *Proc. 25th Int. Conf. Mach. Learn. - ICML ’08*. New York, New York, USA: ACM Press, 2008, pp. 1096–1103 (cit. on p. 44).

[Wan, 2013] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. “Regularization of Neural Networks using DropConnect”. In: *Proc. 30th Int. Conf. Mach. Learn.* 2013, pp. 1058–1066 (cit. on p. 18).

[Wang, 2013] Sida Wang and Christopher Manning. “Fast dropout training”. In: *Proc. 30th Int. Conf. Mach. Learn.* 28.2 (2013). Ed. by Sanjoy Dasgupta and David McAllester, pp. 118–126 (cit. on pp. 18, 41, 82).

[Wang, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Nathan Srebro. “Stochastic optimization for deep CCA via nonlinear orthogonal iterations”. In: *2015 53rd Annu. Allert. Conf. Commun. Control. Comput.* IEEE, Sept. 2015, pp. 688–695 (cit. on p. 42).

[Wei, 2019] Wen Wei, Emilie Poirion, Benedetta Bodini, Stanley Durrleman, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot. “Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis”. In: *Med. Image Anal.* 58 (2019) (cit. on p. 33).

[Wei, 2020] Wen Wei, Emilie Poirion, Benedetta Bodini, Matteo Tonietto, Stanley Durrleman, Olivier Colliot, Bruno Stankoff, and Nicholas Ayache. “Predicting PET-derived myelin content from multisequence MRI for individual longitudinal analysis in multiple sclerosis”. In: *Neuroimage* 223 (Dec. 2020), p. 117308 (cit. on pp. 6, 33).

[Weiner, 2013] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, Enchi Liu, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Mark E. Schmidt, Leslie Shaw, Li Shen, Judith A. Siuciak, Holly Soares, Arthur W. Toga, and John Q. Trojanowski. "The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception". In: *Alzheimer's Dement.* 9.5 (Sept. 2013), e111–e194 (cit. on p. 3).

[Wu, 2018] Mike Wu and Noah Goodman. "Multimodal Generative Models for Scalable Weakly-Supervised Learning". In: *Adv. Neural Inf. Process. Syst. 31*. Ed. by S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett. Curran Associates, Inc., 2018, pp. 5575–5585 (cit. on p. 34).

[Yang, 2020] Junlin Yang, Xiaoxiao Li, Daniel Pak, Nicha C Dvornek, Julius Chapiro, MingDe Lin, and James S Duncan. "Cross-Modality Segmentation by Self-supervised Semantic Alignment in Disentangled Content Space". In: *Domain Adapt. Represent. Transf. Distrib. Collab. Learn.* Ed. by Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, Nicola Rieke, Holger Roth, Daguang Xu, and Ziyue Xu. Cham: Springer International Publishing, 2020, pp. 52–61 (cit. on p. 33).

[Yeung, 2017] Serena Yeung, Anitha Kannan, Yann Dauphin, and Li Fei-Fei. "Tackling Over-pruning in Variational Autoencoders". In: (June 2017). arXiv: 1706.03643 (cit. on p. 13).

[Zhang, 2012] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W Laird, and Xianghong Jasmine Zhou. "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data". In: *Nucleic Acids Res.* 40.19 (2012), pp. 9379–9391 (cit. on p. 7).

[Zhou, 2019a] Tao Zhou, Kim-Han Thung, Xiaofeng Zhu, and Dinggang Shen. "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis". In: *Hum. Brain Mapp.* 40.3 (2019), pp. 1001–1016 (cit. on p. 6).

[Zhou, 2019b] Tao Zhou, Kim-Han Thung, Xiaofeng Zhu, and Dinggang Shen. "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis". In: *Hum. Brain Mapp.* 40.3 (Feb. 2019), pp. 1001–1016 (cit. on p. 34).

[Zhou, 2020] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. “Hi-Net: Hybrid-Fusion Network for Multi-Modal MR Image Synthesis”. In: *IEEE Trans. Med. Imaging* 39.9 (2020), pp. 2772–2781 (cit. on p. 33).

[Zigel, 2009] Yaniv Zigel, Dima Litvak, and Israel Gannot*. “A Method for Automatic Fall Detection of Elderly People Using Floor Vibrations and Sound—Proof of Concept on Human Mimicking Doll Falls”. In: *IEEE Trans. Biomed. Eng.* 56.12 (Dec. 2009), pp. 2858–2867 (cit. on p. 96).

List of Figures

2.1 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. 22

2.2 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. 22

2.3 Testing benchmark of four variational methods applied to the multi-channel scenarios in Tab. 3.12 (cases snr $= 10$, $l_{\text{fit}} = l$). Sparse Multi-Channel models performs consistently better than non-sparse Multi-Channel ones. 24

2.4 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: (left) disease status among Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Normal Cognition (NC), (right) age, in almost orthogonal directions. Classification accuracy for these subjects is given in the fifth numeric column of Tab. 2.3. 26

2.5 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. Stars indicate the sampling points along trajectories. The trajectories share the same origin. (b) MRI data evolution. (c) FDG-PET. (d) Amyloid-PET. All the trajectories show a plausible evolution across disease and healthy conditions. 27

2.6 Generative parameters $\mathbf{G}_c^{(\mu)}$ (*cfr.* Eq. (2.7)) 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. . . 28

2.7 Negative lower bound (NLB) on the synthetic training set computed at convergence for all the scenarios. Each bar shows mean $\pm$ std.err. of $N = 80$ total experiments as a function of the number of fitted latent dimensions. Red bars represents experiments where the number of true and fitted latent dimensions coincide. (a) Experimental setup $C = 10$, $d_c = 32$: NLB stops decreasing when the number of fitted latent dimension coincide with the generated ones; notable gap between the under-fitted and over-fitted experiments (elbow effect). (b) Experimental setup $d_c = 4$, $l = 4$: increasing the number of channels C makes the elbow effect more pronounced. (c) Experimental setup $C = 10$, $d_c = 500$: with high dimensional data ($d_c = 500$) using the lower bound as a model selection criteria to assess the true number of latent dimensions may end up in overestimation. (d) Restricted ($N = 5$ total experiments) high quality experimental setup $C = 10$, $d_c = 500$, $S = 10000$, $\mathrm{SNR} = 100$: the risk to overestimate the true number of latent dimensions can be mitigated by increasing the SNR and S of the observations in the dataset. 29

2.8 Reconstruction error on synthetic test data reconstructed with the multi-channel model. The reconstruction is better for high SNR and high training data sample size. Scenarios where generated by varying one-at-a-time the dataset attributes listed in Tab. 3.12 for a total of $8\,000$ experiments. (a) Mean squared error from the ground truth test data using the Multi-Channel reconstruction: $\hat{\mathbf{x}}_i = \mathbb{E}_c \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c,\boldsymbol{\phi}_c)} \left[p\left(\mathbf{x}_i|\mathbf{z},\boldsymbol{\theta}_i\right)\right]\right]$. (b) Mean squared error from the ground truth test data using the Single-Channel reconstruction: $\hat{\mathbf{x}}_i = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i,\boldsymbol{\phi}_i)} \left[p\left(\mathbf{x}_i|\mathbf{z},\boldsymbol{\theta}_i\right)\right]$. (c) Ratio between Multi- *vs* Single-Channel reconstruction errors: we notice that the error made in ground truth data recovery with multi-channel information is systematically lower than the one obtained with a single-channel decoder. 30

3.1 General variational framework for our multi-view and multi-dataset model. Compatibly with the MCVAE formulation, for every pair of views w and v there is a prediction path $w \rightarrow v$ composed by two learnable functions: the encoding distribution $q\left(\mathbf{z}|\mathbf{x}_{d,n,w},\boldsymbol{\phi}_w\right)$ and the decoding likelihood $p\left(\mathbf{x}_{d,n,v}|\mathbf{z},\boldsymbol{\theta}_v\right)$. Parameters $\boldsymbol{\phi}_w$ and $\boldsymbol{\theta}_v$ are optimized through Eq. (3.4) to maximize the likelihood of our generative model under the encoding distributions, and at the same time minimize the Kullback-Leibler distance between every encoding distribution and the prior $p\left(\mathbf{z}\right)$. 37

3.2 Simple example of a Multi-Task Model learning scheme in the presence of missing not available (NA) views. Arrows represent learnable functions used as network encoders and decoders, transforming respectively input views (*e.g.,* clinical scores, imaging derived phenotypes, ...) from the observation space to the representation space (circles) and from the representation space back to the observation space. The separability of the loss function $\mathcal{L}_v^{(\mathbf{x}_{d,n})}$ in Eq. (3.2) allows to group together observations into homogeneous learning tasks. For every task, functions associated to missing views (dashed gray arrows) are locally not updated by the learning algorithm. Globally, common latent representations (red circles) across pairs of tasks act as a link allowing the information to flow throughout the views. 39

3.3 Pictorial example of training an imaging dataset with two views: MRI (left side, in gray scale) and FDG-PET (right side, in color scale). In this case we have data from 30 independent observations: 10 with left-views only; 10 with right-views only; 10 with complete views. The fraction of observations with complete views amounts to: $f = 1/3$. 42

3.4 Mean Squared Error (MSE) of imputation in synthetic held-out datasets (5-folds cross-validation). Compared to the best competing methods among k-Nearest Neighbor ($k = \{1, 5\}$) and Denoise Autoencoder (DAE), our model comes out as the best performer, with a mean MSE improvement of 17% in MAR cases (a) and 71% in MNAR cases (b). Stratification by signal-to-noise ratio (SNR) is shown. 45

3.5 Mean Squared Error of test sets predictions in synthetic held-out datasets in MAR scenarios. Stratification by SNR and by the fraction f of data-points with complete views is shown. A value of $f = 0.25$ is enough to reduce the prediction error on testing data-points at the level of the ideal case ($f = 1$). 46

4.1 Correlation matrices (in range $[-1, 1]$) before and after MRI gray matter volume harmonization. ComBat-adjusted matrices are visually indistinguishable from Raw, where each center is characterized by its own data covariance, whereas between-center differences are still conspicuous, meaning that these covariances are not harmonized. With CovBat covariances are harmonized because between-center differences are less noticeable. With DIVA, both linear an non-linear (4 layers architecture), the original data correlations are lost. All these visual clues are quantified with the Frobenius norm in Tab. 4.2 and Tab. 4.3. 70

4.2 The bias inducing matrix $\mathbf{L}$ (see Eq. (4.5)) is the Cholesky decomposition of a symmetric positive definite matrix $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T$. One way to create $\mathbf{\Sigma}$ is to randomly generate a matrix of k d-dimensional loadings $\mathbf{W} \in \mathbb{R}^{d \times k}$ with $k < d$, then form covariance matrix $\mathbf{W}\mathbf{W}^T$ and add to it a random diagonal matrix $\mathbf{D}$ with positive elements to make $\mathbf{W}\mathbf{W}^T + \mathbf{D}$ full rank. The resulting covariance matrix can be normalized to have ones on its diagonal. Here we see examples generated with $k \in \{1, 5, 10\}$ and $d = 50$ features. 72

4.3 Label classification accuracy (the higher the better) of a LDA classifier on harmonized synthetic data created by varying the parameters in Tab. 4.4. Random chance level (acc$= 0.5$) rendered as a dashed line. The accuracy is measured on test observations coming from the dataset not used to train the classifier. 75

4.4 Dataset classification accuracy (the lower the better) of a LDA classifier on harmonized synthetic data created by varying the parameters in Tab. 4.4. Random chance level (acc$= 0.5$) rendered as a dashed line. 76

5.1 Generalization of the (a) Variational Autoencoder (VAE) latent variable model to the (b) multi-channel (or multi-view) case, where multiple related views are encoded into and decoded from the same latent space. In (c) further extension to the multi-task case, where a specific optimization scheme allow missing non available (NA) data in the training phase, to jointly model observations from multiple datasets. Arrows represent learnable functions used as network encoders and decoders, transforming respectively input views (*e.g.,* clinical scores, imaging derived phenotypes, ...) from the observation space to the representation space (circles) and from the representation space back to the observation space. Globally, common latent representations (red circles) across pairs of tasks act as a link allowing the information to flow throughout the views. 81

5.2 Inheritance scheme of the main classes in the `mcvae` package. Class names/attributes/methods are separated by dashed lines. The user can define her/his own VAE module by redefine new `init_encoder()` and `init_decoder()` methods to output the desired distributions (*e.g.,* Normal, Categorical, Bernoulli, *etc.*). The `Mcvae` class builds the MCVAE model based on the input data and user defined architecture. The `MtMcvae` class builds the MT-MCVAE model, where the optimization is guided by the observation identifiers (`ids`) needed to correctly pair observations between channels. 83

5.3 Losses plotted in absolute and relative scale with the `plot_loss` utility (ll: log-likelihood, kl: Kullback-Leibler divergence; total: kl - ll). 88

5.4 Multi-channel observations projected in one selected latent dimension z_i with the `plot_latent_space` utility. The utility can optionally take a grouping variable to highlight clusters of points (diagnosis in this figure). . . . 88

5.5 Dropout probability of a "sparse" model plotted with the `plot_dropout` utility. 89

6.1 (a) Generative model of the Multi-Channel Variational Autoencoder (MCVAE), where a common latent $\mathbf{z}_x$ is the only source for the observations $\mathbf{x}_c$. (b), (c) Possible generative models of a MCVAE with modality specific latent variables, hosting a disentangled, complementary, and richer source of information. 94

6.2 Generative models of: (a) Domain Invariant Variational Autoencoder (DIVA); (b) possible Multi-Channel extension of the DIVA. Samples $\mathbf{z}_{d,\cdot}$ come from the prior $p(\mathbf{z}_{d,c}|d)$ conditioned on the domain label d. Samples $\mathbf{z}_x$ comes from a Standard Gaussian as in VAE and MCVAE. 95

6.3 Generative model for a Domain Invariant Multi-Channel Variational Autoencoder with joint ($\mathbf{z}_x$) and disjoint ($\mathbf{z}_c$) latent spaces. It comes from the combination of the models proposed in Fig. 6.1b and Fig. 6.2b. 95

List of Tables

2.1 Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (2.18). 21

2.2 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. 23

2.3 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 ($p < 0.2$). 25

3.1 The Multi-Task Multi-Channel VAE (MT-MCVAE) extends the MCVAE, which is itself an extension of the VAE. 40

3.2 Mean squared error (MSE) and negative log-likelihood (NLL) - the lower the better - measured as $\text{mean}_{\text{st.dev.}}$ on the reconstructed brain images of the test-set. The MRI were used to infer the FDG-PET slices in the same subject, and *vice versa*. Results stratified by f, the fraction of observations with no missing views in the training set. Notice the immediate drop in the error metrics as soon as f increases. 42

3.3 Number of subjects per view available in each dataset. The last columns provide the size of the intersection ($\cap$) and union ($\cup$) of subjects with available views. Notice how in the jont set no subject has all the modalities. 48

3.4 Mean Squared Error (MSE) of test data from Adni2. All models were trained on all the available datasets by holding-out data from the Adni2 test dataset. 5-folds cross validation of MSE is shown as $\text{mean}_{\text{st.dev.}}$. Best results in boldface are significant with an α level of 0.01 with respect to both competing methods. 49

3.5 Mean Squared Reconstruction Error (the lower the better) measured on test dataset views (clinical scores and imaging derived phenotypes) predicted with the Multi-Channel VAE (MCVAE) and the Multi Task MCVAE (MT-MCVAE). 5-folds cross-validation results shown as $\text{mean}_{\text{st.dev.}}$. Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Experiments were run in two different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; In all cases the MT-MCVAE performs either similarly or statistically better than the MCVAE, with alpha levels at 0.05 (*), 0.01 (**), and 0.001 (***). 51

3.6 Mean Squared Reconstruction Error (the lower the better) measured on test dataset views (clinical scores and imaging derived phenotypes) predicted with our model. 5-folds cross-validation results shown as as $\text{mean}_{\text{st.dev.}}$. Models were trained on all the available views in the training dataset, independently of their presence in the testing dataset. Experiments were run in three different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; 3) when models are trained on all the available datasets except the testing one, or Multi Task Learning (MTL). We measure a better performance in the MTL condition with respect to the STIB (§) in $7/12$ of cases, and in $10/12$ of cases with respect to the average STEB (†) experiments. 53

3.7 Mean Squared Reconstruction Error (mean (st.dev.), the lower the better) measured on clinical scores and imaging derived phenotypes predicted with our MT-MCVAE model in MTL experiments. Results stratified by the number of layers in the encoder-decoder architecture. We measure no significant differences among architectures (anova statistical test at an alpha level of 0.05). Best overall results in boldface. 53

3.8 Number of subjects stratified by dataset and diagnosis: Alzheimer's Disease (AD); Mild Cognitive Impairment (MCI); Normal Cognition (NC). 54

3.9 Experiment of diagnosis classification run with the Multi-Channel VAE (MCVAE) and the Multi Task MCVAE (MT-MCVAE). 5-folds classification accuracy in $\%$ is shown as mean (standard deviation). Since there are no MCI in miriad and oasis3 datasets, the classification tests 'AD vs MCI' and 'MCI vs NC' are meaningless and not reported. Since there are no NC in the geneva dataset, the classification tests 'AD vs NC' and 'MCI vs NC' are meaningless and not reported. Experiments were run in two different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case. In all cases the MT-MCVAE model performs either similarly or statistically better than the MCVAE, with alpha levels at 0.05 (*), 0.01 (**), and 0.001 (***). 55

3.10 Experiment of diagnosis classification run with our model. 5-folds classification accuracy in $\%$ is shown as mean (standard deviation). Experiments were run in three different conditions: 1) when training and testing data are chosen from the same dataset, or Single Task with Internal Benchmark (STIB) learning case; 2) when models trained on one dataset are tested on another dataset, or Single Task with External Benchmark (STEB) case; 3) when models are trained on all the available datasets except the testing one, or Multi Task Learning (MTL). In all cases we measure a better performance in the MTL condition with respect to the average STEB one (†). 57

3.11 Diagnosis classification with our model and the EmbraceNet (EN, [Choi, 2019]). Accuracy in % as mean$_{\text{st.dev.}}$ over 5-folds. Results are stratified by the classification task and by the number of layers in the encoder-decoder architecture. We measure no significant difference among architectures depth (anova test, alpha level 0.05) and between models (t-test, alpha level 0.05). 57

3.12 Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (3.16). 62

4.1 MRI Observations stratified by dataset and diagnosis. 68

4.2 Pairwise Frobenius norms between dataset-specific correlation matrices for every harmonization method. We find that ComBat adjustment does not harmonize the correlation matrices whereas CovBat adjustment shows large reductions in the between-datasets distances. The almost perfect reduction of the Frobenius distances in the DIVA-1 linear cases is spurious, as the original correlation structure is very different from the harmonized ones (see Tab. 4.3, Fig. 4.1). In the DIVA-4 non-linear case Frobenius distances generally increases, which is not ideal for an harmonization method. . . . 69

4.3 Pairwise Frobenius norms between between correlation matrices of harmonized and raw data for every dataset. With ComBat the original covariance matrices are unchanged ($||\cdot||_F = 0$). With CovBat the harmonization tends to slightly change the original covariance matrices ($||\cdot||_F < 10$) to harmonize them, while with both DIVA methods the original covariance structures become very different from the original ones ($||\cdot||_F > 40$). 69
4.4 Parameters, varied one-at-a-time in the prescribed ranges, used to generate synthetic scenarios for the experimental campaign. 73