An interpretable generative model for image-based predictions

Mauri, Chiara

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

An interpretable generative model for image-based predictions

Chiara Mauri

Kongens Lyngby 2022

Technical University of Denmark
Department of Health Technology
Ørsteds Plads, Building 345C,
2800 Kongens Lyngby, Denmark
healthtech-info@dtu.dk
www.healthtech.dtu.dk/

Summary (English)

The last decades have seen a significant development of computational methods that make automatic predictions of variables of interest, such as a subject's diagnosis or prognosis, based on brain Magnetic Resonance Imaging (MRI) scans. Since MRI is able to detect subtle effects in brain anatomy more than clinical assessment, these methods have a huge potential in clinical tasks such as early diagnosis, therapy planning and monitoring, paving the way to personalized treatments. Many different image-based prediction methods have been proposed in the literature, with a special focus on discriminative deep learning techniques in the last years.

In this thesis, we propose an alternative approach for image-based predictions, based on a lightweight generative method, which yields accurate and interpretable predictions. We first demonstrate that the proposed method achieves competitive performances as compared to state-of-the-art benchmarks in age and gender prediction tasks, especially when the sample size is at most of a few thousand subjects, which is the typical scenario in many neuroimaging applications. We then give insight into the interpretability properties of the proposed method: It automatically yields spatial maps displaying morphological effects of the variable of interest, which are straightforward to interpret. Being both accurate and interpretable, the proposed method bridges the gap between classical brain mapping techniques, which produce interpretable maps by studying effects of variables of interest on the brain on a population level, and more recent prediction methods, which provide accurate predictions on a subject-specific level.

We also present possible model extensions and applications, showing that the

proposed method can be easily extended to incorporate known covariates and/or nonlinearities. Finally, we discuss possible future work, such as extending the proposed method to a longitudinal setting, where more than one scan per subject is available.

Summary (Danish)

De sidste årtier har set en betydelig udvikling af beregningsmetoder, der foretager automatiske forudsigelser af variabler af interesse, såsom et forsøgspersons diagnose eller prognose, baseret på Magnetisk resonans (MR)-scanninger. Da MR er i stand til at detektere subtile effekter i hjernens anatomi mere end klinisk vurdering, har disse metoder et enormt potentiale i kliniske opgaver såsom tidlig diagnose, terapiplanlægning og overvågning, hvilket baner vejen for personlige behandlinger. Mange forskellige billedbaserede forudsigelsesmetoder er blevet foreslået i litteraturen, med særligt fokus på diskriminerende deep learning-teknikker i de seneste år.

I dette speciale foreslår vi en alternativ tilgang til billedbaserede forudsigelser baseret på en letvægts generativ metode, som giver nøjagtige og fortolkelige forudsigelser. Vi demonstrerer først, at den foreslåede metode opnår konkurrencedygtige præstationer sammenlignet med state-of-the-art benchmarks i alders- og kønsforudsigelsesopgaver, især når stikprøvestørrelsen højst er på et par tusinde forsøgspersoner, hvilket er det typiske scenarie i mange neuroimaging applikationer. Vi giver derefter indsigt i den foreslåede metodes fortolkningsegenskaber: Den giver automatisk rumlige kort, der viser morfologiske effekter af variabelen af interesse, som er ligetil at fortolke. Da den foreslåede metode både er nøjagtig og kan fortolkes, bygger den bro over kløften mellem klassiske hjernekortlægningsteknikker, som producerer fortolkbare kort ved at studere effekter af variabler af interesse på hjernen på et befolkningsniveau, og nyere forudsigelsesmetoder, som giver nøjagtige forudsigelser om et emne-specifikt niveau.

Vi præsenterer også mulige modeludvidelser og applikationer, hvilket viser, at den foreslåede metode let kan udvides til at inkorporere kendte kovariater og/eller

ikke-lineariteter. Til sidst diskuterer vi muligt fremtidigt arbejde, såsom at udvide den foreslåede metode til en longitudinel indstilling, hvor mere end én scanning pr. emne er tilgængelig.

Preface

This thesis was prepared at the Department of Health Technology at the Technical University of Denmark in partial fulfillment of the Ph.D. degree requirements. Professor Koen Van Leemput, from the Technical University of Denmark and the Athinoula A. Martinos Center for Biomedical Imaging (Harvard Medical School), acted as the main supervisor. Professor Mark Mühlau from Klinikum rechts der Isar (Technical University of Munich) acted as co-supervisor. The thesis focuses on prediction methods based on brain images.

Lyngby, 30-September-2022

Chiara Mauri

Acknowledgements

First, I would like to thank my supervisor Koen Van Leemput for guiding me through this Ph.D. project. His extensive knowledge of medical image analysis and his devotion to high quality research had made this Ph.D. exciting and challenging at the same time.

Secondly, I would like to thank my co-supervisor Mark Mühlau, whose research group I have joined in Munich for an external stay. His broad expertise in MS has widened my knowledge about the disease and its clinical significance.

A special thanks to Stefano Cerri and Oula Puonti, for their precious collaboration to the publications included in this thesis.

I also want to thank my other colleagues at DTU: Mikeal Agn, Sveinn Pálsson, and Ines Meyer, and my "new" office mate Jupeng Zhao for the great moments that we shared at DTU.

Last but not least, I want to thank my family for always giving me unconditioned support, and my friends in Copenhagen, especially Yvonne and Candela, with whom I have spent memorable moments during these years.

Scientific Contributions

Papers included in this thesis

Paper A: Chiara Mauri, Stefano Cerri, Oula Puonti, Mark Mühlau , Koen Van Leemput, Accurate and Explainable Image-based Prediction Using a Lightweight Generative Model, *MICCAI 2022*, LNCS 13438, pp. 1–11, 2022.

Paper B: Chiara Mauri, Stefano Cerri, Oula Puonti, Mark Mühlau , Koen Van Leemput, An Accurate and Interpretable Generative Model for Image-based Prediction, *In preparation.*

Contents

Chapter 1

Introduction

Magnetic Resonance Imaging (MRI) is extensively used in neuro-clinical practice, to support clinicians in making diagnosis and planning treatments. Developing computational methods that are able to automatically predict variables of interest, such as a subject's diagnosis or prognosis, directly from brain MRI scans has received increasing attention in the last decades, for its many potential clinical applications, such as providing early diagnosis and/or personalized treatment. Several image-based prediction methods have been therefore developed and tested, using different prediction techniques, with a boost of discriminative deep learning methods in the last few years, thanks to the increasing availability of very large imagining datasets. While these methods are able to produce accurate results, especially when trained on large amounts of data, they have been proven to be difficult to interpret Haufe et al. (2014); Wilming et al. (2022). This may be problematic, since, in many neuroimaging tasks, it is important not only to predict well, but also to interpret morphological changes underlying predictions. In this thesis, we therefore propose a lightweight generative model for image-based prediction, which yields *interpretable* results, without sacrificing prediction accuracy.

1.1 Contributions

In paper A, we developed an interpretable linear generative model for image-based prediction, and validated it with experiments on the UK Biobank.

In paper B, we explored in more details the proposed prediction method. In particular, in this paper we included more extensive experiments on the UK Biobank, with provided insights into our method's and benchmarks' performances in terms of bias-variance trade-off. We also provided more insights into interpretability aspects, and possible model extensions with inclusion of known variables and/or nonlinearities in the model, with related experiments.

1.2 Overview of the thesis

The remainder of the thesis has the following structure:

- Chapter 2 gives an introduction about image-based prediction methods in neuroimaging: first about their clinical applications, then a general overview of different classes of prediction methods, as well as a motivation for the choice of the method proposed in this thesis and its advantages.
- Chapter 3 describes the proposed prediction method, with details about both training and testing phases.
- Chapter 4 first presents experiments conducted on the UK Biobank, showing prediction performances of the proposed method on age and gender prediction tasks, as well as comparison with three benchmarks. It then illustrates interpretability properties of the proposed method, together with difficulties that other types of models experience in this regard. Finally, it provides more insight into the performances of the proposed method and benchmarks in terms of the bias-variance trade-off.
- Chapter 5 presents some possible model extensions and applications. It first describes how to extend the proposed method to incorporate variables that are possibly known about the subjects and/or nonlinearities. It then presents a possible application, where some model parameters estimated on a large training set are subsequently re-used in an experiment with a smaller cohort.
- Chapter 6 discusses the results presented in this thesis, and describes possible directions for future work.

Chapter 2

Image-based prediction

This chapter provides an overview of image-based prediction methods, specifically based on brain MRI scans, and is structured as follows:

- We first illustrate the clinical relevance and possible applications of these methods; and
- We then provide a general overview of different types of prediction methods proposed in the literature, and describe the motivations that led us to develop the proposed model, as well as its advantages.

2.1 Why image-based prediction?

MRI is a widely used technique for acquiring medical scans, which relies on a strong magnetic field to acquire three dimensional images of anatomy and physiological processes of the body. As compared to other imaging techniques such as Positron Emission Tomography (PET) and Computed Tomography (CT), it has the advantage of not using X-rays, or having to inject radioactive substances in patients, and it also produces images with a better contrast in soft tissues. Regarding its disadvantages, it should be mentioned that it is not suited for

subjects with metal implants, because of the strong magnetic field, and that bones are not well imaged with MRI.

MRI scans are extensively used in clinical practice to detect and monitor various diseases. A specific category of diseases where MRI is particularly significant are brain disorders, which are the most prevalent type of diseases in Europe, representing a burden and a huge cost for society. Since brain MRI scans can detect subtle morphological changes better than clinical assessment, developing computational methods that can predict a variable of interest directly from a subject's brain scan is of great interest. The aim of these image-based prediction methods can be to predict either a continuous variable such as a subject's disability score (*regression* methods), or a categorical variable such as a patient's diagnosis or prognosis (*classification* methods).

The potential clinical applications of these image-based prediction methods are numerous. For instance, automatic prediction of a subject's diagnosis based on brain MRI scans can allow to diagnose brain diseases earlier and more reliably than using only clinical assessment, thanks to the high sensitivity of MRI scans to subtle anatomical changes, with consequent better clinical outcomes. Furthermore, it is also particularly useful for diseases with no standard clinical tests, such as schizophrenia[1]. Another example is automatic prediction of disability scores from brain MRI scans, which can provide more accurate estimates of the scores than clinical tests, which are more noisy due to human factors. Another image-based prediction task of interest is to estimate disease progression of individual patients, for instance by identifying patients at higher risk of future disability accrual. This can lead to better and more personalized treatment planning and is particularly useful for diseases, such as multiple sclerosis, where several possible treatments are available[2], but their efficacy depends on the specific patient and may be hard to predict in the initial phase of the disease. Automatically predicting individual prognosis from brain scans has also the potential of uncovering subtle morphological and temporal mechanisms underlying disease progression.

With this huge potential for applications in clinical practice, prediction methods based on brain MRI scans have seen a huge development in the last decades, with many different prediction techniques proposed in the literature.

[1] https://www.nhs.uk/mental-health/conditions/schizophrenia/diagnosis/
[2] http://nationalMSsociety.org/DMT

2.2 Methods for image-based predictions

Classical methods for analysis of brain scans are the so-called human brain mapping techniques Wright et al. (1995); Ashburner et al. (2000); Davatzikos et al. (2001); Chung et al. (2001); Fischl et al. (2000); Snook et al. (2007); Friston et al. (1994); Worsley and Friston (1995); Friston et al. (1991); Worsley et al. (1992), which were developed since the 1990s. Their aim is to identify on a population level brain areas with significant differences between two groups of subjects (such as patients and healthy controls), regions correlated with specific variables of interest (such as age, disease severity, etc.), or interactions among various effects of interest. This aim is achieved by registering imaging data to a standard template space, and then performing statistical tests on voxel-level measurements independently. Since these brain mapping approaches analyze single voxel-level measurements separately, they are called *mass-univariate* methods. These techniques originally started with functional imaging, first using PET imaging and then functional MRI, using activation maps as input data Friston et al. (1991); Worsley et al. (1992); Friston et al. (1994); Worsley and Friston (1995). Subsequently, they translated to structural MRI, with input data of several types, such as deformations Chung et al. (2001); Davatzikos et al. (2001), tissue density maps, such as gray or white matter segmentations Wright et al. (1995); Ashburner et al. (2000) , cortical thickness measurements Fischl et al. (2000), and voxel-level values in diffusion tensor imaging Snook et al. (2007).

While these techniques generate valuable maps of various effects of interest that are straightforward to interpret, they cannot provide accurate predictions at an individual-level, since they consider each voxel independently. For instance, a single voxel displaying a significant group difference is not necessarily a good classifier at an individual-level. Therefore, with the goal to provide accurate subject-specific predictions, *multivariate* methods have been subsequently developed, which consider all voxel-level measurements simultaneously, capturing multivariate association patterns. These methods are able to leverage the predictive power of many voxels, which may be only poorly predictive if used independently, to achieve accurate predictions.

Several multivariate methods for image-based prediction have been developed in the last decade, most of which are *discriminative* models, which directly predict a variable of interest from a subject's image. As opposed to these models, *generative* methods express the image as function of the target variable, and then need to be "inverted" to provide predictions of the variable of interest. Another possible classification of prediction methods divides them into linear, shallow nonlinear and deep nonlinear models Schulz et al. (2019): linear and shallow nonlinear models are classical machine learning methods, which may include nonlinearities through e.g. kernels, as opposed to deep learning models,

which are characterized by a cascade of sequential nonlinear functions.

Many methods developed for image-based prediction have been tested on a particular application: age prediction based on a subject's brain scan - the so called brain age Cole et al. (2019). This prediction task has been vastly used for model development, since it is straightforward to collect age information about subjects. In this sense, it represents a unique case in image-based prediction, since typical neuroimaging applications count on much smaller datasets. Furthermore, during the last three years (i.e. during this PhD project), the development of brain age prediction methods has received a further boost, thanks to the increasing access to large datasets, comprising thousands or even tens of thousand of data. Apart from its value for model development, predicting brain age has also clinical applications, since the difference between *brain* age and *chronological* age (called the "brain age gap") has been proven to be a potential biomarker of healthy aging and patological deviations. A brain age gap larger than average is in fact associated with several neurological diseases, such as Alzheimer's disease, dementia, schizophrenia, and multiple sclerosis Cole et al. (2019); Kaufmann et al. (2019).

Fig. 2.1 shows a classification of image-based prediction models that have been proposed for age prediction, divided into discriminative vs generative, and linear vs nonlinear models. In each quadrant, we display the number of methods of the corresponding class that have been proposed in the literature, according to Cole et al. (2019) and our subsequent literature review on the topic. For simplicity, we have grouped together linear and shallow nonlinear models.

First, we observe that generative models, in particular linear and shallow nonlinear ones, are almost unexplored for image-based predictions, with the greatest majority of methods being discriminative. In particular, we found only two deep nonlinear generative models proposed for age prediction: Zhao et al. (2019) which employs a Variational Autoencoder (VAE), and Wilms et al. (2020) which models the bidirectional functional relationship between brain morphology and age using normalizaing flows. The only linear generative model that we found for age prediction is He et al. (2020), which uses a kernel regression method. Regarding a possible comparison between discriminative and generative models, prior studies in non-neuroimaging tasks compared classification performances of linear generative methods, such as Naive Bayes classifier, and discriminative models (such as logistic regression or rule learners), showing two distinct regimes of performances: for limited sample sizes, the Naive Bayes classifier achieves better results, while after a certain training size the roles are reserved Domingos et al. (1997); Ng et al. (2002); Domingos (2012). However, a similar comparison in neuroimaging prediction tasks is missing.

Regarding the use of linear vs nonlinear methods, until 2019 the majority of

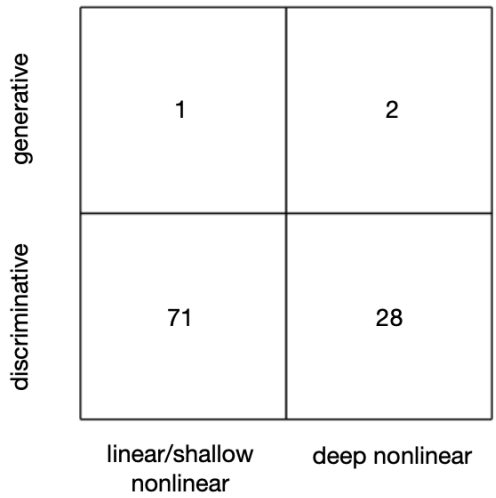


Fig. 2.1: Number of image-based prediction models proposed for age prediction, classified into generative vs discriminative, and linear/shallow nonlinear vs deep nonlinear models.

image-based prediction methods tested on age prediction were linear or shallow-nonlinear discriminative methods, mostly Support Vector Machine, Relevance Vector Machine, Gaussian Process Regression and Elastic Net Cole et al. (2019). The review in Cole et al. (2019) in fact reports 38 studies that use machine learning methods for brain age prediction, and only 5 studies employing neural networks. In the following years, the field has seen a significant development of deep learning discriminative models, especially Convolutional Neural Networks based on 2D image slices or on 3D volumes: we counted 71 studies that use machine learning methods, and 28 studies based on deep learning models, offering a quite different picture from the one provided a few years ago by Cole et al. (2019). About possible advantages of adding deep nonlinearities in image-based prediction methods, there is an ongoing discussion in the neuroimaging field, mostly for the discriminative case, debating if deep learning models are actually beneficial for performances or if they provide the same results as their linear counterparts. On one hand, studies such as He et al. (2020) and Schulz et al. (2019) showed that discriminative neural networks and simple linear methods achieve comparable prediction performances. On the other hand, studies like Peng et al. (2021) employed neural networks to achieve state-of-the-art performances, reporting better results than linear discriminative benchmarks.

This overall picture about image-based prediction methods drew our attention

towards the class of linear and shallow nonlinear generative models, which is essentially yet unexplored in neuroimaging. Therefore, in this PhD project we decided to investigate this class of methods and to explore if they can achieve good prediction performances. Besides the desire to investigate an unexplored class of methods and to corroborate previous findings about discriminative vs generative and nonlinear vs linear models, this choice was motivated by the still unmet need in neuroimaging of developing prediction models that are *interpretable*. In fact, in neuroimaging applications, it is of extreme importance not only to achieve accurate predictions, but also to interpret the underlying anatomical mechanisms, which has proven to be difficult with discriminative methods Haufe et al. (2014); Wilming et al. (2022).

Therefore, in this project we developed a linear or shallow non-linear generative method for image-based predictions, and demonstrated that it achieves accurate and interpretable predictions. The proposed method consists of a *causal* part, directly expressing the effect of the variable of interest on a subject's image, and a *noise* component, which uses latent variables to capture correlations between voxels, and to automatically model the variability in the images which is not due to the target variable. For the choice of a noise model based on latent variables, we took inspiration from the VAE. The resulting *forward* model needs to be subsequently "inverted" to make predictions about the variable of interest. The proposed model can be regarded as an extension of the Naive Bayes classifier, where the strong assumption of feature independence conditioned on the class, which does not suit imaging tasks, is relaxed.

The proposed method for image-based prediction provides the following advantages:

- It combines mass-univariate and multivariate approaches,
- It yields interpretable predictions,
- It works well with limited sample sizes,
- It is simple and fast to use.

We will now give more insight into these four advantages.

Combining mass-univariate and multivariate approaches

The proposed method has the conceptual advantage of bridging the gap between classical *mass-univariate* brain mapping techniques, which have the form of linear generative models that are not subsequently inverted, and

state-of-the-art discriminative *multivariate* methods. In fact, the causal part of the proposed method, like univariate approaches, expresses and fits with a generative model the effect of the variable of interest on each voxel independently. However, unlike these classical approaches, our method has a latent variable noise model capturing correlations between voxels, and is subsequently inverted to make predictions about the target variable. This yields a discriminative predictor that, like multivariate methods, captures correlation patterns and leverages the contribution of all voxels simultaneously.

Note that other methods have been proposed in the literature for combining generative and discriminative approaches Batmanghelich et al. (2011); Varol et al. (2018), in order to achieve accurate single-subject predictions while remaining interpretable. However, these methods had to artificially constrain the models' weights, to keep them interpretable. Instead, the proposed method naturally combines generative univariate and discriminative multivariate approaches, yielding accurate and interpretable predictions without having to impose external interpretability constraints.

Interpretable predictions

In the neuroimaging field, the ability to interpret a model's predictions is of great interest. The proposed method meets this need, by directly modeling the effect that the target variable has on image intensities. This yields weight maps showing target-related changes in neuroanatomy on a population level, which are therefore straightforward to interpret Haufe et al. (2014). These maps are in common with classical brain mapping techniques, but, as opposed to those methods, the proposed model also provides predictions about the target variable.

The straightforward availability of these maps is an advantage of our method, which other prediction models do not offer, be they generative or discriminative. In fact, nonlinear generative models can still produce interpretable maps, but these are not readily available and instead require heavy computations Zhao et al. (2019); Wilms et al. (2020). Regarding discriminative methods, they have proven to be difficult to interpret, both in the linear and nonlinear case Haufe et al. (2014); Arun et al. (2021); Ghassemi et al. (2021); Adebayo et al. (2018); Rudin (2019); Wilming et al. (2022); Lipton (2018); Sixt et al. (2020); Gu and Tresp (2019). In fact, in recent years, the Explainable AI (XAI) field has developed many techniques that aim at explaining a discriminative model's decision, producing maps that highlight important areas for prediction - the so-called saliency maps Ras et al. (2022); Simonyan et al. (2014); Baehrens et al. (2010); Erhan et al. (2009); Shrikumar et al. (2017); Sundararajan et al. (2017); Springenberg et al. (2014); Selvaraju et al. (2017); Smilkov et al. (2017a); Zeiler and Fergus (2014); Bach et al. (2015); Ribeiro et al. (2016);

Lundberg and Lee (2017); Ribeiro et al. (2018); Kindermans et al. (2017); Montavon et al. (2017); Simonyan and Zisserman (2014); Fisher et al. (2019); Zien et al. (2009), but these methods have been shown to suffer from several difficulties. The main conceptual issue is that, while saliency maps identify areas in the input image that were most relevant for making a certain prediction, they do not explain *why* those regions were important Ghassemi et al. (2021); Rudin (2019). This problem has been illustrated in Wilming et al. (2022), where many XAI techniques have been used to locate the signal of interest, on synthetic data with structured noise. The majority of the tested saliency maps failed in retrieving the signal and instead highlighted a mixture of signal and noise, showing that regions that are important for predictions are not necessarily directly affected by the signal. This also holds in the apparently simple case of linear discriminative methods, where nonzero weights are used to amplify the signal of interest or remove noise from the image, making their interpretation problematic Haufe et al. (2014); Wilming et al. (2022).

Limited sample sizes

As we will show in the next chapters, the proposed method achieves competitive performances in prediction tasks, especially when the training set size is limited. This property meets the need in neuroimaging of developing models that can learn efficiently from quite small training sets. In fact, while tasks commonly used to develop prediction models such as age prediction can count on tens of thousands of scans from very large datasets Alfaro-Almagro et al. (2018); German National Cohort Consortium (2014); Breteler et al. (2014); Schram et al. (2014), sample sizes in typical neuroimaging applications are much more limited. For instance, Arbabshirani et al. (2017) contains a review of over 200 studies on predictions of brain diseases from various neuroimaging modalities, showing that sample sizes are typically small, with mean and median of only 186 and 88 subjects, respectively, as illustrated in Fig. 2.2.

Even in extremely large studies as the UK Biobank Sudlow et al. (2015); Miller et al. (2016); Alfaro-Almagro et al. (2018), the number of subjects with quite common disorders is fairly limited. The UK Biobank is the world's largest epidemiological and imaging prospective study, comprising 500.000 subjects, 100.000 of which have been selected for multimodal imaging acquisition Alfaro-Almagro et al. (2018). The scanning process is still underway, with the 50.000th participant scanned in January 2022[3]. In 2022, the UK Biobank should roughly include scans of 900 subjects with stroke, 900 with Alzheimer's Disease, and 600 with Parkinson's Disease, while in 2027, these amount are projected to increase to 4.000 subjects

[3] https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/world-s-largest-imaging-study-scans-50-000th-participant

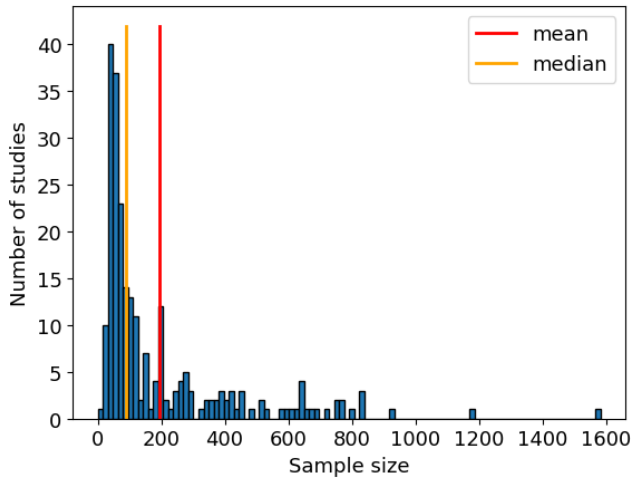


Fig. 2.2: Histogram of sample sizes in brain diseases prediction studies, as reported by Arbabshirani et al. (2017).

with stroke, 6.000 with Alzheimer's Disease, and 2.800 with Parkinson's Disease, considering these diseases' prevalence in the population and the number of scanned subjects Sudlow et al. (2015); Alfaro-Almagro et al. (2018). Regarding other diseases such as Multiple Sclerosis and epilepsy, in 2021, there were only 87 scans of subjects with MS and 185 of subjects with epilepsy in the UK Biobank (according to the imaging data to which we had access). These estimates reveal that, even in the world's biggest imaging project, the number of subjects with quite common disorders is fairly modest, and will not be huge even in the coming years.

Analogous considerations hold for other large prospective imaging studies, such as the the German National Cohort, which plans to scan 30.000 participants to study several major chronic disorders German National Cohort Consortium (2014), the Rhineland Study, which targets to acquire images of 30.000 subjects to investigate neurodegenerative and neuropsychiatric diseases Breteler et al. (2014), and the Maastricht Study, including 10.000 participants, which is however augmented with type 2 diabetes subjects, to better investigate this disease Schram et al. (2014).

Other imaging cohorts gathered for investigating specific disorders and/or healthy aging comprises at most 1000-2000 subjects, such as ADNI Jack Jr et al. (2008), ABIDE Di Martino et al. (2014), AIBL Ellis et al. (2009), CoRR Zuo et al. (2014), HCP Glasser et al. (2016), PING Jernigan et al. (2016), PNC Satterthwaite et al. (2014), SHIP Hosten et al. (2021).

Given this scenario characterizing many potential neuroimaging applica-

tions, it is important to develop prediction models that are able to achieve accurate results on modest sample sizes, and we will demonstrate that the proposed method meets this goal.

Simple and fast to use

The proposed method has also the advantage of being simple and fast to use. In fact, as we will show in the next chapter, it does not need approximations, having closed-form expressions for training and testing, and it is simple to tune, with only one hyperparameter to estimate. Additionally, training the proposed method is fast for typical sample sizes, without the need for GPUs.

Conversely, deep learning methods can be much harder to use, with many more knobs to turn, and time consuming to train, even on GPUs. For example, a deep learning method that was recently proposed for brain age prediction Peng et al. (2021) reports the selection of a good combination of data augmentation technique, optimizer, training loss, batch size, etc., and a training time of 65 hours for about 13.000 data, with two GPUs. We trained this model as benchmark, following the same setting described in Peng et al. (2021), and even found that a special GPU with large memory - which may be difficult to have - was necessary to train the model for the given data resolution and batch size.

Given these advantages, the proposed method is well-suited for image-based prediction applications, yielding interpretable results without sacrificing prediction accuracy, especially in typical scenarios with moderate sample sizes.

Chapter 3

Proposed generative prediction method

In this chapter, we describe the proposed method for image-based predictions, with the following outline:

- Forward generative model;
- Model inversion to make predictions; and
- Model training.

3.1 Forward model

Let $\mathbf{t} \in \mathbb{R}^J$ denote a a vectorized version of a subject's image, containing J voxel-level measurements, and $x \in \mathbb{R}$ a variable of interest about that subject, that we aim to predict. The proposed forward generative model has the form

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \boldsymbol{\eta}, \tag{3.1}$$

where $\boldsymbol{\eta} \in \mathbb{R}^J$ is a vector with random noise, with distribution

$$p(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, \mathbf{C}), \tag{3.2}$$

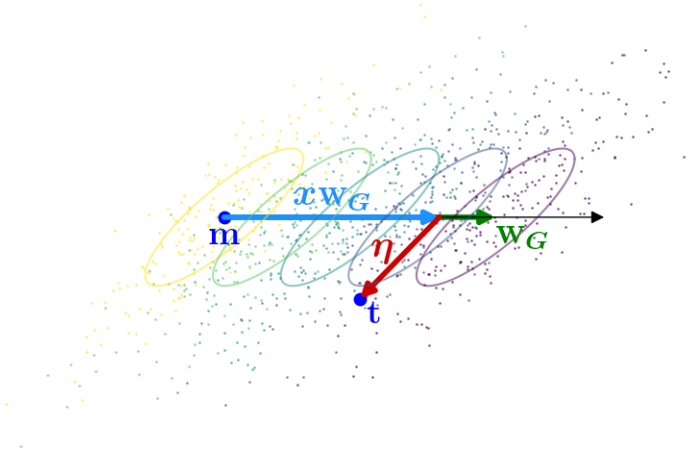


Fig. 3.1: Cartoon illustration of the generative model in (3.1).

and $\mathbf{w}_G, \mathbf{m} \in \mathbb{R}^J$ are two spatial weight maps expressing the effect of the variable of interest x on voxels of $\mathbf{t}$, and a mean effect, respectively. This generative model is illustrated in Fig. 3.1, in a toy 2D example: the model involves decomposing the input signal $\mathbf{t}$ as a sum of a mean effect $\mathbf{m}$, the target effect $x\mathbf{w}_G$, and a subject-specific noise vector $\boldsymbol{\eta}$.

This forward model can also be extended to include known variables about the subject and/or a nonlinear function of x, as we will show in chapter 5. For the rest of the thesis, we will name $\mathbf{w}_G$ the *generative* weight map, and gather the two spatial maps in a single matrix $\mathbf{W} = (\mathbf{m}, \mathbf{w}_G)$.

Note that the model in (3.1) with diagonal $\mathbf{C}$ is the linear generative model used in classical mass-univariate brain mapping techniques, where statistical tests are performed on $\mathbf{w}_G$, to identify brain areas with significant group differences or associated with specific variables of interest. As opposed to these methods, here we consider a non-diagonal $\mathbf{C}$ with spatial structure, as we will show in the next section, which allows to obtain accurate predictions about x once the model is inverted.

3.2 Inverting the model to make predictions

Let us now assume that the parameters $\mathbf{W}$ and $\mathbf{C}$ of the model are known, and that we have an unseen subject with image $\mathbf{t}^*$ and unknown variable of interest x^*. We can then invert the model with Bayes' rule to make predictions about the target x^*.

In the classification case of a binary target variable $x^* \in \{0, 1\}$, assuming the two values have equal prior probability, we get:

$$p(x^* = 1|\mathbf{t}^*, \mathbf{W}, \mathbf{C}) = \sigma\big(\mathbf{w}_D^T \mathbf{t}^* + w_o\big), \tag{3.3}$$

where $\sigma(\cdot)$ is the logistic function, $w_o = -\mathbf{w}_D^T(\mathbf{m} + \mathbf{w}_G/2)$, and

$$\mathbf{w}_D = \mathbf{C}^{-1}\mathbf{w}_G. \tag{3.4}$$

Note that (3.3) has the form of a logistic classifier with *discriminative* weights $\mathbf{w}_D$. We can then predict x^* as 1 if

$$\mathbf{w}_D^T \mathbf{t}^* + w_o > 0, \tag{3.5}$$

and 0 otherwise.

In the regression case of a continuous target variable with a flat prior $p(x^*) \propto 1$, the posterior distribution is Gaussian, of the form:

$$p(x^*|\mathbf{t}^*, \mathbf{W}, \mathbf{C}) = \mathcal{N}(x^*|y(\mathbf{t}^*), \sigma_x^2), \tag{3.6}$$

where we have defined mean

$$y(\mathbf{t}^*) = \sigma_x^2(\mathbf{w}_D^T \mathbf{t}^* + b_0), \tag{3.7}$$

and variance

$$\sigma_x^2 = \big(\mathbf{w}_G^T \mathbf{C}^{-1} \mathbf{w}_G\big)^{-1}, \tag{3.8}$$

with $b_0 = -\mathbf{w}_D^T \mathbf{m}$. Point prediction of x^* can be done with (3.7), which takes the form of a linear discrimitive predictor, entailing the scalar product between discriminative weights $\mathbf{w}_D$ and the input image $\mathbf{t}^*$.

The procedure of inverting the model to predict x^* is displayed in Fig. 3.2, with a cartoon illustration. It entails projecting $\mathbf{t}^*$ orthogonally onto the direction of $\mathbf{w}_D$, to retrieve the signal of interest in presence of noise with covariance $\mathbf{C}$. Note that the direction of $\mathbf{w}_D$ can be very different from $\mathbf{w}_G$, since $\mathbf{w}_D$ also accounts for the noise structure. For example, in Fig. 3.2, $\mathbf{w}_D$ obtains a large y-component, despite $\mathbf{w}_G$ has zero weight in that direction. This gives insight into the problems that arise when trying to interpret the discriminative weights $\mathbf{w}_D$. We will explore this topic in more details in section 4.2.

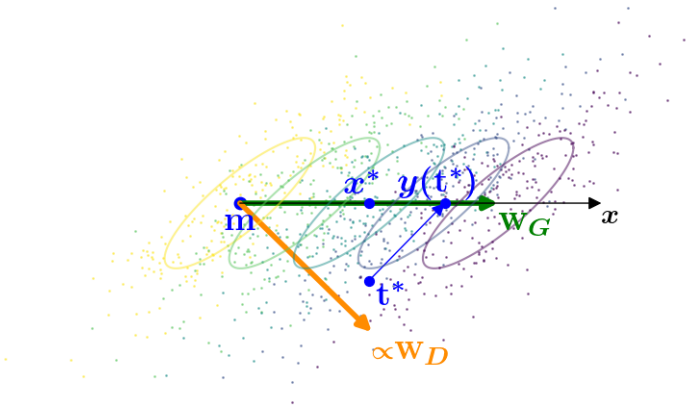


Fig. 3.2: Cartoon illustration of inverting the model to predict x^* based on $\mathbf{t}^*$. by projecting $\mathbf{t}^*$ orthogonally onto the direction of $\mathbf{w}_D$.

3.3 Model training

In practice, the model parameters $\mathbf{W}$ and $\mathbf{C}$ are not known and need to be learned from training data. Assume that we have N training pairs $\{\mathbf{t}_n, x_n\}_{n=1}^N$, the marginal likelihood of these training data is given by:

$$p\left(\{\mathbf{t}_n\}_{n=1}^N | \{x_n\}_{n=1}^N, \mathbf{W}, \mathbf{C}\right) = \prod_{n=1}^N \mathcal{N}\left(\mathbf{t}_n |\ \mathbf{m} + x_n \mathbf{w}_G, \mathbf{C}\right). \tag{3.9}$$

The maximum likelihood (ML) estimate of $\mathbf{W}$ and $\mathbf{C}$ can be then found by maximizing (3.9) with respect to these parameters.

For $\mathbf{W}$, the ML estimate is given in closed form:

$$\mathbf{W} = \left(\sum_{n=1}^N \mathbf{t}_n \boldsymbol{\phi}_n^T\right)\left(\sum_{n=1}^N \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T\right)^{-1} \quad \text{with } \boldsymbol{\phi}_n = (1, x_n)^T. \tag{3.10}$$

Note that estimating $\mathbf{W}$ with (3.10) equals to fitting an Ordinary Least Squares regression model independently in each voxel.

Estimating the noise covariance matrix $\mathbf{C}$ is troublesome, since it has $J(J+1)/2$ free parameters, where J is the number of voxels, and this is larger than

usual training set sizes by orders of magnitude. To elude this problem, we constrain $\mathbf{C}$ to have a specific structure, using a latent variable model known as factor analysis Bishop and Nasrabadi (2006a). This allows us to control the number of parameters while still capturing the dominant correlations in the data. Specifically, we consider a structured noise of the form:

$$\boldsymbol{\eta} = \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon}, \tag{3.11}$$

where $\mathbf{z}$ is a vector of K unknown latent variables with prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbb{I}_K)$, $\mathbf{V}$ includes the corresponding, unknown weights, and $\boldsymbol{\epsilon}$ is a Gaussian error of the form

$$p(\boldsymbol{\epsilon}|\boldsymbol{\Delta}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Delta}), \tag{3.12}$$

where $\boldsymbol{\Delta}$ is unknown and diagonal.

Integrating over $\mathbf{z}$, the noise vector $\boldsymbol{\eta}$ is still distributed as a zero-mean Gaussian with covariance matrix $\mathbf{C}$, while $\mathbf{C}$ is now given by

$$\mathbf{C} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Delta},$$

and is now controlled by a smaller set of parameters $\mathbf{V}$ and $\boldsymbol{\Delta}$. The number of latent variables K, which is also the number of columns in $\mathbf{V}$, is the only hyperparameter of the model, which needs to be determined experimentally.

The parameters $\mathbf{V}$ and $\boldsymbol{\Delta}$ can now be estimated, by maximizing the data marginal likelihood in (3.9). We can do this by plugging in the ML estimate of $\mathbf{W}$ given by (3.10), and then using the Expectation-Maximization (EM) algorithm Rubin et al. (1982). Defining $\tilde{\mathbf{t}}_n = \mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}_n$, this results in an iterative algorithm that repeatedly evaluates the posterior distribution over the latent variables:

$$p(\mathbf{z}_n|\tilde{\mathbf{t}}_n, \mathbf{V}, \Delta) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}) \tag{3.13}$$

with $\boldsymbol{\mu}_n = \boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\tilde{\mathbf{t}}_n$ and $\boldsymbol{\Sigma} = (\mathbb{I}_K + \mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V})^{-1}$, and subsequently updates the parameters:

$$\mathbf{V} \leftarrow \left(\sum_{n=1}^{N} \tilde{\mathbf{t}}_n \boldsymbol{\mu}_n^T\right)\left(\sum_{n=1}^{N} \left(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + \boldsymbol{\Sigma}\right)\right)^{-1} \tag{3.14}$$

$$\boldsymbol{\Delta} \leftarrow \operatorname{diag}\left(\frac{1}{N}\sum_{n=1}^{N} \tilde{\mathbf{t}}_n\tilde{\mathbf{t}}_n^T - \mathbf{V}\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\mu}_n\tilde{\mathbf{t}}_n^T\right). \tag{3.15}$$

The notation $\operatorname{diag}(\cdot)$ means that all non-diagonal entries are set to zero. We detect convergence of the EM algorithm by looking at the relative change in marginal likelihood between iterations. After converge, we have an estimate of all the unknown parameters in the model, and we can then proceed to make predictions on unseen data, as described in section 3.2.

Chapter 4

Results on UK Biobank

In this chapter, we present experiments where the proposed method is employed for age and gender prediction on the UK Biobank data, and we compare the obtained results with selected benchmark methods. In particular, the chapter is structured as follow:

- We first show prediction performances of proposed method and benchmarks for age and gender prediction, on the UK Biobank data;
- We explore the interpretability of the proposed method vs discriminative benchmarks; and
- We then give insight into the considered methods' performances in terms of trade-off between bias and variance.

4.1 Prediction performances and benchmarks

In these experiments, we employed the UK Biobank dataset (Sudlow et al., 2015; Alfaro-Almagro et al., 2018), a huge prospective study that aims at scanning 100.000 subjects. The data release used for these experiments includes 42,180 T1-weighted MRI scans, which after some exclusion criteria (selection of healthy

subjects and baseline scans) yielded MRI scans of 26,127 healthy subjects, aged 44-82 years. We employed these data for a regression task (age prediction) and a classification task (gender prediction), on healthy subjects.

To assess if the proposed method achieves competitive results, we compared its performances with three other prediction methods. Since we aim at investigating properties of nonlinear vs linear and discriminative vs generative models, we selected as benchmarks one method of each type: one discriminative nonlinear method (SFCN Peng et al. (2021)), one discriminative linear model (RVoxM Sabuncu et al. (2012)), and one generative nonlinear model (variational auto-encoder Zhao et al. (2019)).

SFCN: This is a lightweight convolutional neural network proposed in Peng et al. (2021) for brain age prediction. We selected this method as discriminative nonlinear benchmark since it achieves, to the best of our knowledge, state-of-the-art performances for age prediction, and it also won the 2019 Predictive Analysis Challenge for this task. In Peng et al. (2021), the authors train the SFCN on the UK Biobank for age and gender prediction, based on training sets of different sizes. We used a similar setting for our experiment and trained the SFCN and the other selected models on the UK Biobank. We followed the training regime described in Peng et al. (2021), with only the number of epochs used for training left as hyperparameter of the method.

RVoxM: This is a linear discriminative method, proposed in Sabuncu et al. (2012), which we selected as benchmark since its performances are competitive within the class of linear discriminative models. The method provides weight maps that are sparse and spatially smooth as a form of regularization, and the strength of this spatial smoothness is the one hyperparameter of the method.

Variational auto-encoder: We employed a variational auto-encoder (VAE) proposed for age prediction in Zhao et al. (2019) as nonlinear generative benchmark. This model can be seen as a nonlinear version of the proposed method, where its latent variables are decoded nonlinearly through a deep neural network. This model has two hyperparameters, which control the amount of regularization (dropout factor and L2 regularization). In Zhao et al. (2019), the VAE is trained on T1 scans that are cropped around the ventricles, therefore we considered the same setting and compared performances of the proposed model and VAE, both trained on cropped T1 scans.

In the next two sections, we first show the comparison of prediction performances obtained by the proposed method, RVoxM and SFCN on whole T1 scans; then, we present the comparison with VAE in a separate experiment, where both methods are applied on T1 scans cropped around the ventricles.

4.1.1 Comparison with discriminative benchmarks: SFCN and RVoxM

We first compared prediction performances of the proposed method, SFCN and RVoxM for age and gender prediction, based on T1-weighted scans from the UK Biobank. The dataset provides skull-stripped, bias-field corrected T1 scans in subject space, and both an affine and a deformable transformation from subject space to MNI space, which can be used to spatially normalize the images. In Peng et al. (2021), the authors mostly use T1 scans that are affinely registered to template space. However, since the deformable transformation is available and already used for skull-stripping, we deemed more meaningful to use nonlinearly registered T1 scans in our experiments. We anyway also employed the affinely registered T1s, to be able to compare results with Peng et al. (2021). In the remainder, we will call "affine T1s" and "deformable T1s" the scans registered with the affine and deformable transformations, respectively.

We aimed at comparing prediction performances of the considered methods in different training scenarios, therefore we considered training sizes from 100 subjects up to 7800 (for age) and 9800 subjects (for gender), resembling the setting in Peng et al. (2021). In all the experiments, we employed a validation set of 500 subjects, and a test set of 1000 subjects. We performed hyperparameter selection using grid search on the validation set, with validation MAE (for age) and validation accuracy (for gender) as metrics to be optimized. For each training size, we used 10 randomly sampled training sets, (only 3 sets for sizes strictly larger than 1000 subjects), and averaged the obtained prediction metrics on the test set, to get more robust results.

Fig. 4.1 shows the obtained results for age prediction, in terms of MAE and Pearson correlation coefficient between real and predicted values, and Fig. 4.2 displays results for gender classification, in terms of test accuracy. Comparing performances obtained on affine vs deformable T1s, we observe that the proposed method and RVoxM achieve clearly worse results on affine T1s, while the effect of input data type on prediction performances is much less pronounced for SFCN, with almost no effect for age, and a small difference for gender prediction. These findings are not surprising, since the proposed model and RVoxM are linear methods, and therefore they cannot model nonlinear deformations that have not been removed from the input images by the the affine registration, while the SFCN has the capability of doing it. However, since these nonlinear deformations are known and actually used to compute the affine T1s, we deem that it is more meaningful to consider results on deformable T1s. Comparing performances of the proposed method and RVoxM, we observe that their results for age and gender prediction on both affine and deformable T1s are comparable up to a few thousand of training subjects, after which the RVoxM starts out-

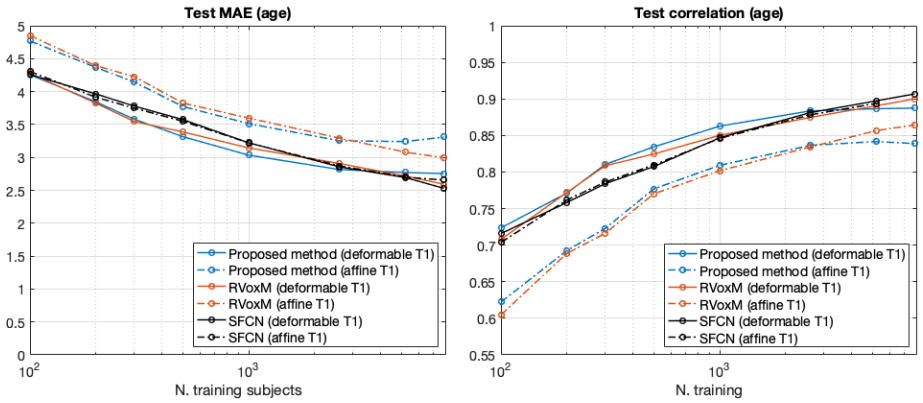


Fig. 4.1: Comparison of performances for age prediction, for the proposed method, RVoxM and SFCN. Results are shown on both affine and deformable T1s, for all methods.

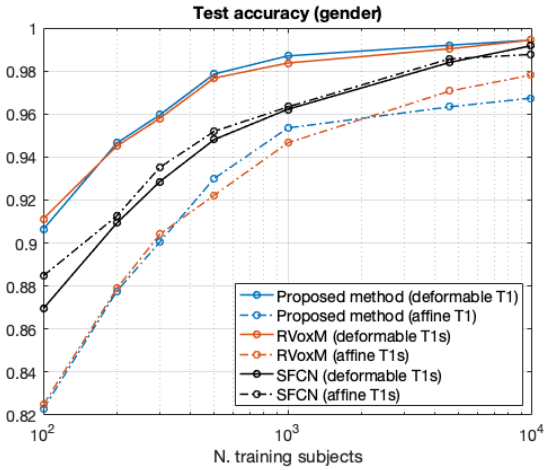


Fig. 4.2: Comparison of performances for gender classification, for the proposed method, RVoxM and SFCN. Results are shown on both affine and deformable T1s, for all methods.

performing the proposed method (except for gender prediction on deformable T1s, where the two methods achieve equal performances even at N=9800). Regarding the comparison between the proposed method and SFCN, we find that, on affine T1s, the SFCN achieves better results for each training size, for both age and gender prediction. When using deformable T1s, the scenario is very

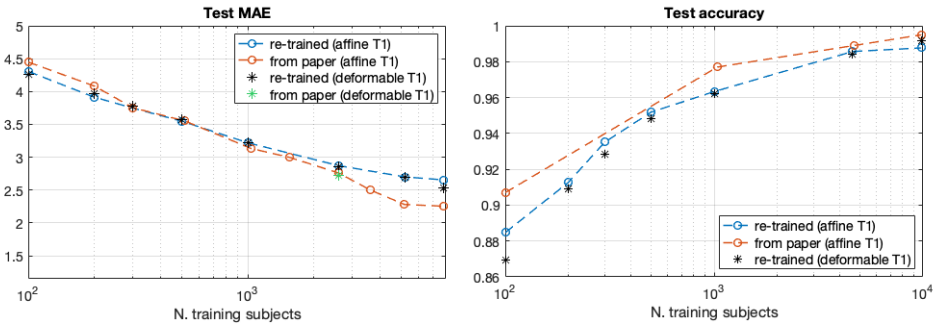


Fig. 4.3: Comparison of performances obtained by SFCN for age prediction (left) and gender prediction (right) when re-trained vs as reported in Peng et al. (2021).

different: for age prediction, the proposed method yields comparable or better results in regimes up to 2600 training subjects, after which it is outperformed by the SFCN. For gender prediction, we find that the proposed method outperforms the SFCN for all training sizes, apart from $N = 9800$ where they perform comparably.

In order to make sure to perform a fair comparison of proposed method vs SFCN, we also checked if the results we obtained by re-training the SFCN are in line with the ones reported in Peng et al. (2021) (cf. Fig. 4.3). For age prediction, we are reproducing the results reported in Peng et al. (2021) for up to $N = 1000$, with a small variability caused by the use of different data, whereas for larger sizes, the errors that we obtain are systematically worse than in Peng et al. (2021). However, this does not change the conclusions drawn from Figure 4.1 about the comparison of proposed method and SFCN for age prediction. Regarding gender prediction, there is a systematic difference between our results and the ones from Peng et al. (2021), with a larger gap for $N = 100$ and $N = 1000$. In this regard, it is worth reminding that we used different training and test set as compared to Peng et al. (2021), and that their results, obtained on only one training set for each size, may depend a lot on the choice of training data when N is small. Anyway, considering both Fig. 4.2 and 4.3, we could conclude that SFCN for gender prediction on deformable T1s performs comparably to the proposed method for each training size.

To give an idea of the usability of the considered methods, we show in Table 4.1 training times for proposed method, RVoxM and SFCN, for age prediction on deformable T1s. The times reported in the table are training times for a single selected value of the models' hyperparameter (for SFCN, this is the training time

	N=100	N=200	N=300	N=500	N=1000	N=2600	N=5200	N=7800
Proposed method	1.20 min	0.67 min	1.94 min	9.53 min	32.18 min	≈3h	≈15h	≈ 69 h
RVoxM	92.42 min	66.46 min	75.36 min	76.21 min	129.05 min	126.55 min	≈ 22 h	≈ 21 h
SFCN	≈ 8h	≈ 11 h	≈ 16 h	≈ 18 h	≈ 34h	≈ 76h	≈ 69h	≈ 102 h

Table 4.1: Training times of proposed method, RVoxM and SFCN, for age prediction on deformable T1s, averaged across all training runs. For proposed method and RVoxM, the reported times are CPU times obtained with Matlab on a state-of-the-art desktop, while for SFCN they are obtained with a NVIDIA A100 SXM4 GPU (40 GB of RAM).

	N=100	N=200	N=300	N=500	N=1000	N=2600	N=5200	N=7800
Proposed method	19.80	20.40	52.00	86.00	120.00	366.67	1833.33	3333.33
RVoxM	24950	18250	23400	19400	31000	31666.67	200000	200000
SFCN	218.10	217.10	221.50	248.33	226.50	262.67	146.00	123.67

Table 4.2: Values of hyperparameters selected on the validation set for proposed method, RVoxM and SFCN, averaged across all runs, for age prediction on deformable T1s.

up to the selected epoch). We observe that for small N, the proposed method is much faster than the other models, with training times of only a few minutes. With 2600 and 5200 training subjects, our method becomes comparable to the RVoxM, while for $N = 7800$ it is slower. Furthermore, training the SFCN takes more time than the other models, for each size. When performing this comparison, we need to take into account that for proposed method and RVoxM these are CPU times with 3mm T1 scans, while for SFCN they are GPU times with 1mm T1s, and that training the SFCN in the same setting as described in Peng et al. (2021) required to use a GPU with particularly large memory (40 GB RAM).

Furthermore, Table 4.2 displays the selected values of the methods' hyperparameters, averaged across all runs of each size, for age prediction on deformable T1s. We observe that for the proposed method, the number of latent variables on average increases monotonically as the training set size grows, as expected since larger sizes allow more flexible models. Conversely, for RVoxM and SFCN, the hyperparameter behaviour is not monotone, which can be due to different hyperparameter values giving similar results.

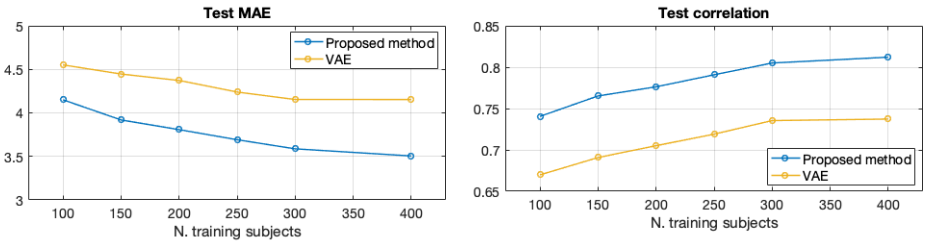


Fig. 4.4: Test MAE and Pearson correlation coefficient obtained by the proposed method and VAE for age prediction, on deformable T1s cropped around the ventricles.

4.1.2 Comparison with generative benchmark: VAE

We then compared results obtained by the proposed method with the generative benchmark (VAE), for age prediction, on the UK Biobank. Since in Zhao et al. (2019), the method is applied to T1 scans cropped around the ventricles, we trained the proposed method on the same input type, to perform a fair comparison. In this experiment, we used only deformable T1s, and the same validation and test set of 500 and 1000 subjects respectively as in the previous experiment. We again selected the models' hyperparameters with grid search, by optimizing the validation MAE. Since the VAE was used in Zhao et al. (2019) with around 200 training subjects, with the number of latent variables chosen accordingly, we considered in this experiment only training sets of similar sizes (from 100 to 400 subjects). For each training size, we employed 10 randomly sampled training sets, and averaged the obtained test MAEs and correlations, for both proposed method and VAE. In Fig. 4.4 we show the results: the proposed method achieves better results than the VAE for each considered training size. These findings suggest that, at least for training sets of a few hundred subjects, including more flexibility into the model with the nonlinear expansion of latent variables is not advantageous, and it may even decrease performances.

Concerning training times, the VAE training took on average 9.40 minutes for N=200, using a NVIDIA GeForce RTX 2080 Ti GPU (11 GB of RAM), while training the proposed method took on average 1.16 minutes with the same size, using Matlab on a state-of-the-art desktop. In both cases, these are training times for the selected values of the hyperparameters.

Additionally, Table 4.3 displays which values of the hyperparameters are selected on the validation set on average, for both proposed method and VAE. For the proposed method, the number of latent variables is almost always in-

	N=100	N=150	N=200	N=250	N=300	N=400
Proposed method	16.28	21.78	25.56	49.11	43.06	49.17
VAE (dropout, L2 regularization)	(0.64, 0.31)	(0.61, 0.24)	(0.57, 0.14)	(0.57, 0.07)	(0.63, 0.03)	(0.54, 0.12)

Table 4.3: Values of hyperparameters selected on the validation set for proposed method and VAE, averaged across all runs.

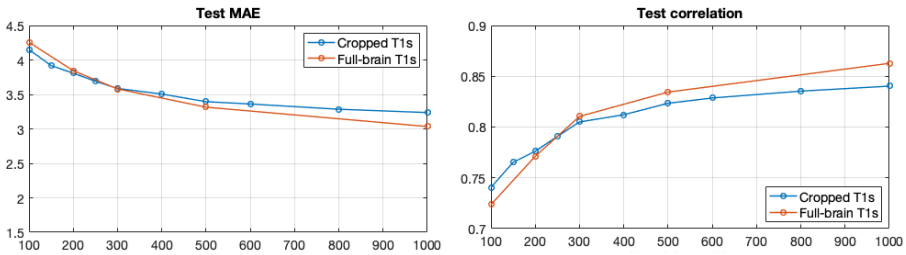


Fig. 4.5: Performances of proposed method for age prediction on T1s cropped around the ventricles vs full-brain T1s (both deformable). Results are averaged across 10 different training sets for each size.

creasing with the training set size, as already observed for the comparison with discriminative benchmarks. As opposed to Table 4.2 where the trend of the latent space size was monotone, here there is a fluctuation for $N = 300$, probably due to the smaller gap between the considered training sizes. For the VAE, we note that the dropout factor (fraction of units to drop) and L2 regularization are almost always decreasing with the training set size, which is expected since larger training sizes require a smaller degree of regularization.

As a side experiment, we can compare the proposed method's performances for age prediction on deformable full-brain T1s (used when comparing with discriminative benchmarks) vs deformable T1s cropped around the ventricles (used when comparing with VAE). Results are shown in Fig. 4.5, for training sizes from 100 to 1000 subjects. We observe that for very small training sets, performances on cropped T1s are better than on full-brain T1s, while for training sizes bigger than around 300 subjects, the roles are reversed. A possible explanation is that increasing the input dimensionality when the training size is very small raises the chance of overfitting and therefore of worse results, considering also that the ventricle area is already informative. Including more input features may become beneficial only if there are enough data to exploit the additional information.

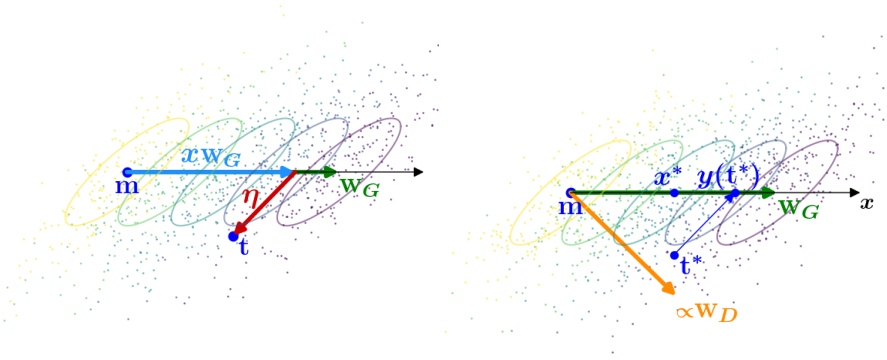


Fig. 4.6: Cartoon illustration of signal decomposition $\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \boldsymbol{\eta}$ (left), and inversion of the model to make predictions (right).

4.2 Interpretability

One of the main perks of the proposed method is that it produces an interpretable spatial map ($\mathbf{w}_G$), showing the direct effect of the variable of interest on image intensities, on a population level. This map is obtained through a decomposition of the signal (image) into an average anatomy, the effect of the variable of interest and a subject-specific noise. This decomposition is illustrated in Fig. 4.6 in a 2D toy example, and in Fig. 4.7 with images, for the age prediction task. Thanks to the form of this decomposition, the generative map $\mathbf{w}_G$ expresses how the variable of interest affects a subject's image, encoding target-related neuroanatomical changes Haufe et al. (2014). Consistently with this, the generative map for age displayed in Fig. 4.7 expresses known age-related effects, such as gray matter atrophy and enlargement of ventricles Fjell et al. (2009); Fjell and Walhovd (2010).

When the model is subsequently inverted, the discriminative weight map $\mathbf{w}_D$ (3.4) is obtained by combining the generative maps with the noise covariance matrix, and used to make predictions through a scalar product with the test subject's image, as illustrated in Fig. 4.6 for vectors and Fig. 4.7 for images. This discriminative map contains the weights given to voxels for predicting the variable of interest, and therefore highlights image areas the model uses for predictions. However, since it includes both generative effect and noise pattern, it does not directly express target-related changes in neuroanatomy and it results in an uninterpretable spatial pattern Haufe et al. (2014). This concept is shown in Fig. 4.6, illustrating how the y channel has a large component in $\mathbf{w}_D$, although it is not affected by the target variable since its weight in $\mathbf{w}_G$ is zero. Similarly,

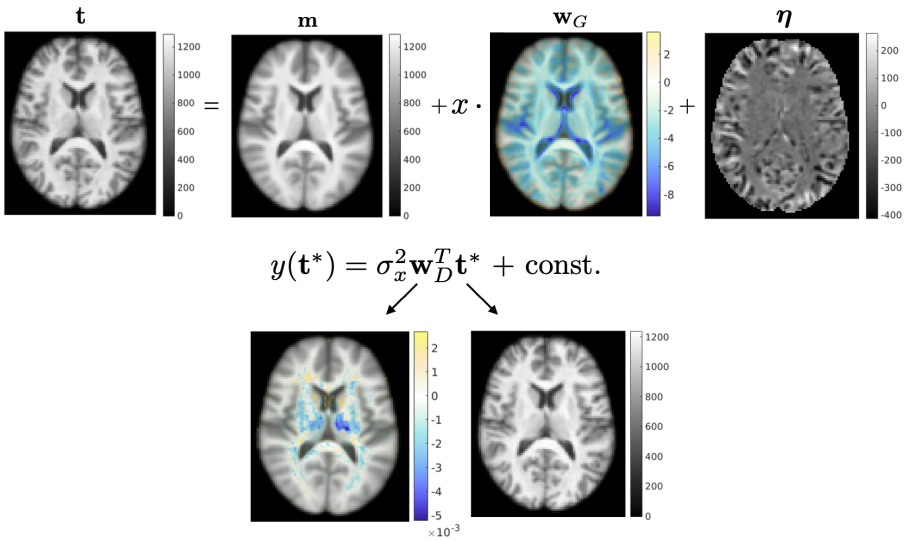


Fig. 4.7: Discriminative linear regression (bottom) is *mathematically* the same as decomposing the signal into its constituents in the model (top), but *not in terms of interpretability*. Maps obtained for age prediction on the UK Biobank, on a training set of 300 subjects (with deformable T1s).

the age discriminative map shown in Fig. 4.7 does not present the typical age-related patterns that characterize the generative map, but it mostly highlights white matter areas.

Fig. 4.8 shows other 2D slices of both generative and discriminative maps obtained for age prediction. We observe again the large difference between the two maps: while the generative maps display typical age-related effects, mostly highlighting gray matter borders and ventricles, the discriminative maps focus on very different areas, mainly within white matter, with both positive and negative weights.

It is also interesting to compare the generative map $\mathbf{w}_G$ not only with its discriminative counterpart $\mathbf{w}_D$, but also with spatial maps of the other discriminative methods employed as benchmarks. Additionally, we are interested in analyzing the stability of these spatial maps when changing training set, both of same and different sizes. For these reasons, we display in Fig. 4.9 spatial maps of the proposed method (both $\mathbf{w}_G$ and $\mathbf{w}_D$), RVoxM and SFCN, obtained on training sets of 300, 2600 and 7800 subjects. To also investigate the behaviour of these maps when changing training data within the same cohort size, in Fig. 4.10 we

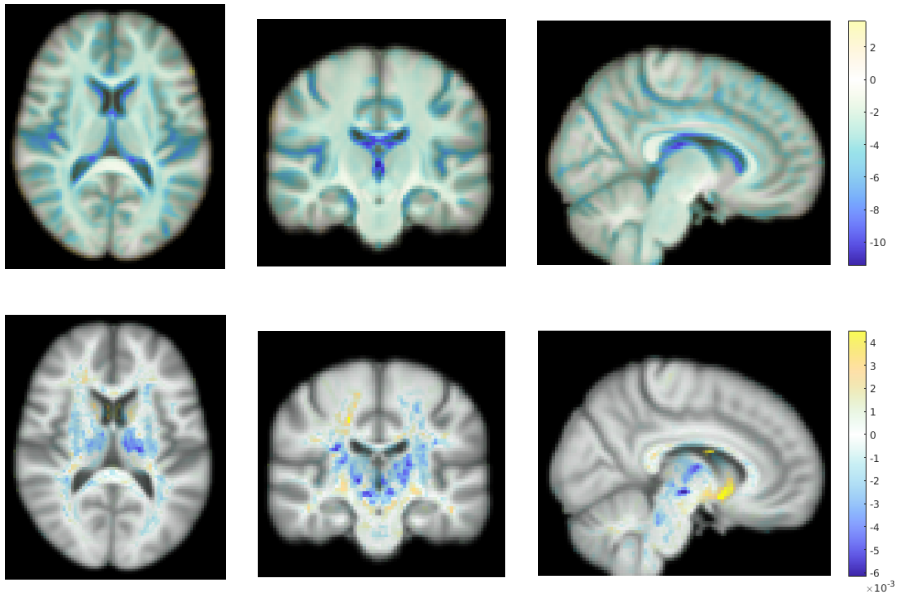


Fig. 4.8: 2D slices of generative map $\mathbf{w}_G$ (top) and discriminative map $\mathbf{w}_D$ (bottom), obtained for age prediction on UK Biobank data, on a training set of 300 subjects (with deformable T1s).

display maps obtained by the methods on three randomly sampled training sets of 2600 subjects. Since SFCN is a neural network and therefore does not automatically provide spatial maps, we used SmoothGrad Smilkov et al. (2017b) to compute saliency maps for this method. The SmoothGrad maps are commonly used as post-hoc explanations of deep learning models, and they can be seen as a generalization of linear methods' weight maps (i.e. the SmoothGrad map computed for a linear discriminative method would correspond exactly to the model's weight map Adebayo et al. (2018)). Note that, since SmoothGrad provides subject-specific maps, in order to obtain a single template that could be compared with the other methods' ones, we averaged the SmoothGrad maps of all test subjects. This technique is considered a relevant way to produce population-level maps for instance-based XAI methods, since it removes the noise characterizing single-subject maps Wilming et al. (2022), consistent with the finding that it's the aggregate use of saliency maps rather than the individual one that can yield significant results Ghassemi et al. (2021).

First, from Fig. 4.9 and 4.10, we note that the discriminative spatial patterns of $\mathbf{w}_D$, RVoxM, and SFCN are much less intuitive than $\mathbf{w}_G$, which shows known age-related effects. These findings seem to support what we discussed in Chapter

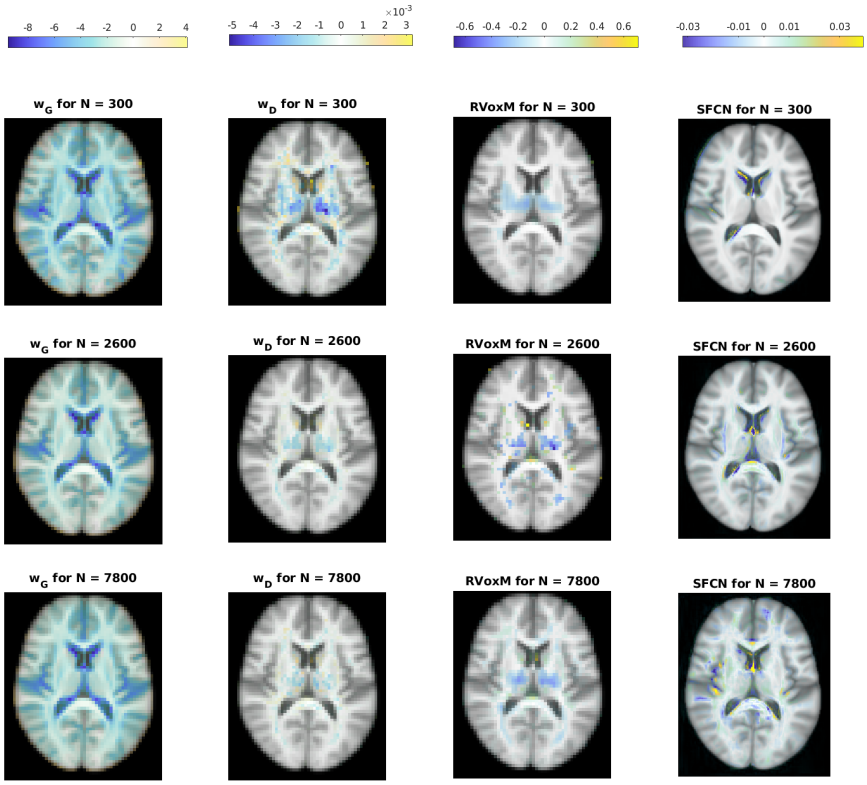


Fig. 4.9: Maps of proposed method ($\mathbf{w}_G$ and $\mathbf{w}_D$), RVoxM, and SFCN (with SmoothGrad), for different training set sizes. Voxels with zero weight are transparent. Discriminative maps are displayed for the optimal value of the hyperpameter, selected as described in section 4.1.

2, namely that discriminative maps highlight regions that are most important to make predictions, but they do not directly express anatomical changes caused by the variable of interest, and this holds even in the apparently simple case of a linear discriminative method Haufe et al. (2014); Wilming et al. (2022). Additionally, regarding the dependency of maps on specific training data, we note that the generative maps $\mathbf{w}_G$ are quite stable across different training samples. The discriminative maps $\mathbf{w}_D$ and RVoxM's also show some consistency, especially when keeping the same training set size, but with more differences than the generative maps. A possible explanation is that $\mathbf{w}_D$ and the RVoxM's maps depend on an hyperparameter, which in general varies with the specific training data, especially when changing training size (cf. Table 4.2), while $\mathbf{w}_G$ does not. Additionally, generative maps are produced by estimating two weights from N

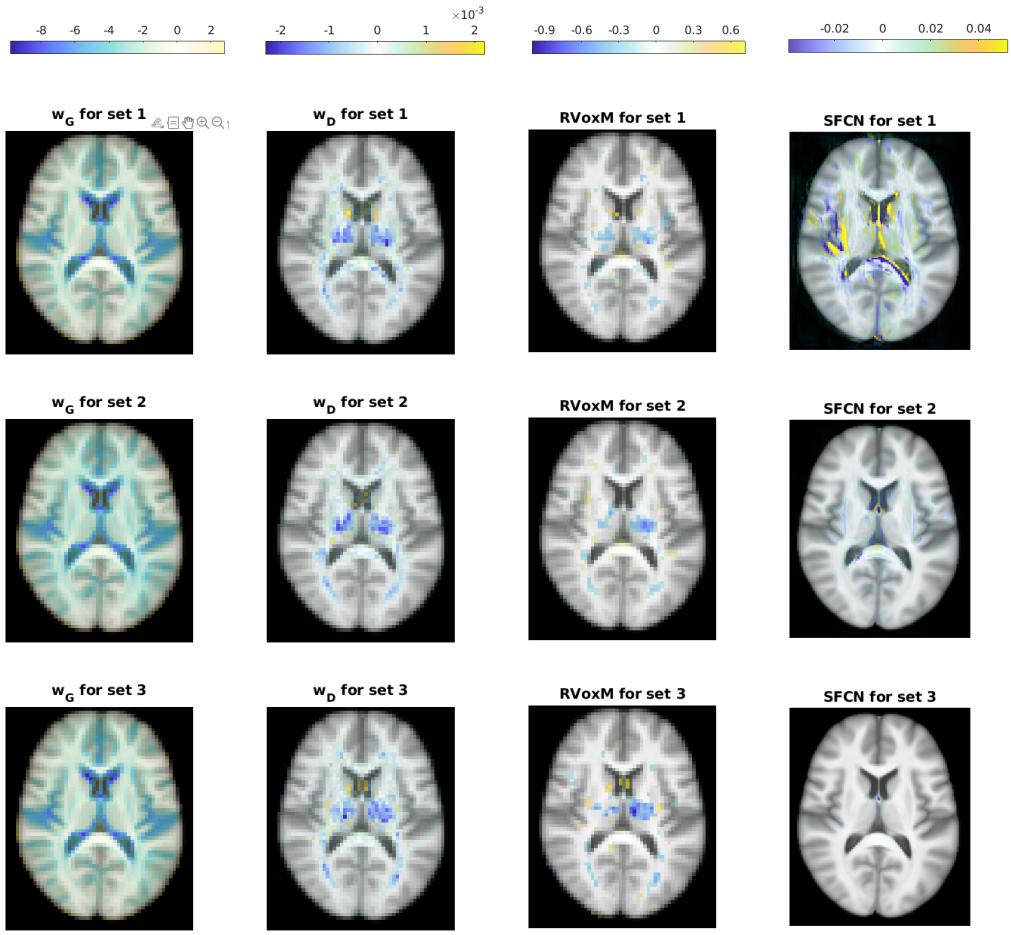


Fig. 4.10: Maps of proposed method ($\mathbf{w}_G$ and $\mathbf{w}_D$), RVoxM, and SFCN (with SmoothGrad), computed on 3 different training sets of 2600 subjects. Voxels with zero weight are transparent. Discriminative maps are displayed for the optimal value of the hyperpameter, selected as described in section 4.1.

data point in each voxel independently, which is expected to yield rather stable fittings. Instead, $\mathbf{w}_D$ involves estimating many more basis functions, together with their coefficients, in a multivariate way, and RVoxM's maps entail fitting a very high-dimensional hyperplane from N data - which are both likely to be less stable operations. Regarding SFCN, we observe that there is a huge variability in its maps when changing training sets, both of same and different sizes. These

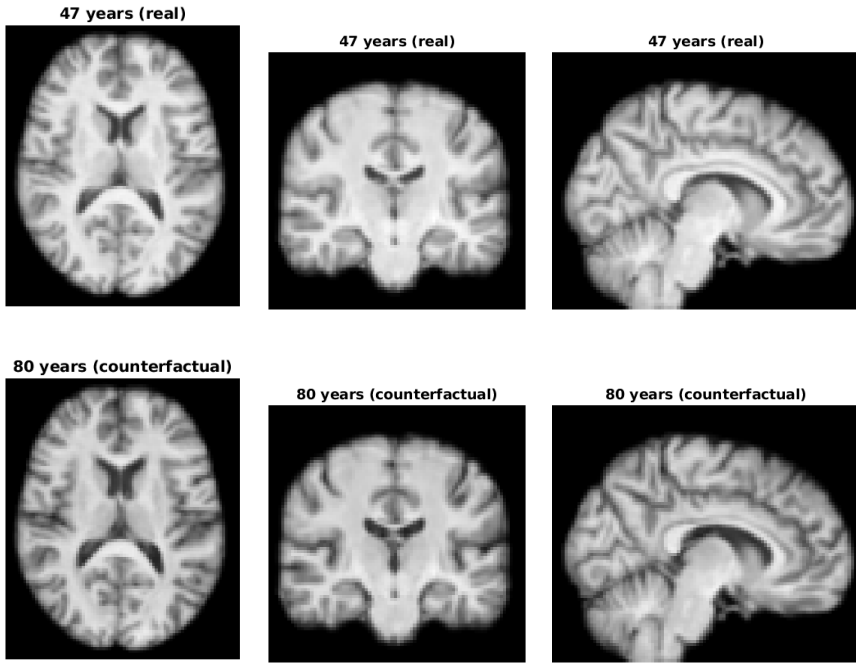


Fig. 4.11: Top: original images of a 47 years old subject. Bottom: counterfactual images of the same subject at the age of 80. All images are on the same intensity scale.

findings seem consistent with Arun et al. (2021), which shows that many commonly used saliency maps methods, including SmoothGrad, did not pass a test of reproducibility when the model is retrained with a different random initialization or with a different architecture yielding similar prediction performances. All these results illustrate some difficulties that arise in the interpretation of discriminative maps - both *theoretical* difficulties, since such maps do not express the causal effects of interest for interpretation purposes, and *empirical* ones, since especially SFCN maps rely heavily on the specific choice of training data.

Like all generative models, the proposed method can also generate counterfactual images. Figure 4.11 displays an example of counterfactual image generated by the proposed method, where the brain of a 47 years old subject is artificially aged to 80 years. The aging patterns shown in the counterfactual image are consistent with the effects encoded by $\mathbf{w}_G$: the aged brain is characterized by larger ventricles, enhanced gray matter atrophy, and a general slight decrease in image intensities. Counterfactual images are useful for showing target-related

effects at a subject-specific level. Furthermore, Pearl and Mackenzie (2018) illustrates that counterfactual reasoning is typical of human thinking: answering counterfactual questions is the highest degree of human casual inference, and the advantages that humans have had in evolution from being able to do it are huge. This seems to suggest that the way a generative model captures target-related effects is similar to the human perception of causation.

It is interesting to note that any generative model can generate counterfactual images, showing target-related effects on specific subjects. However, the proposed method has the big advantage of providing spatial maps of target-related patterns on a population level, which may be more difficult to obtain with complex nonlinear generative models such as the ones proposed in Zhao et al. (2019); Wilms et al. (2020). For instance, nonlinear generative methods can produce age-specific templates, from which global aging effects can be retrieved by computing the jacobian determinant of the deformation between pairs of templates, as in Wilms et al. (2020); Zhao et al. (2019). However, producing such maps requires a lot of computation, and depends on the selected age-gap. Furthermore, Wilms et al. (2020) also uses another technique to generate a population-level interpretable spatial map, by computing the partial derivative of the model's inverse map with respect to age. But again, this technique involves heavy computations, and it provides a less intuitive spatial pattern Wilms et al. (2020). As opposed to these nonlinear generative models, it is straightforward for the proposed method to produce population-level spatial maps, which can be computed in sub-second speed.

To further investigate how the proposed model works, we can also compute the eigenvectors of the noise covariance matrix $\mathbf{C}$. Fig. 4.12 shows the effect of the first three eigenvectors. We notice that the first eigenvector expresses a scaling factor in image intensities, the second one encodes bias field in the top part of brain that has not been removed in the data pre-processing, and the third one encodes differences in the lateral ventricles size. It is worth noting that that the bias-field is modelled by the proposed method in the noise component and therefore it is disentangled from the signal of interest. Conversely, discriminative methods would implicitly consider it when estimating their spatial weights, therefore relying on an intensity pattern that is unrelated to the task at hand.

4.3 Bias-variance trade-off

In section 4.1, we showed that the proposed method and the two discriminative benchmarks perform quite similarly on age prediction based on deformable T1s, in training regimes up to a few thousands of subjects. This finding is perhaps

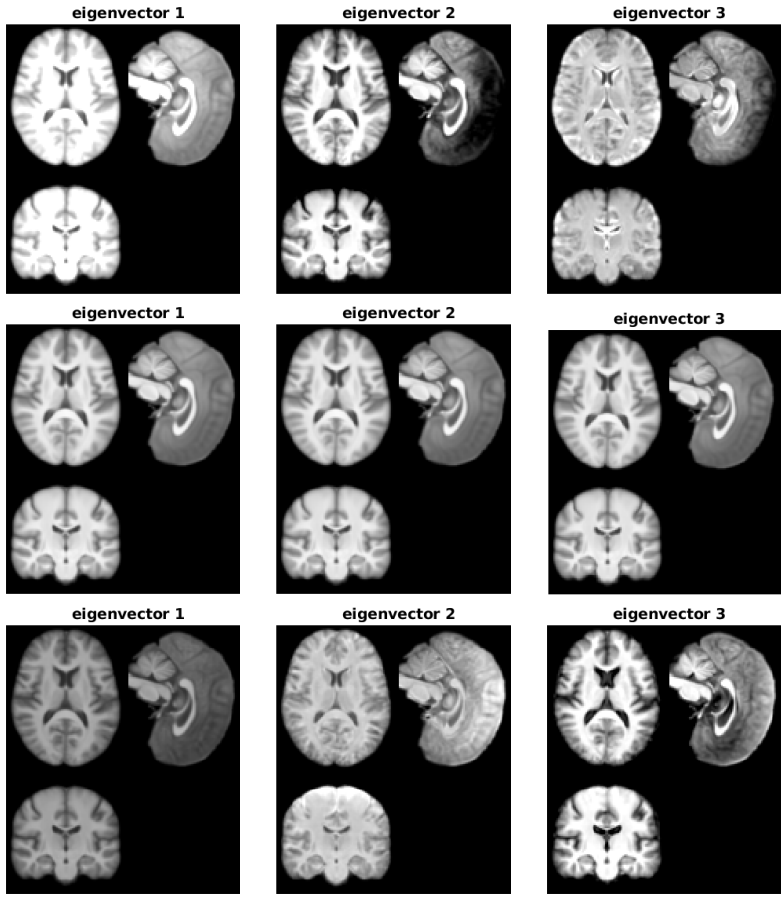


Fig. 4.12: Effect of the first three eigenvectors of the covariance matrix $\mathbf{C}$. The middle line displays slices of the average volume. The top ad bottom lines show the average volume modified in direction of the eigenvectors, with negative and positive sign respectively.

surprising if we consider the vastly different numbers of parameters in the methods. In fact, on one hand, the proposed method has $J(K+3) - K(K-1)/2$ free parameters (2 columns of J elements in $\mathbf{W}$, K columns of J parameters in $\mathbf{V}$ and J diagonal elements in $\mathbf{\Delta}$, which are reduced by $K(K-1)/2$ because any rotation in the latent space provides the same model Bishop and Nasrabadi (2006a)), with $J \approx 80,000$ in our experiments and K that varies from tens to thousands, depending on the training set size (Cf. Table 4.2). On the other hand, the RVoxM has J free parameters, and the SFCN has 3 million parameters Peng et al. (2021).

However, several factors determine a method's prediction performances besides the number of parameters, and it is in general not straightforward to predict which method will perform best a priori. For instance, the choice of the optimizer and how it explores the hypothesis space has an impact on a method's performances, where optimizers that try fewer hypothesis, restrict the hypothesis space and act as regularizers Domingos (2012). Additionally, the posterior distribution of a "wrong" generative model can still give correct predictions Domingos et al. (1997), and even on simulated data, an incorrect model can achieve better prediction performances than the "true" model in certain regimes Domingos (2012), making an a priori guess of the best-performing method extremely hard.

Given these difficulties, to gain more insight into the performances of considered methods, we can perform the so-called bias-variance decomposition of prediction errors. We compute it for age prediction, since it is more straightforward for a continuous variable than in a classification case. For a given method, the prediction mean squared error (MSE) can be decomposed into a *bias* term, which denotes how well the method performs on average, and a *variance* term, which indicates how much predictions for the same test subjects change across different training runs Hart et al. (2000); Bishop and Nasrabadi (2006b).

In particular, if we consider a test pair $(\mathbf{t}^*, x^*)$, and we denote with $y(\mathbf{t}^*; D)$ the prediction made by the model trained on a dataset D for test subject $\mathbf{t}^*$, we obtain the following decomposition:

$$\underbrace{\mathbb{E}_{\mathbf{t}^*,D}\left[(x^* - y(\mathbf{t}^*; D))^2\right]}_{MSE} = \underbrace{\mathbb{E}_{\mathbf{t}^*}\left[(x^* - \mathbb{E}_D\left[y(\mathbf{t}^*; D)\right])^2\right]}_{bias} + \underbrace{\mathbb{E}_{\mathbf{t}^*,D}\left[(y(\mathbf{t}^*; D) - \mathbb{E}_D\left[y(\mathbf{t}^*; D)\right])^2\right]}_{variance} \quad (4.1)$$

where $\mathbb{E}_D[\cdot]$ denotes the expected value over all training sets D of a fixed size, and $\mathbb{E}_{\mathbf{t}^*}[\cdot]$ denotes the expected value over all possible inputs $\mathbf{t}^*$. In practice, if we consider M test pairs $\{\mathbf{t}^*_m, x^*_m\}_{m=1}^M$, and B different training sets $\{D_b\}_{b=1}^B$ of a given size, we can write for test subject m:

$$\frac{\sum_{b=1}^B \left(x^*_m - y(\mathbf{t}^*_m; D_b)\right)^2}{B} = \left(x^*_m - \bar{y}(\mathbf{t}^*_m)\right)^2 + \frac{\sum_{b=1}^B \left(y(\mathbf{t}^*_m; D_b) - \bar{y}(\mathbf{t}^*_m)\right)^2}{B} \quad (4.2)$$

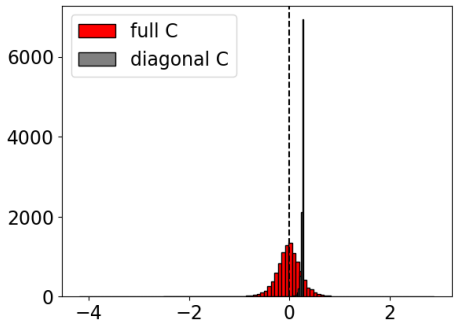
where we have defined the mean prediction for test subject m as $\bar{y}(\mathbf{t}^*_m) = \sum_{b=1}^B y(\mathbf{t}^*_m; D_b)/B$. We can than average the decomposition in (4.2) over all test subjects.

Typically, a very flexible model will have a large variance and a low bias, reflecting an *overfitting* of the training data, while a strongly constrained method will have the opposite behaviour, resulting in *underfitting* of the training data Hart et al. (2000); Bishop and Nasrabadi (2006b). Finding the right balance in the bias-variance trade-off is a key point to achieve good results in a given setting, and there is no method that can be *in absolute* better than others ("no free lunch" theorem) Domingos (2012); Hart et al. (2000).

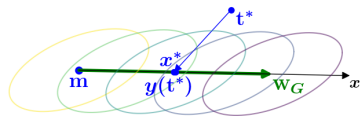
The bias-variance decomposition principle is illustrated in Fig. 4.13, in a 2D toy example from the proposed method. We generate 5 data points for several training sets using a full covariance matrix, and we consider one specific test subject $(\mathbf{t}^*, x^*)$ drawn from the same distribution. We then predict the target variable for the test subject, fitting both a diagonal (*wrong* model) and a full (*correct* model) covariance matrix to the training sets. The histogram shows the distribution of the signed prediction error $y(\mathbf{t}^*; D) - x^*$ for the two models, over 10.000 training runs. The overall MSE is similar in the two cases (MSE= 0.068 vs MSE= 0.072), but the error distribution is very different: the model with diagonal $\mathbf{C}$ yields predictions that are very similar across training runs but systematically wrong (low variance, high bias), while predictions obtained by the more flexible model with full $\mathbf{C}$ vary more across training runs but they are on average correct (high variance, low bias). This is also illustrated by the three shown examples.

In order to gain a similar insight in the real-data experiment, we computed MSE, bias and variance with (4.2), for proposed method, RVoxM and SFCN, using the same training runs described in section 4.1 for age prediction on deformable T1s. We therefore averaged across the training sets that we already have (10 for sizes up to $N = 1000$ and 3 for larger sizes), and used the same test set of 1000 subjects. The computed decomposition is shown in Fig. 4.14. Let us first analyze the decomposition of our method (blue lines). We can see that the bias is reduced as the training set size increases, and the variance slightly decreases. This behaviour is achieved through the method's regularization hyperparameter K: for small training sizes, the hyperparameter constrains the models in order to control the variance, and this results in a larger bias. As the training size increases, the variance is naturally reduced thanks to the larger number of training subjects Hart et al. (2000). This allows the method to be less regularized - as we saw in Table 4.2, the value of K increases as N becomes larger - and thus to achieve a smaller bias.

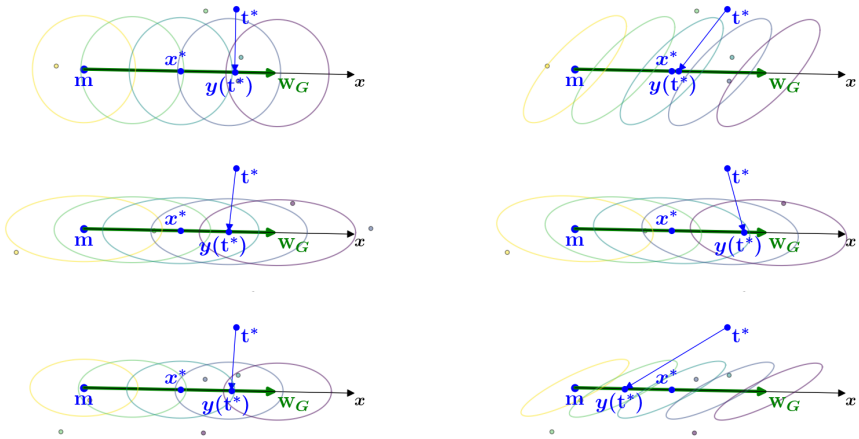
If we now consider the decomposition of RVoxM and SFCN (red and black lines, respectively), we observe a similar behaviour as in our method, with decreasing bias and variance as N increases. If we compare the decomposition of the three methods, we observe that the our method's variance is smaller than the others, except for very large N, where the benchmarks' variances reach (and



histogram of $y(\mathbf{t}^*; D) - x^*$



true model has full $\mathbf{C}$



fit with diagonal $\mathbf{C}$ fit with full $\mathbf{C}$

Fig. 4.13: Visualization of bias-variance principle: fitting a model with diagonal $\mathbf{C}$ (wrong model) yields predictions that are consistent across training runs, but systematically wrong (low variance, high bias), while fitting a model with full $\mathbf{C}$ (correct model) yields predictions that are more variable but on average correct. This principle is illustrated by the histogram of prediction errors, and by the cartoon examples, which display model inversion as in figure 3.2: the test data point $\mathbf{t}^*$ is projected orthogonally onto the direction of $\mathbf{w}_D$ to obtain predictions $y(\mathbf{t}^*)$, while x^* indicates the real target.

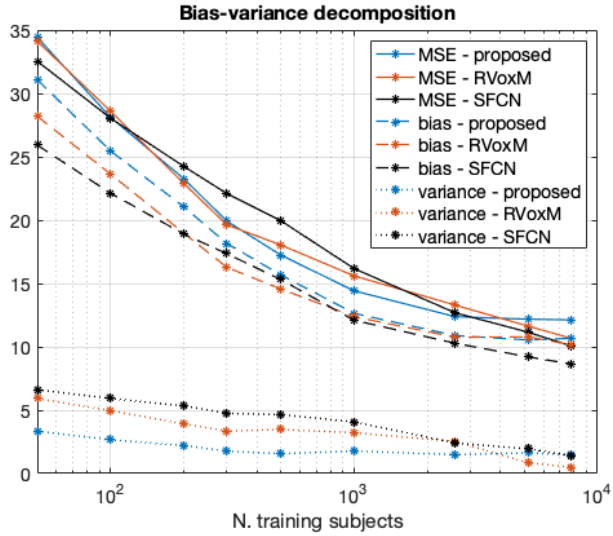


Fig. 4.14: Bias-variance decomposition for the proposed method, RVoxM and SFCN. We used the same training runs as in section 4.1.

become smaller than, in the RVoxM case) the proposed method's one. Instead, the proposed method's bias is larger than the other methods' counterpart, with some training sizes where they are comparable, especially for the RVoxM. This behaviour is in general expected since the proposed method is less flexible than other two, and therefore has a higher bias and a smaller variance, while for larger N, all methods can achieve a small variance thanks to the large number of subjects. The bias-variance trade-off is therefore a tool for interpreting prediction performances: a simpler model like the proposed method is competitive or even outperforms the much more powerful SFCN with training sizes up to a few thousand subjects, because, although its strong assumptions make it on average incorrect (large bias), they also prevent it to overfit (small variance), and this compensates and possibly overcomes the large bias. Conversely, for larger training sizes, there is less risk of overfitting even for a flexible model such as the SFCN, and thus its smaller bias becomes decisive to obtain better prediction errors. These findings are in line with previous studies showing that a more powerful method is not necessarily better than a simpler one, and that, when the training size is limited, models with stronger assumptions - even if incorrect - may yield better performances than more flexible methods, because the latter overfit more Domingos (2012).

We also computed the bias-variance decomposition of age prediction errors for the VAE. We again used the same training runs as in section 4.1, i.e. deformable

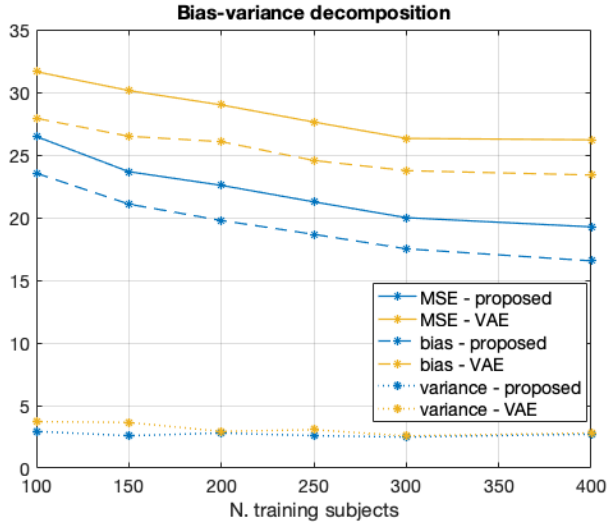


Fig. 4.15: Bias-variance decomposition for the proposed method and VAE. We used the same training runs as in section 4.1.

T1s cropped around the ventricular area as input data, and 10 training sets for each size, in a reduced range (from $N = 100$ to $N = 400$). We also computed the decomposition of our method, trained in the same setting. The computed MSE, bias and variances and displayed in Fig. 4.15. We observe that the VAE has a slightly larger variance and a much larger bias then the proposed method. Therefore the VAE's worse performances than our method's for age prediction reported in section 4.1 are explained mostly by the VAE's higher bias.

Chapter 5

Model extensions and other applications

In this chapter, we present some extensions that can be added to the proposed model, and some possible applications of the method. The chapter is structured as follow:

- We first present model extensions, such as inclusion of known covariates and/or nonlinear effects in the causal model, with relative experimental results.
- We then show results of a possible application, where we try to improve classification performances of the proposed method on a small training set by reusing weights of the model previously trained on a larger cohort.

5.1 Model extensions

In this section we show the effect of extending the proposed model to incorporate known subject-specific variables and nonlinear dependencies on the variable of interest.

Model	CV accuracy	CV AUC	CV sensitivity	CV specificity
Without covariates	0.7023	0.7645	0.6718	0.7328
With extra hyperparameter regulating covariates (age and gender selected in all CV folds)	0.7214	0.7669	0.6794	0.7634

Table 5.1: Performances achieved on the MS vs healthy classification task on 262 subjects from the Munich dataset, with two nested 5-fold CV loops. AUC denotes the area under the ROC curve.

5.1.1 Additional known covariates

In some cases, subject-specific variables are available, such as demographic variables or information about patients, like disease duration, and it can be useful to take them into account. In the proposed model, it is straightforward to do it by simply adding the known variables in the forward model. Note that having a principled way to incorporate known covariates is one of the advantages of the proposed method, as compared to discriminative deep learning models. In fact, for discriminative neural networks, adding known variables is not straightforward and requires to choose in which network layer to infuse them, as in Armanious et al. (2021).

In order to analyze the effect of adding known covariates into the model, we considered the task of classifying multiple sclerosis (MS) patiens vs healthy controls, and added age and gender as covariates. This experiment was conducted using gray matter segmentations from a private dataset owned by Klinikum rechts der Isar (Munich, Germany), from which we selected a cohort of 262 subjects (131 MS patients and 131 age- and sex-matched healthy controls). To assess the contribution of the covariates in an unbiased way, we added an extra binary hyperparameter in the model, regulating the possible inclusion of age and gender, and estimated it together with the number of latent variables with cross-validation (CV). Specifically, we performed two-nested 5-fold CV loops to estimate hyperparameter and assess prediction performances on this small cohort, in an unbiased way: one CV loop is needed to assess prediction performances, since the cohort is too small to have an independent test set, while a second CV loop is used to estimate hyperparmaters, because of the lack of a separate validation set. We then compared the obtained results with the baseline version of the model with no additional covariates (including only the "MS effect").

Table 5.1 displays the obtained results. The model with the additional hyperparameter regulating the use of covariates selected the inclusion of age and gender

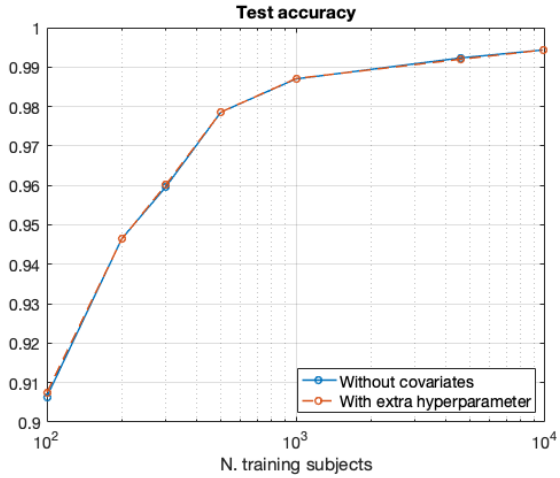


Fig. 5.1: Comparison of test accuracy for gender prediction on UK Biobank, with binary hyperparmater regulating inclusion of age, and without covariates.

in all the CV folds, and obtained slightly better test results then the baseline model.

We also tested the inclusion of known variables on UK Biobank, by incorporating age as covariate in gender classification tasks based on T1 scans. In this experiment, we used the same setting as in section 4.1, and the binary hyperparameter regulating the possible inclusion of age is estimated on the validation set. Fig. 5.1 shows the results, compared with the baseline model without any covariate. We see that, in this experiment, adding age does not improve prediction accuracy. A possible explanation is that, when the covariate is not included in the model, the variability in images due to age is automatically modelled by the method in the noise component.

5.1.2 Nonlinear forward model

The proposed generative method described in Chapter 3 models image intensities as linear functions of the variable of interest. However, in some scenarios it may be useful to incorporate nonlinear dependencies on the target variable. This can be easily done in the proposed method, by adding nonlinear terms in the causal model.

Model	CV MAE	CV RMSE	CV correlation
Linear	4.7335	5.9283	0.9330
With extra hyperparameter for quadratic vs linear (quadratic model selected in all CV folds)	4.3627	5.4322	0.9445

Table 5.2: Performances for age prediction from GM images, on the IXI dataset, with two nested 5-fold CV loops.

We explored the effect of adding nonlinearities, by using a quadratic forward model for age prediction. We selected this test case, since it has been shown that aging has an approximately quadratic effect on several brain structures across the entire lifespan Walhovd et al. (2005) Fjell et al. (2013). For this experiment, we considered a dataset with a larger age span than the UK Biobank, which makes the quadratic effect more visible. In particular, we used the IXI dataset, an open access collection of around 600 T1-weighted MRI scans of healthy subjects, aged 20-86 years. Since gray matter is one of the brain tissues displaying a quadratic behaviour Walhovd et al. (2005) Fjell et al. (2013), we employed gray matter segmentations as input data.

As for adding known covariates, we implemented the possible inclusion of a quadratic term with an extra binary hyperparameter that is estimated with CV, and then compared this model against the linear version. Table 5.2 displays the results. When using the extra hyper-parameter for the choice of linear vs quadratic model, the quadratic version is selected in all the CV folds, and it achieves better test results than the baseline model.

We also tested the quadratic model for age prediction on T1 scans from UK Biobank data. For this experiment, we used the same validation and test set as in section 4.1, and we considered several training sizes. The validation set is used to estimate the binary hyperparameter which selects the linear or quadratic model, together with the number of latent variables. Table 5.3 shows the obtained results, and comparison with the linear model. Here, the quadratic model was selected for all training sizes, while results on the test set present some variability, but are anyway quite similar for the two models. N=7800 is the only case where the quadratic model outperforms the linear version both in terms of MAE and correlation. These results seem to suggest that adding a quadratic term in the model for age prediction on UK Biobank's T1 scans does not impact performances, except for very large training sizes where it yields a certain improvement. This is probably due to the limited age range in the UK Biobank (44-82 years), which makes the quadratic effect much less pronounced. Therefore, adding a nonlinear term may become beneficial only when the training size

Model	Test MAE			Test correlation		
	N=300	N=2600	N=7800	N=300	N=2600	N=7800
Linear	3.5840	2.7410	2.7053	0.8159	0.8892	0.8938
With extra hyperparameter (quadratic model selected for each N)	3.5758	2.7658	2.6226	0.8047	0.8852	0.8972

Table 5.3: Performances for age prediction on T1 scans from the UK Biobank, for different training sizes.

is large enough to actually observe the quadratic trend in the data.

As a side note, if we compare MAEs obtained on IXI data and UK Biobank for similar training sizes, we observe that errors obtained on the IXI dataset are larger. Besides the different input features (GM vs T1), this is caused by the larger age span in the IXI data, which automatically results in bigger prediction errors Cole et al. (2019).

5.2 Reusing part of the model

Let us consider a task that is characterized by a small cohort, such as classification of a certain disease vs healthy controls, where small sample sizes are the typical scenario. A possible advantage of the proposed method in this scenario is the possibility to train the method on a huge cohort of healthy subjects (that may be available), and then re-use the estimated noise model in the small cohort task. We wanted to explore if this technique can yield more accurate results than performing the classification task using only the small cohort, since the noise model would be estimated on a much larger sample.

In order to do this, we used T1 scans from the UK Biobank, and we considered the classification task of MS patients vs healthy controls. The UK Biobank contains 87 scans of MS patients, and we selected the same amount of age- and sex-matched healthy subjects, yielding a cohort of 174 data. We then compared results of our method trained on this cohort (with two-nested 5-fold cross validation loops: one for assessing prediction performances and the other for hyperparameter selection), with the following version: the model is first trained on $N = 9800$ healthy subjects from the UK Biobank, with age and gender as variables in the causal model. Subsequently, the estimated noise model, together with age and gender effects, are re-used in the MS vs healthy classification task, with only the "disease effect" being estimated on the small cohort, and prediction performances assessed with a single 5-fold CV loop.

	CV accuracy	CV AUC	CV sensitivity	CV specificity
Model fit on small cohort	0.7356	0.7685	0.6207	0.8506
With pre-trained weights	0.5632	0.5920	0.1264	1.0000

Table 5.4: Results of MS vs healthy classification on 174 data from UK Biobank. The model trained entirely on the small cohort is compared to the model with weights pre-trained on 9800 healthy subjects. AUC denotes the area under the ROC curve.

	CV accuracy	CV AUC	CV sensitivity	CV specificity
Model fit on small cohort	0.7356	0.7685	0.6207	0.8506
With pre-trained weights and scaling	0.7874	0.8575	0.7471	0.8276

Table 5.5: Same comparison as in Table 5.4, but the model with pre-trained weights now includes a scaling factor in the disease effect. The selected (cross-validated) scaling factor is 0.1.

Results are shown in Table 5.4. We see that, conversely to what we expected, the model with pre-trained weights completely overfit to the training data. However, we noticed that if we consider the weights of the disease effect (the only weights that are actually estimated on the small cohort) and scale them by a certain factor, the classifier starts to work well. We therefore cross-validated the estimate of this factor, and results are shown in Table 5.5. We can see that the classifier with pre-trained weights now outperforms the model that is trained entirely on the small cohort. At the time of writing, we were not able to fully understand why the model with pre-trained weights does not work at all without the scaling factor, which can be seen as a form of regularization (with a prior over the weights that encourages smaller estimates).

Chapter 6

Conclusions and future work

As discussed in chapter 1, in this thesis we wanted to explore if a linear or shallow nonlinear generative model, which provides interpretable predictions, could also achieve good prediction performances in neuroimaging tasks. For this reason, we developed a generative method for image-based predictions, with a forward model composed by two parts: On one hand, the causal part expresses the effect of the target variable on brain anatomy, yielding interpretable maps. On the other hand a linear noise model captures the dominant correlations between voxels, allowing us, when the model is inverted, to obtain accurate predictions. The proposed model can also be extended to include shallow nonlinearities in the variable of interest and/or known subject-specific covariates.

We demonstrated that the proposed method achieves good performances in prediction tasks, with no trade-off between accuracy and interpretability. In fact, in the experiments performed for age prediction based on brain MRI scans, we showed that it was competitive with discriminative state-of-the-art methods, especially for moderate sample sizes (up to 2600 training subjects), which is the typical setting in many neuroimaging applications. In gender classification experiments, it was even competitive for every considered sample size (up to 9800 training subjects). We also gave insight into the performances of our method and benchmarks in terms of bias-variance trade-off: we found that the good performances of our method derive from its low variance, which offsets the large bias that characterizes a simple method.

The proposed method achieved good performances when compared to nonlinear generative benchmarks as well. We demonstrated this by comparing our method with the VAE proposed in Zhao et al. (2019) for age prediction, which can be regarded as a deep nonlinear version of our method. We found that our method achieved better results than the VAE in the age prediction task for every tested training size, suggesting that adding deep nonlinearities in the model in not beneficial in this scenario. Another nonlinear generative model proposed in the literature for age prediction is Wilms et al. (2020), which uses normalizìng flows. We did not explicitly compare this method with our model, but its reported MAE as percentage of age range (6.3% with N=4281) is similar to ours (7.4% with N=2600 and 7.3% with N=5200).

In addition to competitive performances, we showed that the proposed method has the advantage of being interpretable, producing maps that show target-related morphological changes on a population-level. Conversely, discriminative methods have been proven to be problematic to interpret, both in the linear and nonlinear case. For deep nonlinear generative methods, as discussed in section 4.2, it is still possible to produce global interpretable maps, but it is much more difficult than for our method, for which they are readily available. Furthermore, we also illustrated that the proposed method has also the advantage of being simpler to use than discriminative and generative benchmarks, and it requires less time and resources for training it.

A downside of our method is that, while it fits very well settings with up to a few thousands of training data, training the model on larger samples becomes moderately slow (on a CPU), and the achieved performances are not competitive in regression tasks.

As future work, the proposed method can be extended to operate in a longitudinal scenario, where more than one image per subject is acquired. Being generative, our method can easily adapt to the longitudinal setting, which usually presents variability in number of scans per subject and in time intervals between scans. Conversely, discriminative methods have to impute missing data or discard observations in order to tackle this kind of data, and are therefore less suited for this scenario. The longitudinal version of the proposed method can be obtained using mixed-effect models, which take into account the temporal correlation between different images of the same subject.

Finally, in the experiments performed in this thesis, the number of latent variables in the model was tuned using grid search. A possible extension is to directly estimate this hyperparameter on the training set, with variational methods Bishop and Nasrabadi (2006a), which gives the advantage of avoiding repeated model training.

Chapter 7

Paper A

Accurate and Explainable Image-based Prediction Using a Lightweight Generative Model

Chiara Mauri[1⋆], Stefano Cerri[2], Oula Puonti[3], Mark Mühlau[4], and Koen Van Leemput[1,2]

[1] Department of Health Technology, Technical University of Denmark, Denmark
[2] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA
[3] Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Denmark
[4] Department of Neurology and TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Germany

Abstract. Recent years have seen a growing interest in methods for predicting a variable of interest, such as a subject's age, from individual brain scans. Although the field has focused strongly on nonlinear discriminative methods using deep learning, here we explore whether linear generative techniques can be used as practical alternatives that are easier to tune, train and interpret. The models we propose consist of (1) a causal forward model expressing the effect of variables of interest on brain morphology, and (2) a latent variable noise model, based on factor analysis, that is quick to learn and invert. In experiments estimating individuals' age and gender from the UK Biobank dataset, we demonstrate competitive prediction performance even when the number of training subjects is in the thousands – the typical scenario in many potential applications. The method is easy to use as it has only a single hyperparameter, and directly estimates interpretable spatial maps of the underlying structural changes that are driving the predictions.

1 Introduction

Image-based prediction methods aim to estimate a variable of interest, such as a subject's diagnosis or prognosis, directly from a medical scan. Predicting a subject's age based on a brain scan – the so called brain age – in particular has seen significant interest in the last decade [12], with the gap between *brain* age and *chronological* age being suggested as a potential biomarker of healthy aging and/or neurological disease [12, 25].

Methods with state-of-the-art prediction performance are currently based on *discriminative learning*, in which a variable of interest x is directly predicted from an input image $\boldsymbol{t}$. Although there are ongoing controversies in the literature regarding whether nonlinear or linear discriminative methods predict better [23,

⋆ Corresponding author. Email address: cmau@dtu.dk

32, 28], recent years have seen a strong focus on nonlinear variants based on deep learning (DL), with impressive performances especially when the training size is very large [28]. Nevertheless, these powerful methods come with a number of potential limitations:

The available training size is often limited: While methods for predicting age and gender can be trained on thousands of subjects using large imaging studies [4, 21, 24, 14, 16], in many potential applications the size of the training set is much more modest. In a recent survey on single-subject prediction of brain disorders in neuroimaging, the mean and median samples size was only 186 and 88 subjects, respectively [5]. Even in such ambitious imaging projects as the UK Biobank [4], the number of subjects with diseases such as multiple sclerosis is only projected to be in the hundreds in the coming years.

Discriminative methods are hard to interpret: As opposed to *generative* methods that explicitly model the effect a variable of interest x has on a subject's image $\boldsymbol{t}$, correctly interpreting the internal workings of discriminative methods is known to be difficult [22, 6, 20, 3]. Whereas the spatial weight maps of linear discriminative methods, or more generally the saliency maps of nonlinear ones [35, 8, 17, 34, 38, 37, 33, 36], are useful for highlighting which image areas are being used in the prediction process [20, 29], they do not explain *why* specific voxels are given specific attention: Amplifying the signal of interest, or suppressing noninteresting noise characteristics in the data [22].

DL can be more difficult to use: Compared to less expressive techniques, DL methods are often harder to use, as they can be time consuming to train, and have many more "knobs" that can be turned to obtain good results (e.g., the choice of architecture, data augmentation, optimizer, training loss, etc. [28]).

In this paper, we propose a lightweight generative model that aims to be easier to use and more straightforward to interpret, without sacrificing prediction performance in typical sample size settings. Like in the mass-univariate techniques that have traditionally been used in human brain mapping [7, 13, 11, 18], the method has a causal forward model that encodes how variables of interest affect brain shape, and is therefore intuitive to interpret. Unlike such techniques, however, the method also includes a linear-Gaussian latent variable noise model that captures the dominant correlations between voxels. As we will show, this allows us to efficiently "invert" the model to obtain accurate predictions of variables of interest, yielding an effective linear prediction method without externally enforced interpretability constraints [9, 39].

The method we propose can be viewed as an extension of prior work demonstrating that naive Bayesian classifiers can empirically outperform more powerful methods when the training size is limited, even though the latter have asymptotically better performance [15, 27]. Here we show that these findings translate to prediction tasks in neuroimaging when the strong conditional independence assumption of such "naive" methods is relaxed. Using experiments on age and gender prediction in the UK Biobank imaging dataset, we demonstrate

empirically that, even when the number of training subjects is the thousands, our lightweight linear generative method yields prediction performance that is competitive with state-of-the-art nonlinear discriminative [28], linear discriminative [31], and nonlinear generative [40] methods.

2 Method

Let $\boldsymbol{t}$ denote a vectorized version of a subject's image, and $\boldsymbol{\phi} = (x, \boldsymbol{\phi}_{\setminus x}^T)^T$ a vector of variables specific to that subject, consisting of a variable of interest x (such as their age or gender), along with any other known[5] subject-specific covariates $\boldsymbol{\phi}_{\setminus x}$. A simple generative model is then of the form

$$\boldsymbol{t} = \boldsymbol{W}\boldsymbol{\phi} + \boldsymbol{\eta}, \tag{1}$$

where $\boldsymbol{\eta}$ is a random noise vector, assumed to be Gaussian distributed with zero mean and covariance $\boldsymbol{C}$, and $\boldsymbol{W} = (\boldsymbol{w}_x \ \boldsymbol{W}_{\setminus x})$ is a matrix with spatial weight maps stacked in its columns. The first column, $\boldsymbol{w}_x$, expresses how strongly the variable of interest x is expressed in the voxels of $\boldsymbol{t}$; we will refer to it as the *generative* weight map. Taking everything together, the image $\boldsymbol{t}$ is effectively modeled as Gaussian distributed:

$$p(\boldsymbol{t}|\boldsymbol{\phi}, \boldsymbol{W}, \boldsymbol{C}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{W}\boldsymbol{\phi}, \boldsymbol{C}).$$

Making predictions

When the parameters of the model are known, the unknown target variable x^* of a subject with image $\boldsymbol{t}^*$ and covariates $\boldsymbol{\phi}_{\setminus x}^*$ can be inferred by inverting the model using Bayes' rule. For a binary target variable $x^* \in \{0, 1\}$ where the two outcomes have equal prior probability, the target posterior distribution takes the form of a logistic regression classifier:

$$p(x^* = 1|\boldsymbol{t}^*, \boldsymbol{\phi}_{\setminus x}^*, \boldsymbol{W}, \boldsymbol{C}) = \sigma\big(\boldsymbol{w}_D^T\boldsymbol{t}^* + w_o\big),$$

where

$$\boldsymbol{w}_D = \boldsymbol{C}^{-1}\boldsymbol{w}_x$$

are a set *discriminative* spatial weights, $\sigma(\cdot)$ denotes the logistic function, and $w_o = -\boldsymbol{w}_D^T(\boldsymbol{W}_{\setminus x}\boldsymbol{\phi}_{\setminus x}^* + \boldsymbol{w}_x/2)$. The prediction of x^* is therefore 1 if $\boldsymbol{w}_D^T\boldsymbol{t}^* + w_o > 0$, and 0 otherwise.

For a continuous target variable with Gaussian prior distribution $p(x^*) = \mathcal{N}(x^*|0, \sigma^2)$, the posterior distribution is also Gaussian with mean

$$\sigma_x^2(\boldsymbol{w}_D^T\boldsymbol{t}^* + b_0), \tag{2}$$

where $b_0 = -\boldsymbol{w}_D^T\boldsymbol{W}_{\setminus x}\boldsymbol{\phi}_{\setminus x}^*$ and $\sigma_x^2 = \big(\sigma^{-2} + \boldsymbol{w}_x^T\boldsymbol{C}^{-1}\boldsymbol{w}_x\big)^{-1}$. The predicted value of x^* is therefore given by (2), which again involves taking the inner product of the discriminative weights $\boldsymbol{w}_D$ with $\boldsymbol{t}^*$.

[5] For notational convenience, we include 1 as a dummy "covariate".

Model training

In practice the model parameters $\boldsymbol{W}$ and $\boldsymbol{C}$ need to be estimated from training data. Given N training pairs $\{\boldsymbol{t}_n, \boldsymbol{\phi}_n\}_{n=1}^N$, their maximum likelihood (ML) estimate is obtained by maximizing the marginal likelihood

$$p\left(\{\boldsymbol{t}_n\}_{n=1}^N | \{\boldsymbol{\phi}_n\}_{n=1}^N, \boldsymbol{W}, \boldsymbol{C}\right) = \prod_{n=1}^N \mathcal{N}\left(\boldsymbol{t}_n |\ \boldsymbol{W}\boldsymbol{\phi}_n, \boldsymbol{C}\right) \tag{3}$$

with respect to these parameters. For the spatial maps $\boldsymbol{W}$, the solution is given in closed form:

$$\boldsymbol{W} = \left(\sum_{n=1}^N \boldsymbol{t}_n \boldsymbol{\phi}_n^T\right)\left(\sum_{n=1}^N \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T\right)^{-1}. \tag{4}$$

Obtaining the noise covariance matrix $\boldsymbol{C}$ directly by ML estimation is problematic, however: For images with J voxels, $\boldsymbol{C}$ has $J(J+1)/2$ free parameters – orders of magnitude more than there are training samples. To circumvent this problem, we impose a specific structure on $\boldsymbol{C}$ by using a latent variable model known as factor analysis [10]. In particular, we model the noise as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{z} + \boldsymbol{\epsilon},$$

where $\boldsymbol{z}$ is a small set of K unknown latent variables distributed as $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \mathbb{I}_K)$, $\boldsymbol{V}$ contains K corresponding, unknown spatial weight maps, and $\boldsymbol{\epsilon}$ is a zero-mean Gaussian distributed error with unknown diagonal covariance $\boldsymbol{\Delta}$. Marginalizing over $\boldsymbol{z}$ yields a zero-mean Gaussian noise model with covariance matrix

$$\boldsymbol{C} = \boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{\Delta},$$

which is now controlled by a reduced set of parameters $\boldsymbol{V}$ and $\boldsymbol{\Delta}$. The number of columns in $\boldsymbol{V}$ (i.e., the number of latent variables K) is a hyperparameter in the model that needs to be tuned experimentally.

Plugging in the ML estimate of $\boldsymbol{W}$ given by (4), the parameters $\boldsymbol{V}$ and $\boldsymbol{\Delta}$ maximizing the marginal likelihood (3) can be estimated using an Expectation-Maximization (EM) algorithm [30]. Applied to our setting, this yields an iterative algorithm that repeatedly evaluates the posterior distribution over the latent variables:

$$p(\boldsymbol{z}_n|\boldsymbol{t}_n, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{\Delta}) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}_n = \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)$ and $\boldsymbol{\Sigma} = (\mathbb{I}_K + \boldsymbol{V}^T\boldsymbol{\Delta}^{-1}\boldsymbol{V})^{-1}$, and subsequently updates the parameters:

$$\boldsymbol{V} \leftarrow \left(\sum_{n=1}^N (\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)\boldsymbol{\mu}_n^T\right)\left(\sum_{n=1}^N \left(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + \boldsymbol{\Sigma}\right)\right)^{-1}$$

$$\boldsymbol{\Delta} \leftarrow \operatorname{diag}\left(\frac{1}{N}\sum_{n=1}^N (\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)(\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)^T - \boldsymbol{V}\frac{1}{N}\sum_{n=1}^N \boldsymbol{\mu}_n(\boldsymbol{t}_n - \boldsymbol{W}\boldsymbol{\phi}_n)^T\right).$$

3 Experiments

In our implementation, we initialize the EM algorithm by using a matrix with standard Gaussian random entries for $\boldsymbol{V}$, and a diagonal matrix with the sample variance in each voxel across the training set for $\boldsymbol{\Delta}$. For continuous target variables, we de-mean the target and use the sample variance as the prior variance σ^2. Convergence is detected when the relative change in the log marginal likelihood is smaller than 10^{-5}.

The method has a single hyperparameter, the number of latent variables K, that we set empirically using cross-validation on a validation set, by optimizing the mean absolute error (MAE) for regression and the accuracy for classification. Running times vary with the size of the training set N, which also influences the selected value of K – in our implementation, typical training runs in the full-brain experiments described below took between 2.8 and 16.3 minutes for $N = 200$ and $N = 1000$, respectively (CPU time for a single selected value of K; Matlab on a state-of-the-art desktop). Once the model is trained, testing is fast: typically 0.01 seconds per subject when trained on $N = 1000$.

Comparing performance of an image-based prediction method with state-of-the-art benchmark methods is hampered by the dearth of publicly available software implementations, and the strong dependency of attainable performance on the datasets that are used [12]. Within these constraints, we conducted the following comparisons of the proposed linear generative method:

Nonlinear discriminative benchmark: As the main benchmark method, we selected the convolutional neural network SFCN proposed in [28], which is, to the best of our knowledge, currently the best performing image-based prediction method. The paper reports performance for age and gender prediction over a wide range of training sizes in preprocessed UK Biobank data (14,503 healthy subjects, aged 44-80 years), using a validation set of 518 subjects and a test set of 1036 subjects. For a training size of 12,949 subjects, the authors report a training time of 65 hours on two NVIDIA P100 GPUs [28]. Although the method uses affinely registered T1-weighted scans as input ("affine T1s"), these are in fact skull-stripped and subsequently bias-field-corrected based on deformable registrations that are also available [4]. Because of this reason, and because the authors report only very minor improvements of their method when deformable T1s are used instead ($\sim 2.5\%$ decrease in MAE for age prediction on 2590 training subjects), we compared our method using both affine and deformable T1s, based on a set-up that closely resembles theirs (validation set of 500 subjects, test set of 1000 subjects).

Linear discriminative benchmark: In order to compare against a state-of-the-art *linear* discriminative method, we selected the RVoxM method [31] because its training code is readily available [1] and its performance is comparable to the best linear discriminative method tested in [28]. RVoxM regularizes its linear discriminant surface by encouraging spatial smoothness and sparsity of its weight maps, using a regularization strength that is the one

hyperparameter of the method. In our experiments, we selected the optimal value of this hyperparameter in the same way as we do it for the proposed method, i.e., by cross-validation on our 500-subject validation set. Typical training times were between 66 and 122 minutes for $N = 200$ and $N = 1000$, respectively (CPU time for a single selected value of the model's hyperparameter; Matlab on a state-of-the-art desktop).

Nonlinear generative benchmark: As a final benchmark, we compared against a variational auto-encoder (VAE) [40] that was recently proposed for age prediction, and that has training code publicly available [2]. It is based on a generative model that is similar to ours, except that its latent variables are expanded ("decoded") *nonlinearly* using a deep neural network, which makes the EM training algorithm more involved compared to our closed-form expressions [26]. In [40], the authors use T1 volumes that are cropped around the ventricular area (cf. Fig. 1 right), and they train their method on ∼200 subjects. We closely follow their example and train both the VAE and the proposed method on similarly sized training sets of warped T1 scans from the UK Biobank, cropped in the same way. There are two hyperparameters in the VAE model (dropout factor and L2 regularization), which we optimized on our validation set of 500 subjects using grid search. The training time for this method was on average 9.40 minutes for N=200 with the optimal set of hyperparameters, using a NVIDIA GeForce RTX 2080 Ti GPU.

For each training size tested, we trained each method three times, using randomly sampled training sets, and report the average test MAE and accuracy results. For gender classification, we used age as a known covariate in $\boldsymbol{\phi}_{\setminus x}$, while for age prediction no other variables were employed. All our experiments were performed on downsampled (to 2mm isotropic) data, with the exception of RVoxM where 3mm was used due to time constraints – we verified experimentally that results for RVoxM nor the proposed method would have changed significantly had the downsampling factor been changed (max difference of 0.32% in MAE between 2mm and 3mm across multiple training sizes between 100 and 1000). Since training code for SFCN is not publicly available, we report the results as they appear in [28], noting that the method was tuned on a 518-subject validation set as described in the paper.

4 Results

Fig. 1 shows examples of the *generative* spatial map $\boldsymbol{w}_x$ estimated by the proposed method, along with the the corresponding *discriminative* map $\boldsymbol{w}_D$. The generative map shows the direct effect age has on image intensities, and reflects the typical age-related gray matter atrophy patterns reported in previous studies [19]. The discriminative map, which highlights voxels that are employed for prediction, is notably different from the generative map and heavily engages white matter areas instead. This illustrates the interpretation problem in discriminative models: the discriminative weight map does not directly relate to changes in neuroanatomy, but rather summarizes the net effect of decomposing

the signal as a sum of age-related changes and a typical noise pattern seen in the training data (1), resulting in a non-intuitive spatial pattern [22].

Fig. 2 shows the performances obtained by the proposed method, compared to the discriminative benchmarks RVoxM and SFCN, for age and gender prediction. Both our method and RVoxM achieve clearly worse results when they are applied to affine T1s compared to deformable T1s, whereas SFCN's performance is virtually unaffected by the type of input data (at least for age prediction with 2590 training subjects – the only available data point for SFCN with deformable T1s [28]). These results are perhaps not surprising, since both our method and RVoxM are *linear* predictors that do not have the same capacity as neural networks to "model away" nonlinear deformations that have not been removed from the input images (even though these are actually known and were used for generating the affine T1s).

Comparing the performances of the different methods, our generative model generally outperforms the linear discriminative RVoxM for both age and gender prediction, except when using very large training sets of affine T1s. For *nonlinear* discriminative SFCN, the situation is more nuanced: For age prediction, SFCN starts outperforming our method for training sets larger than 2600 subjects, while for more moderate training sizes our method achieves better performances when deformable T1s are used. For gender prediction, our method based on deformable T1s is competitive with SFCN even on the biggest training set sizes, although it should be noted that SFCN's results are based on affine T1s as its performance on deformable T1s for gender prediction was not tested[6] in [28].

Finally, Fig. 3 compares the age prediction results of our linear generative model with the nonlinear generative VAE, both trained on cropped deformable T1s. Our method clearly outperforms the VAE for all the considered training sizes, suggesting that, at least when only a few hundred training subjects are available, adding nonlinearities in the model is not beneficial.

5 Discussion

In this paper, we have introduced a lightweight method for image-based prediction that is based on a linear generative model. The method aims to be easier to use, faster to train and less opaque than state-of-the-art nonlinear and/or discriminative methods. Based on our experiments in predicting age and gender from brain MRI scans, the method seems to attain these goals without sacrificing prediction accuracy, especially in the limited training size scenarios that are characteristic of neuroimaging applications.

Although the method presented here is linear in both its causal forward model and in its noise model, it would be straightforward to introduce nonlinearities in the forward model while still maintaining numerical invertibility. This may be beneficial in e.g., age prediction in datasets with a much wider age range than the UK Biobank data used here. The method can also be generalized to longitudinal

[6] Nevertheless, SFCN's gender prediction, based on affine T1s, is reported by its authors to be the best in the literature.

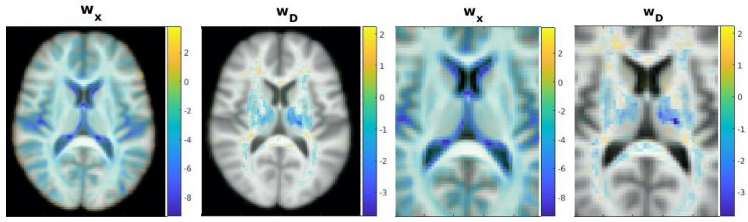


Fig. 1: Examples of *generative* maps $\boldsymbol{w}_x$ encoding age effects vs. the corresponding *discriminative* maps $\boldsymbol{w}_D$ predicting age, obtained on deformable T1s from 300 subjects and overlaid on the average T1 volume. Voxels with zero weight are transparent. Left: results on whole T1 images (used for comparing the proposed method with SFCN and RVoxM). Right: results on cropped T1s (used for comparing with VAE).

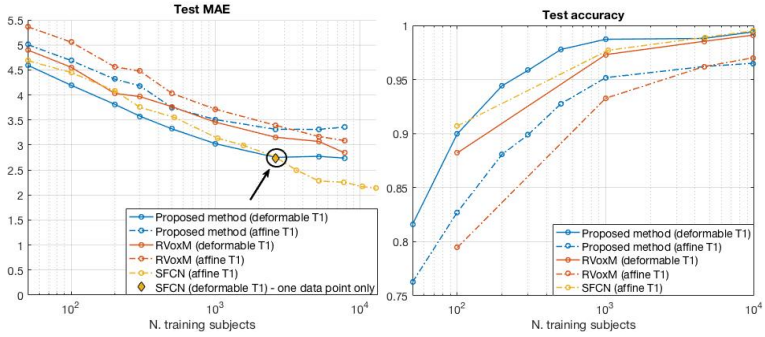


Fig. 2: Comparison of the proposed method, RVoxM and SFCN on an age prediction task (left) and on a gender classification task (right). For each method, results are shown for both affine and deformable T1 input data – except for SFCN for which the result for deformable T1s is only known for age prediction, in a single point (indicated by an arrow at 2590 subjects).

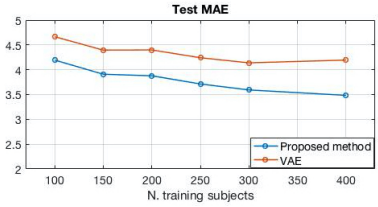


Fig. 3: Test MAE for age prediction obtained by the proposed method and VAE on cropped, deformable T1s.

data, where addressing the intersubject variability in both the timing and the number of follow-up scans is well suited for generative models such as the one proposed here.

Acknowledgments This research has been conducted using the UK Biobank Resource under Application Number 65657. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreements No. 765148 and No. 731827, as well as from the National Institutes Of Health under project numbers R01NS112161 and 1RF1MH117428.

References

1. https://sabuncu.engineering.cornell.edu/software-projects/relevance-voxel-machine-rvoxm-code-release/
2. https://github.com/QingyuZhao/VAE-for-Regression
3. Adebayo, J., et al.: Sanity checks for saliency maps. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
4. Alfaro-Almagro, F., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage **166**, 400–424 (2018)
5. Arbabshirani, M.R., et al.: Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage **145**, 137–165 (2017)
6. Arun, N., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6), e200267 (2021)
7. Ashburner, J., et al.: Voxel-based morphometry–the methods. Neuroimage **11**(6), 805–821 (2000)
8. Baehrens, D., et al.: How to explain individual classification decisions. The Journal of Machine Learning Research **11**, 1803–1831 (2010)
9. Batmanghelich, N.K., et al.: Generative-discriminative basis learning for medical imaging. IEEE transactions on medical imaging **31**(1), 51–69 (2011)
10. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4, chap. 12. Springer (2006)
11. Chung, M., et al.: A unified statistical approach to deformation-based morphometry. NeuroImage **14**(3), 595–606 (2001)
12. Cole, J.H., et al.: Quantification of the biological age of the brain using neuroimaging. In: Biomarkers of human aging, pp. 293–328. Springer (2019)
13. Davatzikos, C., et al.: Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage **14**(6), 1361–1369 (2001)
14. Di Martino, A., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19**(6), 659–667 (2014)
15. Domingos, P., et al.: On the optimality of the simple bayesian classifier under zero-one loss. Machine learning **29**(2), 103–130 (1997)
16. Ellis, K.A., et al.: The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
17. Erhan, D., et al.: Visualizing higher-layer features of a deep network. University of Montreal **1341**(3), 1 (2009)
18. Fischl, B., et al.: Measuring the thickness of the human cerebral cortex from magnetic resonance images. PNAS **97**(20), 11050 (2000)
19. Fjell, A.M., et al.: High Consistency of Regional Cortical Thinning in Aging across Multiple Samples. Cerebral Cortex **19**(9), 2001–2012 (2009). https://doi.org/10.1093/cercor/bhn232
20. Ghassemi, M., et al.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (2021)
21. Glasser, M.F., et al.: The human connectome project's neuroimaging approach. Nature neuroscience **19**(9), 1175–1187 (2016)

22. Haufe, S., et al.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage **87**, 96–110 (2014)
23. He, T., et al.: Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage **206**, 116276 (2020)
24. Jack Jr, C.R., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
25. Kaufmann, T., et al.: Common brain disorders are associated with heritable patterns of apparent aging of the brain. Nature neuroscience **22**(10), 1617–1623 (2019)
26. Kingma, D.P., et al.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
27. Ng, A.Y., et al.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: Advances in neural information processing systems. pp. 841–848 (2002)
28. Peng, H., et al.: Accurate brain age prediction with lightweight deep neural networks. Medical image analysis **68**, 101871 (2021)
29. Ras, G., et al.: Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research **73**, 329–397 (2022)
30. Rubin, D.B., et al.: Em algorithms for ml factor analysis. Psychometrika **47**(1), 69–76 (1982)
31. Sabuncu, M.R., et al.: The Relevance Voxel Machine (RVoxM): A Self-Tuning Bayesian Model for Informative Image-based Prediction. IEEE transactions on medical imaging **31**(12), 2290–2306 (2012)
32. Schulz, M.A., et al.: Deep learning for brains?: Different linear and nonlinear scaling in uk biobank brain images vs. machine-learning datasets. BioRxiv p. 757054 (2019)
33. Selvaraju, R.R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
34. Shrikumar, A., et al.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
35. Simonyan, K., et al.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations (2014)
36. Smilkov, D., et al.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
37. Springenberg, J.T., et al.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
38. Sundararajan, M., et al.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
39. Varol, E., et al.: Generative discriminative models for multivariate inference and statistical mapping in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 540–548. Springer (2018)
40. Zhao, Q., et al.: Variational autoencoder for regression: Application to brain aging analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 823–831. Springer (2019)

Chapter 8

Paper B

An Accurate and Interpretable Generative Model for Image-based Prediction

Chiara Mauri, Stefano Cerri, Oula Puonti, Mark Mühlau, Koen Van Leemput

***Abstract*— Recent years have seen a significant development of computational methods for predicting a variable of interest, such as a subject's diagnosis or prognosis, based on brain Magnetic Resonance Imaging (MRI) scans. While the field has mainly focused on deep discriminative learning techniques, here we propose an alternative approach for image-based prediction based on a lightweight generative method, which yields accurate and interpretable predictions, and which is also simple and fast to use, with only one hyperparameter to tune. The proposed method consists of (1) a causal forward model expressing the direct effect of the variable of interest on brain anatomy, and (2) a linear latent variable noise model, based on factor analysis, which captures dominant correlations in the data and allows to obtain accurate predictions once the model is inverted. In experiments estimating individuals' age and gender from the UK Biobank dataset, we demonstrate competitive prediction performance as compared to state-of-the-art benchmarks, even when the number of training subjects is in the thousands, which is the typical scenario in many potential applications. Using the task of age prediction, we also demonstrate that the proposed method is interpretable, providing spatial maps that display known age-related effects on brain morphology. We finally investigate possible model extensions and applications, where the proposed method is easily extended to incorporate known covariates and/or nonlinearities in the target variable.**

***Index Terms*—**

I. Introduction

Image-based prediction methods aim to estimate a variable of interest directly from a medical scan - either a continuous variable such as a subject's disease score (*regression* methods), or a categorical variable such as a patient's diagnosis or prognosis (*classification* methods). The ability to make reliable image-based predictions at individual level is of clinical interest; For example, methods for automatic prediction of a subject's diagnosis can leverage subtle anatomical changes detected by MRI scans and diagnose disorders earlier than clinical assessment, with consequent better clinical outcomes. Image-based diagnosis is also particularly useful for diseases with no standard clinical tests, such as schizophrenia[1], and it can provide new understanding of disorders and their underlying mechanisms. Another relevant task is to predict individual disease progression, for example by identifying patients at higher risk of future disability accrual, allowing better counseling and more personalized treatments. This is particularly useful for diseases, such as multiple sclerosis, where several possible treatments are available[2], and it is hard to foresee in the initial stages the efficacy of different treatments on a specific patient and the disease time course and outcome. Image-based prediction of individual prognosis has also the potential of giving insight into subtle morphological and temporal dynamics underlying disease progression.

A specific application that has seen a significant development in the last decade is prediction of a subject's age based on the brain scan - the so called brain age [1]. In particular, the last three years have seen a further increase of brain age prediction studies, encouraged by the growing availability of large datasets, containing thousands or even tens of thousand subjects. Besides its utility for developing and testing image-based prediction methods, brain age prediction has shown clinical interest, with the gap between *brain* age and *chronological* age being suggested as a potential biomarker of healthy aging and/or neurological disease [1], [2].

Image-based prediction methods with state-of-the-art prediction performance are currently based on *discriminative learning*, in which a variable of interest x is directly predicted from an input image $\mathbf{t}$. Although there are ongoing controversies in the literature regarding whether nonlinear or linear discriminative methods predict better [3]–[5], recent years have seen a strong focus on nonlinear variants based on deep learning (DL), with impressive performances especially when the training size is very large [5]. Nevertheless, these powerful methods come with a number of potential limitations:

- **The available training size is often limited**:
  While methods for predicting age and gender can be trained on many thousands of healthy subjects using large imaging studies [6]–[9], in many potential applications the size of the training set is much more modest. For

Chiara Mauri is with the Department of Health Technology, Technical University of Denmark, Denmark (e-mail:cmau@dtu.dk).

Stefano Cerri is with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA.

Oula Puonti is with the Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Denmark.

Mark Mühlau is with the Department of Neurology and TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Germany.

Koen Van Leemput is with the Department of Health Technology, Technical University of Denmark, Denmark and with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA.

[1] https://www.nhs.uk/mental-health/conditions/schizophrenia/diagnosis/

[2] http://nationalMSsociety.org/DMT

instance, in a recent survey of over 200 papers on single-subject prediction of brain disorders in neuroimaging, the mean and median samples size was only 186 and 88 subjects, respectively [10], as shown in Fig. 1. Even in such ambitious imaging projects as the UK Biobank [6], [11], which aims at scanning 100.000 participants, the number of subjects with fairly common diseases is quite modest. In fact, in 2022 the UK Biobank should contain images of 900 subjects with stroke, 900 with Alzheimer's Disease, and 600 with Parkinson's Disease, given these diseases' prevalence in the population [12] and the fact that the scanning process is still half way, with 50.000 subjects scanned[3]. In 2027, these numbers are expected to rise to 4.000 subjects with stroke, 6.000 with Alzheimer's Disease, and 2.800 with Parkinson's Disease [6]. These estimates show that, even in the world's largest imaging study, the amount of subjects with fairly common diseases is quite moderate, and it will not be massive even in the coming years.

Similar considerations can be applied to other prospective cohort imaging studies, such as the German National Cohort, which aims at scanning 30.000 subjects for investigating several major chronic diseases [7], the Rhineland Study, which plans to scan 30.000 participants to study neurodegenerative and neuropsychiatric diseases [8], and the Maastricht Study composed by 10.000 subjects, which is however artificially enriched with type 2 diabetes participants, to increase efficiency in the study of this disease [9]. Other imaging datasets collected for studying specific diseases and/or healthy aging have sizes at most of the order of 1000-2000 subjects, such as ADNI [13], ABIDE [14], AIBL [15], CoRR [16], HCP [17], PING [18], PNC [19], SHIP [20].

Therefore, given the amount of available imaging data in many practical, e.g. diseases-related, applications, there is a continued need for methods that can learn efficiently from fairly small sample sizes.

- **Discriminative methods are hard to interpret:**
  Generative methods explicitly model the effect a variable of interest x has on a subject's image $\boldsymbol{t}$, directly expressing how it affects brain anatomy [21]. Conversely, for discriminative methods, gaining insight into the ĺink between the predicted value and the input image has proven to be difficult, both in the linear and non-linear case [21]–[29]. Many methods have been proposed in the Explainable AI (XAI) field, to provide explanations to a model's decision, through maps highlighting important areas for prediction - the so-called saliency maps [30]–[48]. However, these methods suffer from both *theoretical* and *empirical* difficulties. Theoretically, the main problem of these maps is that they locate the regions within the input image that were most important to make a certain prediction, being therefore useful for debugging purposes [23], but they do not reveal *why* the model was looking at that area [23], [25]. In an attempt to demonstrate this, several XAI methods have been tested in the task of retrieving the signal of interest in presence of noise with a specific pattern, on a linearly generated synthetic dataset [26]: many of them were not able to locate the signal, but rather extracted a mixture of signal and noise, demonstrating the difference between areas used for predictions and areas that are directly related to the signal. This difficulty also applies to spatial weight maps of seemingly simple *linear* discriminative methods, where voxels can be assigned a nonzero weight to amplify the signal of interest or to suppress noninteresting noise from the image, resulting in a non-interpretable spatial pattern [21], [26]. This also holds when inverting the proposed generative method in order to make predictions: voxels with zero weight in the generative model, and therefore not affected by the target variable, can obtain a large weight after model inversion, as we will show in the next section.

  The literature has also pointed out empirical difficulties with XAI methods: studies have reported cases of saliency maps that are insensitive to model parameters [24] or to input-output relationships [23], [24], [29], tending to highlight features, such as edges, that are unrelated to the prediction task [24], [28], and being often similar across different classes [25]. Additionally, cases of XAI methods failing in a localization task in medical images have been reported [22], therefore questioning even their ability of localizing interesting areas, as well as examples of instability of the highlighted features across different training runs [22].

  Given these difficulties in interpreting discriminative methods, and the importance in neuroimaging of providing insights into the underlying causes of predictions, a possible solution is to use methods that are *inherently* interpretable, such as generative models, instead of *post-hoc* explanations, without necessarily sacrificing prediction performances [25].

- **DL can be more difficult to use:**
  Compared to less expressive techniques, DL methods

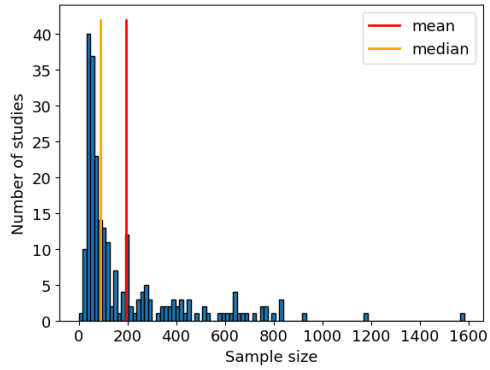


Fig. 1: Histogram of sample sizes in brain disorders prediction studies, as reported by [10].

[3]https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/world-s-largest-imaging-study-scans-50-000th-participant

are often harder to use, as they can be time consuming to train, even when using GPUs, and have many more "knobs" that need to be turned to obtain good results. For instance, a recent "lightweight" DL method that achieves state-of-the-art performances for brain age prediction [5] reports a training regime that includes choosing a good combination of data augmentation scheme, optimizer, training loss, batch size and other factors, and a training time of 65 hours for around 13,000 subjects, using two (NVIDIA P100) GPUs. We re-trained this method as state-of-the-art benchmark, in the same setting as in [5], and found that training the model was time consuming even for moderate sizes (almost one and two days on a (NVIDIA A100 SXM4) GPU, for 300 and 1000 training subjects, respectively) . Furthermore, for the given data resolution and batch size, training the model required to use a specialized GPU with large memory (40GB RAM), to which even state-of-the-art GPU clusters may not have access. Given these difficulties, there may be an intrinsic value in developing prediction methods that are easier to use, especially if they are less opaque and more efficient at learning from small sample sizes.

In this paper, we propose a lightweight generative model that aims to be easier to use and more straightforward to interpret, without sacrificing prediction performance in typical sample size settings. An early version of this work can be found in [49]. Like in the mass-univariate techniques that have traditionally been used in human brain mapping [50]–[53], the proposed method has a causal forward model that encodes how variables of interest affect brain shape, and is therefore intuitive to interpret. Unlike such techniques, however, the method also includes a linear-Gaussian latent variable noise model that captures the dominant correlations between voxels. As we will show, this allows us to efficiently "invert" the model to obtain accurate predictions of variables of interest, yielding an effective linear prediction method without externally enforced interpretability constraints [54], [55].

The method we propose can be viewed as an extension of prior work demonstrating that naive Bayesian classifiers can empirically outperform more powerful methods when the training size is limited, even though the latter have asymptotically better performance [56], [57]. Here we show that these findings translate to prediction tasks in neuroimaging when the strong conditional independence assumption of such "naive" methods is relaxed. Using experiments on age and gender prediction in the UK Biobank imaging dataset, we demonstrate empirically that, even when the number of training subjects is the thousands, our lightweight linear generative method yields prediction performance that is competitive with state-of-the-art nonlinear discriminative [5], linear discriminative [58], and nonlinear generative [59] methods. We then further investigate this comparison of performances in terms of bias-variance trade-off, giving insight into the reasons underlying our competitive prediction performances. Finally, with experiments on age prediction and on classification of multiple sclerosis patients vs healthy controls, we also demonstrate that the proposed method can be easily modified to incorporate possibly known covariates or nonlinearities in the target variable.

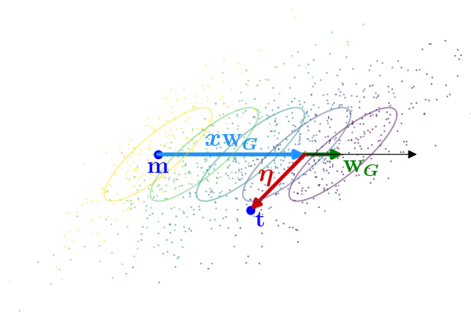


Fig. 2: Toy 2D illustration of the generative model in (1).

II. Method

In this section, we describe the core version of the proposed method, while possible extensions, such as inclusion of known subject-specific covariates or nonlinear dependencies on the variable of interest, will be discussed in section IV.

A. Generative model

Let $\mathbf{t} \in \mathbb{R}^J$ denote a a vector that contains the intensities in the J voxels of a subject's image, and x a scalar variable of interest about that subject (such as their age or gender). A simple generative model, illustrated in Fig. 2 and 3, is then of the form

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \boldsymbol{\eta}, \tag{1}$$

where $\boldsymbol{\eta} \in \mathbb{R}^J$ is a random noise vector, assumed to be Gaussian distributed with zero mean and covariance $\mathbf{C}$, and $\mathbf{w}_G, \mathbf{m} \in \mathbb{R}^J$ are two spatial weight maps that reflect how strongly the variable of interest x is expressed in the voxels of $\mathbf{t}$, and the baseline intensities (i.e., when $x = 0$), respectively. For the remainder of the paper, we will refer to $\mathbf{w}_G$ as the *generative* weight map, and collect the two spatial weight maps in a single matrix $\mathbf{W} = (\mathbf{m}, \mathbf{w}_G)$ for notational convenience.

Note that this is the model commonly assumed in traditional mass-univariate brain mapping techniques, such as voxel- and deformation-based morphometry [50], [52], where diagonal $\mathbf{C}$ is assumed and $\mathbf{w}_G$ is analyzed with statistical tests to reveal brain regions with significant group differences or related to specific variables of interest. In contrast, here we assume that $\mathbf{C}$ has spatial structure, allowing us, besides interpreting $\mathbf{w}_G$, to accurately predict x from $\mathbf{t}$ by inverting the model, as shown in the remainder.

B. Making predictions

When the parameters of the model ($\mathbf{W}$ and $\mathbf{C}$) are known, the unknown target variable x^* of a subject with image $\mathbf{t}^*$ can be inferred by inverting the model using Bayes' rule. For a binary target variable $x^* \in \{0, 1\}$, it is well-known that

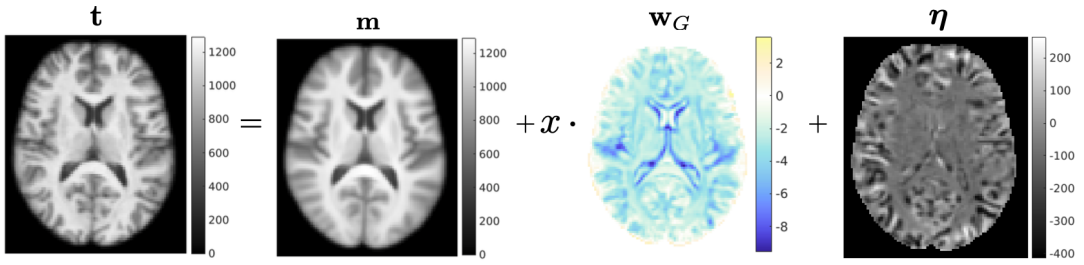


Fig. 3: Example of the image decomposition in (1) applied to age estimation (x denotes the difference between the age of the subject and the average age in the training set), for a 47 year old subject. The model parameters were estimated on a training set of 300 subjects (see section III-A).

the target posterior distribution takes the form of a logistic regression classifier [60]: Assuming the two outcomes have equal prior probability, we obtain (cf. Appendix I)

$$p(x^* = 1|\mathbf{t}^*, \mathbf{W}, \mathbf{C}) = \sigma\big(\mathbf{w}_D^T\mathbf{t}^* + w_o\big), \tag{2}$$

where

$$\mathbf{w}_D = \mathbf{C}^{-1}\mathbf{w}_G \tag{3}$$

are a set *discriminative* spatial weights, $\sigma(a) = 1/(1+e^{-a})$ denotes the logistic function, and $w_o = -\mathbf{w}_D^T(\mathbf{m} + \mathbf{w}_G/2)$. The maximum a posteriori (MAP) estimate of x^* is therefore 1 if

$$\mathbf{w}_D^T\mathbf{t}^* + w_o > 0, \tag{4}$$

and 0 otherwise.

For a continuous target variable with a flat prior $p(x^*) \propto 1$, the posterior distribution is Gaussian with variance

$$\sigma_x^2 = \big(\mathbf{w}_G^T\mathbf{C}^{-1}\mathbf{w}_G\big)^{-1} \tag{5}$$

and mean

$$y(\mathbf{t}^*) = \sigma_x^2(\mathbf{w}_D^T\mathbf{t}^* + b_0), \tag{6}$$

where $b_0 = -\mathbf{w}_D^T\mathbf{m}$ (cf. Appendix I). The predicted value of x^* is therefore given by (6), which again involves taking the inner product of the discriminative weights $\mathbf{w}_D$ with $\mathbf{t}^*$.

C. Model training

In practice the model parameters $\mathbf{W}$ and $\mathbf{C}$ need to be estimated from training data. Given N training pairs $\{\mathbf{t}_n, x_n\}_{n=1}^N$, their maximum likelihood (ML) estimate is obtained by maximizing the marginal likelihood

$$p\left(\{\mathbf{t}_n\}_{n=1}^N|\{x_n\}_{n=1}^N, \mathbf{W}, \mathbf{C}\right) = \prod_{n=1}^N \mathcal{N}\left(\mathbf{t}_n|\ \mathbf{m} + x_n\mathbf{w}_G, \mathbf{C}\right) \tag{7}$$

with respect to $\mathbf{W}$ and $\mathbf{C}$. For the spatial maps, the solution is given in closed form (cf. Appendix I):

$$\mathbf{W} = \left(\sum_{n=1}^N \mathbf{t}_n\boldsymbol{\phi}_n^T\right)\left(\sum_{n=1}^N \boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\right)^{-1} \text{ with } \boldsymbol{\phi}_n = (1, x_n)^T. \tag{8}$$

This amounts to performing a linear regression with two basis functions independently in each voxel. Obtaining the noise covariance matrix $\mathbf{C}$ directly by ML estimation is problematic, however: $\mathbf{C}$ has $J(J+1)/2$ free parameters – orders of magnitude more than there are training samples. To circumvent this problem, we impose a specific structure on $\mathbf{C}$ by using a latent variable model known as factor analysis [61]. In particular, we model the noise as

$$\boldsymbol{\eta} = \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon}, \tag{9}$$

where $\mathbf{z}$ is a small set of K unknown latent variables distributed as $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbb{I}_K)$, $\mathbf{V}$ contains K corresponding, unknown spatial weight maps, and $\boldsymbol{\epsilon}$ is a zero-mean Gaussian distributed error with unknown diagonal covariance $\boldsymbol{\Delta}$. Marginalizing over $\mathbf{z}$ yields a zero-mean Gaussian noise model with covariance matrix

$$\mathbf{C} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Delta},$$

which is now controlled by a reduced set of parameters $\mathbf{V}$ and $\boldsymbol{\Delta}$. The number of columns in $\mathbf{V}$ (i.e., the number of latent variables K) is a hyperparameter in the model that needs to be tuned experimentally.

Plugging in the ML estimate of $\mathbf{W}$ given by (8), the parameters $\mathbf{V}$ and $\boldsymbol{\Delta}$ maximizing the marginal likelihood (7) can be estimated using an Expectation-Maximization (EM) algorithm [62]. Defining $\tilde{\mathbf{t}}_n = \mathbf{t}_n - \mathbf{W}\boldsymbol{\phi}_n$, this yields an iterative algorithm that repeatedly evaluates the posterior distribution over the latent variables:

$$p(\mathbf{z}_n|\tilde{\mathbf{t}}_n, \mathbf{V}, \boldsymbol{\Delta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}) \tag{10}$$

where $\boldsymbol{\mu}_n = \boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\tilde{\mathbf{t}}_n$ and $\boldsymbol{\Sigma} = (\mathbb{I}_K + \mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V})^{-1}$, and subsequently updates the parameters:

$$\mathbf{V} \leftarrow \left(\sum_{n=1}^N \tilde{\mathbf{t}}_n\boldsymbol{\mu}_n^T\right)\left(\sum_{n=1}^N \left(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + \boldsymbol{\Sigma}\right)\right)^{-1} \tag{11}$$

$$\boldsymbol{\Delta} \leftarrow \text{diag}\left(\frac{1}{N}\sum_{n=1}^N \tilde{\mathbf{t}}_n\tilde{\mathbf{t}}_n^T - \mathbf{V}\frac{1}{N}\sum_{n=1}^N \boldsymbol{\mu}_n\tilde{\mathbf{t}}_n^T\right). \tag{12}$$

Here $\text{diag}(\cdot)$ sets all the non-diagonal entries to zero.

D. Practical implementation

With the proposed method, both making predictions and training the model involves manipulating matrices of size $J \times J$. Despite the high dimensionality (recall that J is the number of voxels), computations can be performed efficiently by exploiting the structure of these matrices: As detailed in Appendix III, training and predicting can be implemented in a way that only involves the posterior covariance of the latent variables $\mathbf{\Sigma}$, which is of much smaller size $K \times K$.

In our implementation, we center the target variable x, i.e., we subtract the sample mean of the training set $(\sum_{n=1}^{N} x_n)/N$ from x during both training and testing. This has the advantage that the estimated $\mathbf{m}$ is a template that reflects the average anatomy of the subjects in the training set. We initialize the EM algorithm by first computing, for each voxel, the variance across the training subjects in the centered training images $\{\tilde{\mathbf{t}}_n\}_{n=1}^{N}$. Each row in $\mathbf{V}$ is then initialized with random entries drawn from a zero-mean Gaussian with the corresponding variance. Similarly, the diagonal elements in $\mathbf{\Delta}$ are initially set to the corresponding voxel's variance. Convergence of the EM procedure is detected by checking whether the relative change in the log marginal likelihood drops below 10^{-5} between iterations.

III. Experiments and Results

In this section, we present experiments about age and gender prediction, performed with the core method described in section II. Possible extensions, including experimental results, will be described in section IV.

A. Data and experimental set up

We employed the proposed generative model for predicting age and gender using the UK Biobank dataset, which comprises MRI T1-weighted scans of 26,127 healthy subjects , aged 44-82 years. These scans were already preprocessed with skull stripping and bias field correction [6], and the nonlinear registration used to perform accurate skull stripping is provided. We therefore used it to nonlinearly warp the scans to MNI space, obtaining "deformable T1s". Since [5] makes predictions based on skull-stripped, bias-field corrected T1 scans from the UK Biobank, that are instead *affinely* registered to MNI space, for completeness we also trained our method on such affinely registered T1s ("affine T1s"). We however note that the use of affine T1s is very strange set up, since nonlinear deformations are provided and already used to perform accurate skull-stripping, whose result is in turn used in subsequent processing steps, such as bias field correction. Additionally, the use of affine T1s is expected to disadvantage linear methods (which are also used as benchmarks in [5]), since – unlike neural networks – they are not able to model nonlinear deformations that have not been removed from the input data by the the affine registration. Therefore, in our experiments we mainly focused on deformable T1s, although we also report results on affine T1s for completeness sake.

Following the study design in [5], for all the experiments, we selected a validation set of 500 subjects, and a test set of 1000 subjects. Additionally, in order to investigate prediction performances and explainability for different number of training subjects, we trained the proposed method on sets of different sizes from 100 subjects up to 7800 (for age) and 9800 subjects (for gender). For each size, we trained on 10 randomly sampled training sets (on only 3 sets for sizes larger than 1000 (i.e., $N > 1000$)). The method's one hyperparameter K is selected by using grid search on the validation set as the one that yields the smallest validation Mean Absolute Error (MAE) (for age) and the largest validation accuracy (for gender).

To speed up computations, we trained on downsampled data, specifically on 3mm isotropic - we empirically found that performances with 3mm and 2mm are comparable for the proposed method, and therefore we used 3mm to reduce training times.

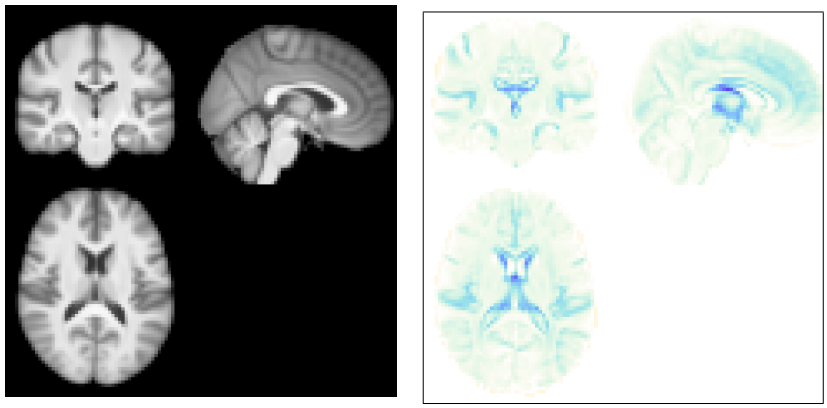
Before investigating prediction performances, we can give insight into the working of the porposed method by displaying the estimated forward model for age prediction on $N = 2600$ training subjects. The estimates obtained for $\mathbf{m}$ and $\mathbf{w}_G$ are displayed in Fig. 4: $\mathbf{m}$ represents the average image, while the generative weights $\mathbf{w}_G$ encode the direct effect of age in image intensities. Additionally, to give insight into the features captured by the noise model, we display in Fig. 5 the major modes of variation, encoded by the first three eigenvectors of the noise covariance matrix $\mathbf{C}$. Details about the computation of eigenvectors in high dimension are provided in Appendix I. We observe that the first eigenvector encodes a general darkening/brightening of image intensities, the second one seems to model residual bias field that has not been removed from the data in the preprocessing, and the third one expresses differences in the size of the lateral ventricles.

Regarding the estimate of the model's hyperparameter K which regulates the number of free parameters in the noise model, in Table I we report its optimal values selected on the validation set, averaged across all runs of each size, for age prediction on deformable T1s. We observe that the optimal value of K increases with the training size N; This is not surprising since in general larger training sets allow the use of more flexible models.

B. Benchmark methods

To compare prediction performances, we selected three benchmark methods: a discriminative nonlinear model (SFCN [5]), a discriminative linear method (RVoxM [58]) and a generative nonlinear one (variational auto-encoder [59]).

- **SFCN**: as discriminative nonlinear benchmark, we selected the SFCN, a lightweight convolutional neural network proposed in [5]. It is, to the best of our knowledge, the best performing method for image-based prediction, and it won the 2019 Predictive Analysis Challenge for brain age prediction. In [5], the SFCN is employed to predict age and gender on UK Biobank data, using several training sizes. Since the SFCN's training code is not publicly available (only the model code has been



$\mathbf{m}$ $\mathbf{w}_G$

Fig. 4: Estimate of $\mathbf{m}$ and $\mathbf{w}_G$ obtained for age prediction, by the proposed method trained on $N = 2600$ subjects on deformable T1s.

	N=100	N=200	N=300	N=500	N=1000	N=2600	N=5200	N=7800
K for age prediction	19.80	20.40	52.00	86.00	120.00	366.67	1833.33	3333.33

TABLE I: Optimal hyperparameters selected for the proposed method on the validation set, averaged across all the runs, for age prediction on deformable T1s.

released[4]), we used the SFCN implementation[5] provided by [63], with some modifications described in Appendix VI. To match the setting described in [5], we trained this method on 1mm isotropic scans. This method has one hyperparameter, which is the number of epochs used for training.

- **RVoxM**: It is a discriminative linear method, proposed in [58], which encourages sparsity and spatial smoothness of its weight map as a form of regularization. The strength of the spatial smoothness is controlled by the one hyperparameter of the model. We selected this method as discriminative linear benchmark because it provides competitive performances among the class of such methods, and it is comparable or it outperforms the best linear discriminative model tested in [5]. As implementation for this method, we used the code that is publicly available[6], with some adaptations described in Appendix VI. Similarly to our method, to speed up computations we trained the RVoxM on 3mm scans, after assessing that performances with 3mm and 2mm are comparable.
- **Variational auto-encoder**: as nonlinear generative benchmark, we selected a variational auto-encoder (VAE) that was recently proposed for age prediction [59]. This method is similar to ours, except that its latent variables contribute nonlinearly to the generative model, through a deep neural network, which makes the EM training algorithm more elaborate than our closed-form expressions [64]. This method contains two regularization hyperparameters (dropout factor and L2 regularization) and its training code is publicly available[7]. As in [59], we trained this model on 2mm T1 scans.

We trained the two discriminative benchmarks for predicting age and gender, using the same experimental set-up as for our method, i.e. training on varying sizes and using 10 different training sets for each size (only 3 sets for $N > 1000$). As for the proposed method, we selected these models' hyperparameters on the validation set, using grid search, by minimizing the validation MAE for age, and validation accuracy for gender.

Regarding the VAE, we employed a different set-up: we trained it only for age prediction, on T1 scans cropped around the ventricular area, following the same setting as in [59]. Additionally, since in [59] the VAE was proposed for around 200 training subjects, with number of latent variables hard-coded (to 12), we tested only training sizes in a similar range (from 100 to 400 training subjects). For each of these sizes, we trained the model on 10 randomly sampled training sets, and we estimated the model's hyperparameters on the validation set, as for the other methods. We also trained the proposed method using this same setting, to perform a comparison.

C. Prediction performance and training time

After training, we computed the average test MAE (for age) and test accuracy (for gender) across all the training runs of each size, for our method, RVoxM and SFCN. Results are

[4]https://github.com/ha-ha-ha-han/UKBiobank_deep_pretrain

[5]https://github.com/pmouches/Multi-modal-biological-brain-age-prediction/blob/main/sfcn_model.py

[6]https://sabuncu.engineering.cornell.edu/software-projects/relevance-voxel-machine-rvoxm-code-release/

[7]https://github.com/QingyuZhao/VAE-for-Regression

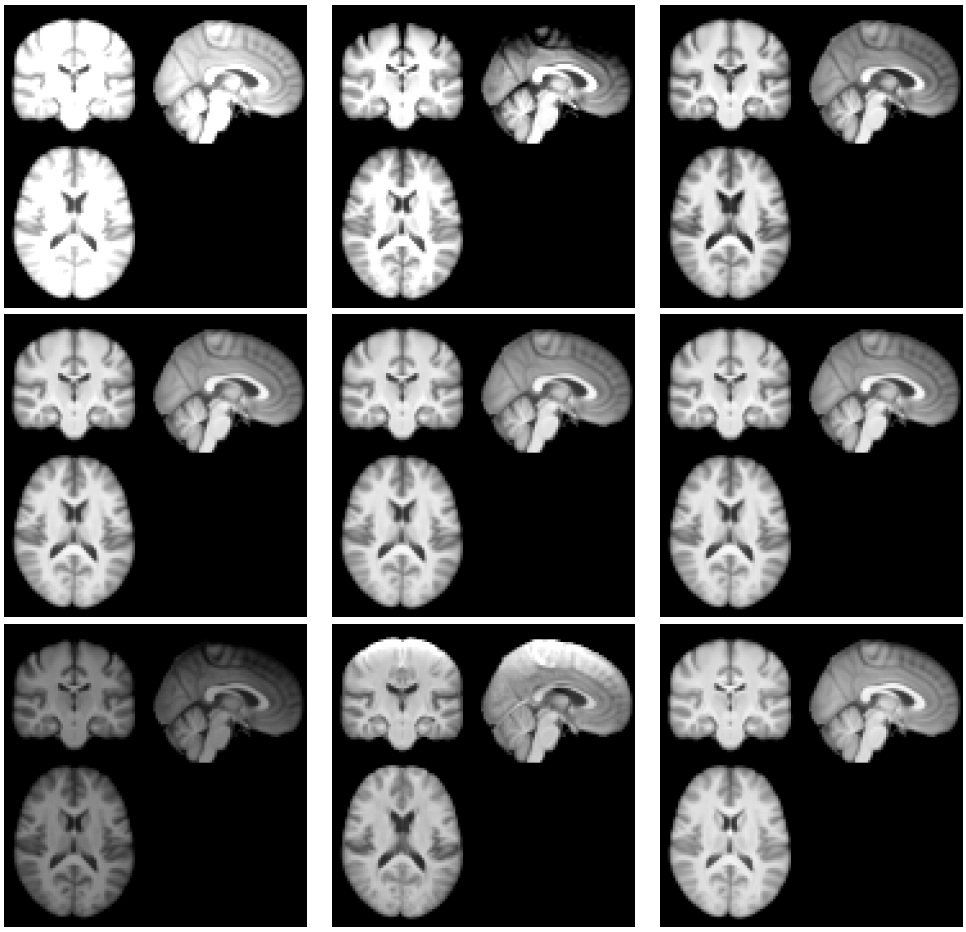


Fig. 5: Modes of variation encoded by the first three eigenvectors of the covariance matrix $\mathbf{C}$, obtained for age prediction, by the proposed method trained on N = 2600 subjects on deformable T1s. The middle line shows slices of the average image $\mathbf{m}$. The top ad bottom line display the average volume modified in direction of the eigenvectors, with negative and positive sign respectively.

shown in Fig. 6, for both age and gender prediction, based on deformable T1s. Because we retrained SFCN ourselves, for completeness we also show results as reported in [5]. The experimental setting in [5] is slightly different than in our re-implementation: only one training set for each size is employed, while we use multiple sets, and the specific choice of training sets, validation and test set is in general different from ours. Additionally, results in [5] are obtained using affine T1s, while performances displayed here for our own implementation are based on deformable T1s. However, the paper also shows that performances obtained on affine and deformable T1s are very similar, and we replicated this finding with our own implementation (Cf. Appendix V).

Comparing performances of different methods for age prediction, we observe that the proposed method and RVoxM achieve comparable performances, except for very large N, where the RVoxM starts achieving better results. Regarding the SFCN, we note that we are able to reproduce results in [5] for up to 1000 training subjects, with a small variability probably due to the use of different training and test sets, after which we start getting systematically larger errors. In this regard, we point out that we did our best to reproduce SFCN's results despite the unavailability of the training code. Comparing our method's and SFCN's performances, we observe that the proposed method performs comparably or better than the SFCN in regimes up to 2600 training subjects, after which the SFCN achieves better results (more markedly for the SFCN as reported in the paper).

Regarding gender prediction, we find that our method and the RVoxM perform equally for all tested training sizes. For the

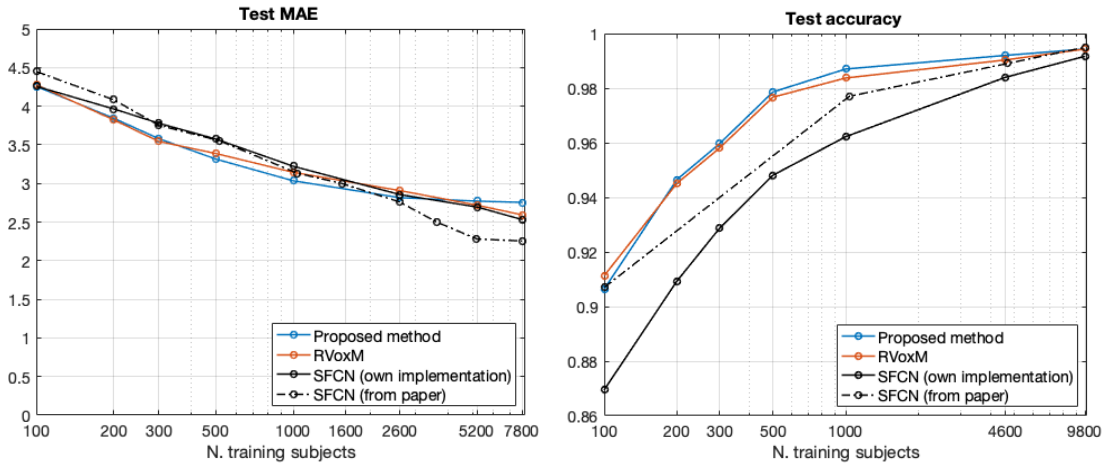


Fig. 6: Comparison of the proposed method, RVoxM and SFCN on an age prediction task (left) and on a gender classification task (right). For SFCN, we also display performances as reported in [5].

SFCN, we observe a systematic gap between our performances and results from [5], with a larger difference for $N = 100$ and $N = 1000$. However, the SFCN as reported in the paper, although it performs better than our own re-implementation, is not able to achieve the same prediction performances as our method, except for very large training sizes where the performances are comparable.

Additionally, training times for the three methods are shown in Table II, for age prediction. For small sizes, training the proposed method takes only a few minutes, being much faster than the other methods. As the training size increases, our method becomes comparable to the RVoxM for $N = 2600$ and $N = 5200$, and then slower for $N = 7800$. The SFCN is slower than the other methods, for any size. It should be noted that for proposed method and RVoxM, these are CPU times (with 3mm T1 scans), while for SFCN they are GPU times (for 1mm T1s).

Regarding the comparison with the VAE, in Fig. 7 we display the test MAE averaged across all training sets of each size, for both our method and VAE. We observe that the proposed method achieves better results for every tested training size, suggesting that, at least when the training set consists of a few hundred subjects, adding more flexibility to the model is not beneficial - it may even hurt performances. Regarding training times, training the VAE took on average 9.40 minutes for N=200 with the optimal set of hyperparameters, using a NVIDIA GeForce RTX 2080 Ti GPU (11 GB of RAM), while training time for the proposed method with N=200 was 1.16 minutes, with the selected value of the hyperparameter, using Matlab on a state-of-the-art desktop.

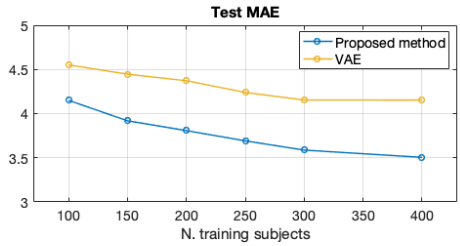


Fig. 7: Test MAE obtained by the proposed method and VAE for age prediction, on deformable T1s cropped around the ventricles.

D. Explainability

One of the main perks of the proposed method is that it produces an interpretable spatial map ($\mathbf{w}_G$), showing the direct effect of the variable of interest on image intensities, on a population level. This map is obtained through a decomposition of the signal (image) into an average anatomy, the effect of the variable of interest and a subject-specific noise. This decomposition is illustrated in Fig. 2 in a 2D toy example, and in Fig. 3 with images, for the age prediction task. Thanks to the form of this decomposition, the generative map $\mathbf{w}_G$ expresses how the variable of interest affects a subject's image, encoding target-related neuroanatomical changes [21]. Consistently with this, the generative map for age displayed in Fig. 3 expresses known age-related effects, such as gray matter atrophy and enlargement of ventricles [65], [66].

When the model is subsequently inverted, the discriminative weight map $\mathbf{w}_D$ (3) is obtained by combining the generative maps with the noise covariance matrix, and used to make predictions through a scalar product with the test subject's image, as illustrated in Fig. 8 for vectors and Fig. 9 for images. This discriminative map contains the weights given to voxels for predicting the variable of interest, and therefore highlights image areas the model uses for predictions. However, since it includes both generative effect and noise pattern, it does

	N=100	N=200	N=300	N=500	N=1000	N=2600	N=5200	N=7800
Proposed method	1.20 min	0.67 min	1.94 min	9.53 min	32.18 min	≈3h	≈15h	≈ 69 h
RVoxM	92.42 min	66.46 min	75.36 min	76.21 min	129.05 min	126.55 min	≈ 22 h	≈ 21 h
SFCN	≈ 8h	≈ 11 h	≈ 16 h	≈ 18 h	≈ 34h	≈ 76h	≈ 69h	≈ 102 h

TABLE II: Training times for age prediction on deformable T1s, for proposed method, RVoxM and SFCN, averaged across all training runs. For proposed method and RVoxM, the table displays CPU time for a single selected value of the models' hyperparameter, obtained with Matlab on a state-of-the-art desktop. For SFCN, the reported time is the training time up to the selected epoch, obtained with a NVIDIA A100 SXM4 GPU (40 GB of RAM).

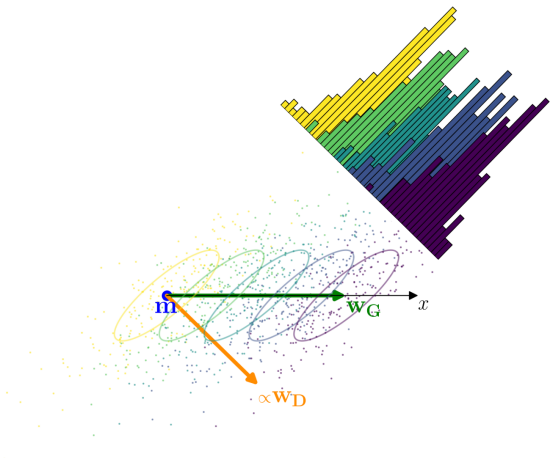


Fig. 8: Illustration of the inversion process of the toy generative model shown in Fig. 2: Measurements (indicated by individual points) are orthogonally projected onto the direction $\mathbf{w}_D = \mathbf{C}^{-1}\mathbf{w}_G$, as the resulting 1-dimensional signal optimally disentangles the variable of interest (illustrated by the histogram) in the presence of noise.

not directly express target-related changes in neuroanatomy and it results in an uninterpretable spatial pattern [21]. This concept is shown in Fig. 8, illustrating how the y channel has a large component in $\mathbf{w}_D$, although it is not affected by the target variable since its weight in $\mathbf{w}_G$ is zero. Similarly, the age discriminative map shown in Fig. 9 does not present the typical age-related patterns that characterize the generative map, but it mostly highlights white matter areas.

Fig. 10 shows other 2D slices of both generative and discriminative maps obtained for age prediction. We observe again the large difference between the two maps: while the generative maps display typical age-related effects, mostly highlighting gray matter borders and ventricles, the discriminative maps focus on very different areas, mainly within white matter, with both positive and negative weights.

It is also interesting to compare the generative map $\mathbf{w}_G$ not only with its discriminative counterpart $\mathbf{w}_D$, but also with spatial maps of the other discriminative methods employed as benchmarks. Additionally, we are interested in analyzing the stability of these spatial maps when changing training set, both of same and different sizes. For these reasons, we display in Fig. 11 spatial maps of the proposed method (both $\mathbf{w}_G$ and $\mathbf{w}_D$), RVoxM and SFCN, obtained on training sets of 300, 2600 and 7800 subjects. To also investigate the behaviour of these maps when changing training data within the same cohort size, in Fig. 12 we display maps obtained by the methods on three randomly sampled training sets of 2600 subjects. Since SFCN is a neural network and therefore does not automatically provide spatial maps, we used SmoothGrad [67] to compute saliency maps for this method. The SmoothGrad maps are commonly used as post-hoc explanations of deep learning models, and they can be seen as a generalization of linear methods' weight maps (i.e. the SmoothGrad map computed for a linear discriminative method would correspond exactly to the model's weight map [24]). Note that, since SmoothGrad provides subject-specific maps, in order to obtain a single template that could be compared with the other methods' ones, we averaged the SmoothGrad maps of all test subjects. This technique is considered a relevant way to produce population-level maps for instance-based XAI methods, since it removes the noise characterizing single-subject maps [26], consistent with the finding that it's the aggregate use of saliency maps rather than the individual one that can yield significant results [23].

First, from Fig. 11 and 12, we note that the discriminative spatial patterns of $\mathbf{w}_D$, RVoxM, and SFCN are much less intuitive than $\mathbf{w}_G$, which shows known age-related effects: Discriminative maps highlight regions that are most important to make predictions, but they do not directly express anatomical changes caused by the variable of interest. Additionally, regarding the dependency of maps on specific training data, we note that the generative maps $\mathbf{w}_G$ are quite stable across different training samples. The discriminative maps $\mathbf{w}_D$ and RVoxM's also show some consistency, especially when keeping the same training set size, but with more differences than the generative maps. A possible explanation is that $\mathbf{w}_D$ and the RVoxM's maps depend on an hyperparameter, which in general varies with the specific training data, especially when changing training size (cf. Table I), while $\mathbf{w}_G$ does not. Additionally, generative maps are produced by estimating two weights from N data point in each voxel independently, which is expected to yield rather stable fittings. Instead, $\mathbf{w}_D$ involves estimating many more basis functions, together with their coefficients, in a multivariate way, and RVoxM's maps entail fitting a very high-dimensional hyperplane from N data - which are both likely to be less stable operations. Regarding SFCN, we observe that there is a huge variability in its

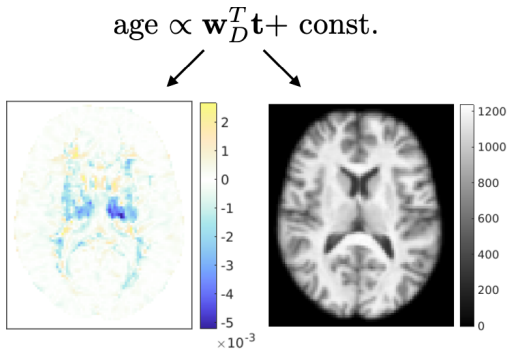


Fig. 9: Illustration of how a subject's age is estimated by inverting the model shown in Fig. 3. Note that the resulting discriminative linear regression is *mathematically* the same as decomposing the signal into the individual constituents shown in Fig. 3, but *not in terms of interpretability.*

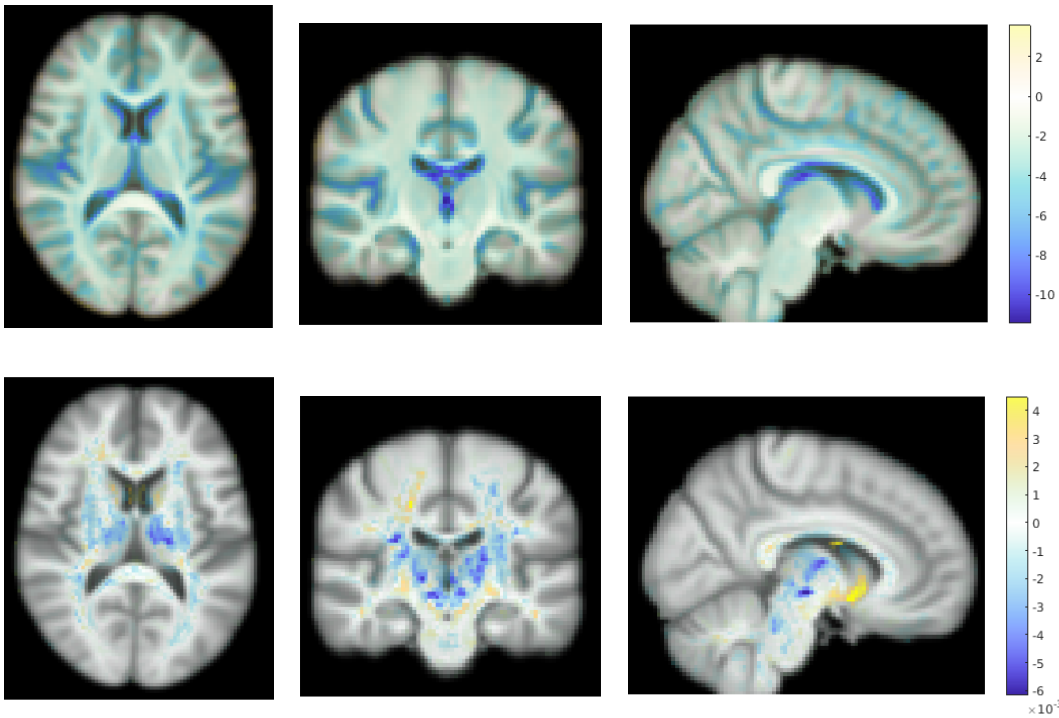


Fig. 10: 2D slices of generative map $\mathbf{w}_G$ (top) and discriminative map $\mathbf{w}_D$ (bottom), obtained for age prediction on UK Biobank data, on a training set of 300 subjects (with deformable T1s).

maps when changing training sets, both of same and different sizes. These findings seem consistent with [22], which shows that many commonly used saliency maps methods, including SmoothGrad, did not pass a test of reproducibility when the model is retrained with a different random initialization or with a different architecture yielding similar prediction performances. All these results illustrate some difficulties that arise in the interpretation of discriminative maps - both *theoretical* difficulties, since such maps do not express the causal effects of interest for interpretation purposes, and *empirical* ones, since especially SFCN maps rely heavily on the specific choice of training data.

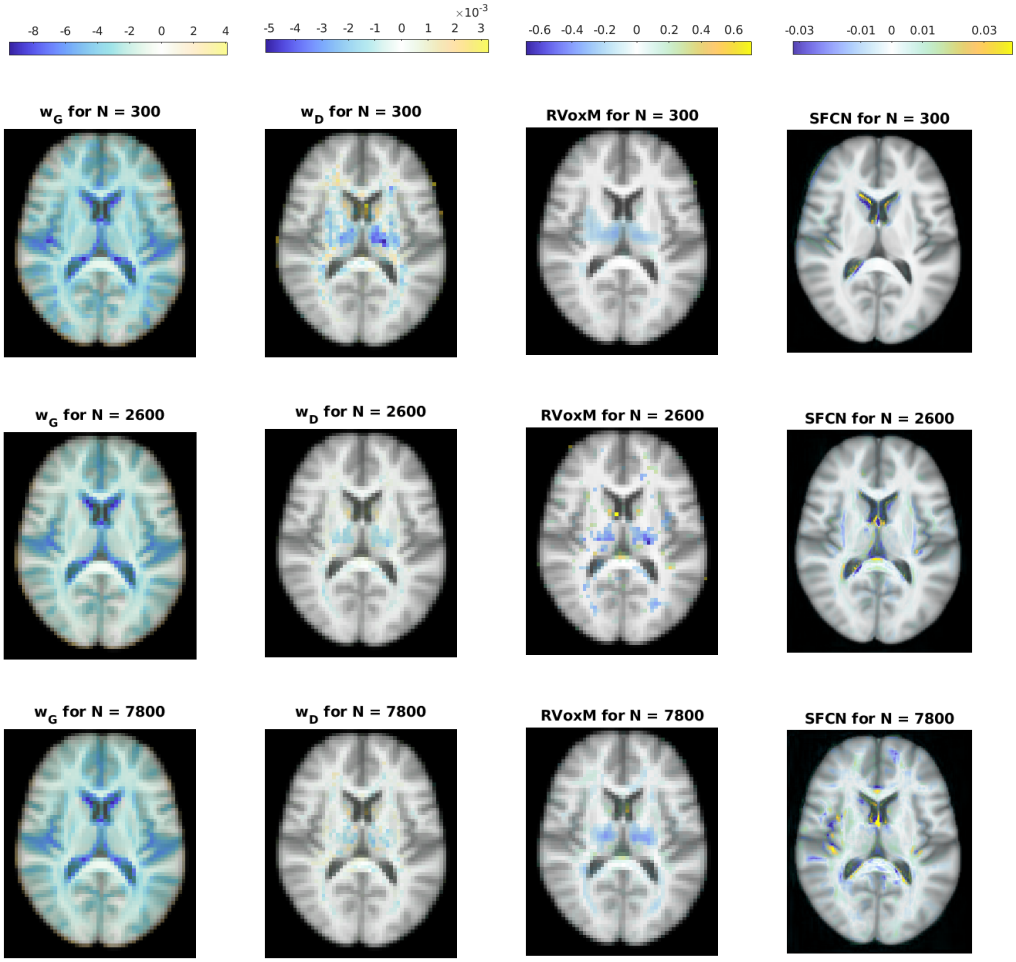


Fig. 11: Maps of proposed method ($\mathbf{w}_G$ and $\mathbf{w}_D$), RVoxM, and SFCN (with SmoothGrad), for different training set sizes. Voxels with zero weight are transparent. Discriminative maps are displayed for the optimal value of the hyperpameter, selected as described in section III-A and III-B.

In addition to visualizing $\mathbf{w}_G$, which reflects the average causal effect of age on brain morphometry on a population level, our generative model also allows us to generate subject-specific *counterfactuals* [68] – imaginary images of a specific individual if they had been younger or older than they really are. Specifically, given an image $\mathbf{t}$ and their real age x, (1) can be used to compute the noise vector $\boldsymbol{\eta}$, which captures the subject's individual idiosyncracies that are not explained by the population-level causal model. Counterfactual images can then by obtained by simply re-assembling the forward model ((1) and Fig. 3) from its constituent components, using a different, imaginary age x instead of the real one. An example of this process is shown in Fig. 13, where the brain of a 47 years old subject is artificially changed to 80 years, with the expected aging-related changes occurring. Such counterfactuals can therefore be used as an intuitive way to explain the age estimation process of our method in a particular individual, one that has been argued to closely match human intuition [68]: *A specific age is estimated because, had the subject been older or younger, their scan would have looked different.*

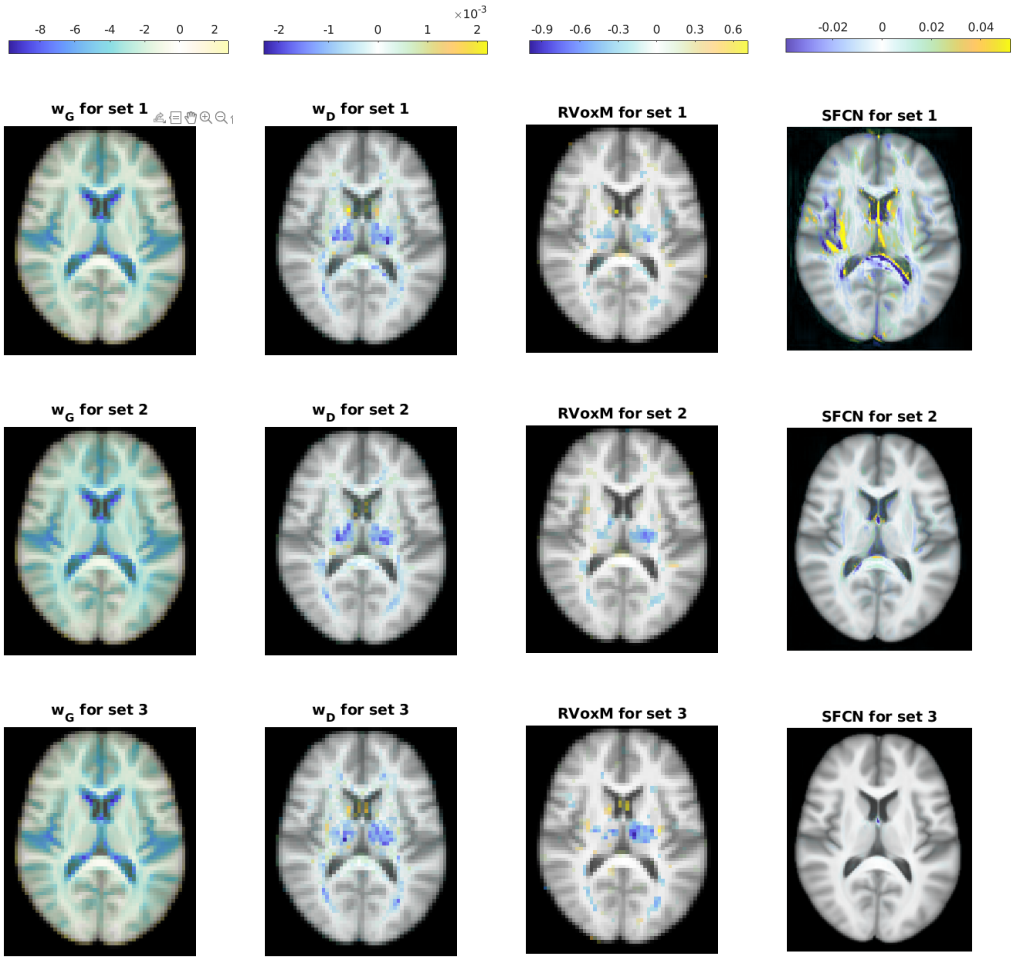


Fig. 12: Maps of proposed method ($\mathbf{w}_G$ and $\mathbf{w}_D$), RVoxM, and SFCN (with SmoothGrad), computed on 3 different training sets of 2600 subjects. Voxels with zero weight are transparent. Discriminative maps are displayed for the optimal value of the hyperpameter, selected as described in section III-A and III-B.

E. Bias-variance trade-off (gaining further insight)

In section III-C, we showed that prediction performances for age prediction of the proposed method and the two discriminative benchmarks are quite similar, in training regimes up to a few thousands of subjects. This finding is perhaps surprising if we consider the vastly different numbers of parameters in the methods. In fact, on one hand, the proposed method has $J(K+3) - K(K-1)/2$ free parameters (2 columns of J elements in $\mathbf{W}$, K columns of J parameters in $\mathbf{V}$ and J diagonal elements in $\mathbf{\Delta}$, which are reduced by $K(K-1)/2$ because any rotation in the latent space provides the same model [61]), with $J \approx 80,000$ in our experiments and K that varies from tens to thousands, depending on the training set size (Cf. Table I). On the other hand, the RVoxM has J free parameters, and the SFCN has 3 million parameters [5].

However, there are a lot of factors underlying a method's prediction performances besides the number of parameters, and in general it is not straightforward to predict which method will perform best a priori. For instance, the choice of the optimizer and how it explores the hypothesis space has an impact on a method's performances, where optimizers that try fewer hypothesis, restrict the hypothesis space and act as regularizers [69]. Additionally, the posterior distribution of a "wrong" generative model can still give correct predictions [56], and even on simulated data, an incorrect model can achieve better prediction performances than the "true" model in certain regimes [69], making an a priori guess of the best-performing method extremely hard.

Given these difficulties, to gain more insight in the perfor-

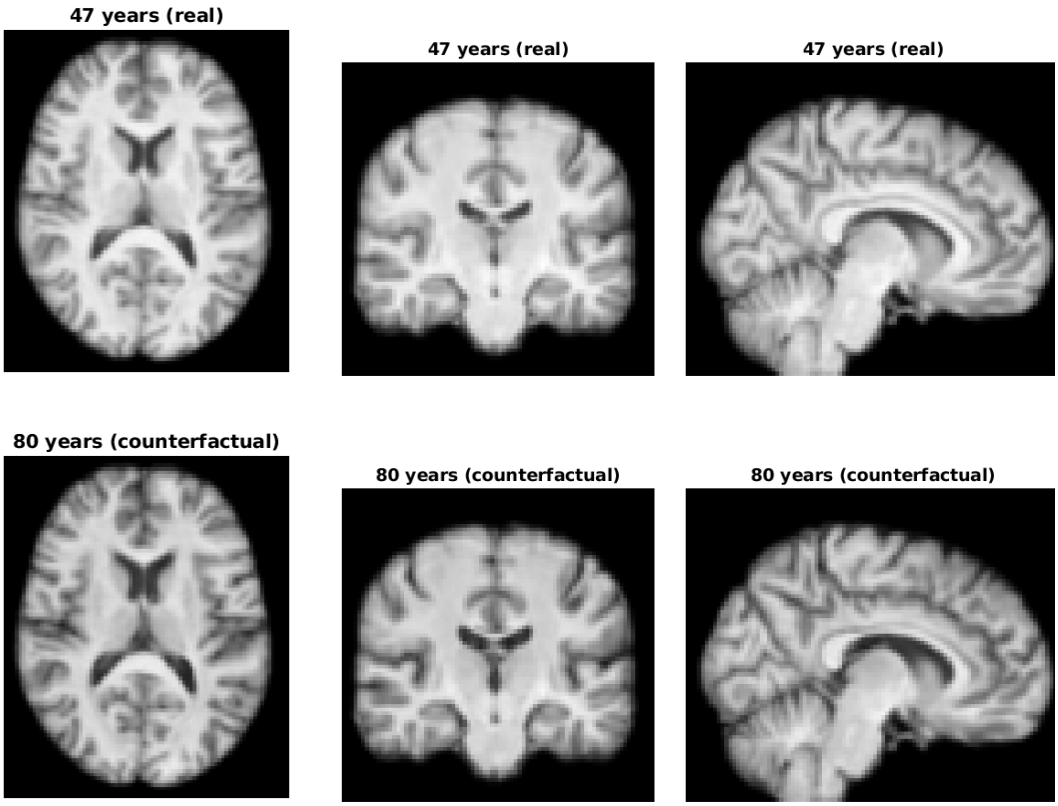


Fig. 13: Top: original images of a 47 years old subject. Bottom: counterfactual images of the same subject at the age of 80. All images are on the same intensity scale. The aging patterns shown in the counterfactual image are consistent with the effects encoded by $\mathbf{w}_G$: the aged brain is characterized by larger ventricles, enhanced gray matter atrophy, and a general slight decrease in image intensities.

mances of the considered methods, we can perform the so-called bias-variance decomposition of prediction errors. We compute it for age prediction, since it is more straightforward for a continuous variable than in a classification case. For a given method, the prediction mean squared error (MSE) can be decomposed into a *bias* term, which denotes how well the method performs on average, and a *variance* term, which indicates how much predictions for the same test subjects change across different training runs [70], [71]. In particular, if we consider a test pair $(\mathbf{t}^*, x^*)$, and we denote with $y(\mathbf{t}^*; D)$ the prediction made by the model trained on a dataset D for test subject $\mathbf{t}^*$, we obtain the following decomposition:

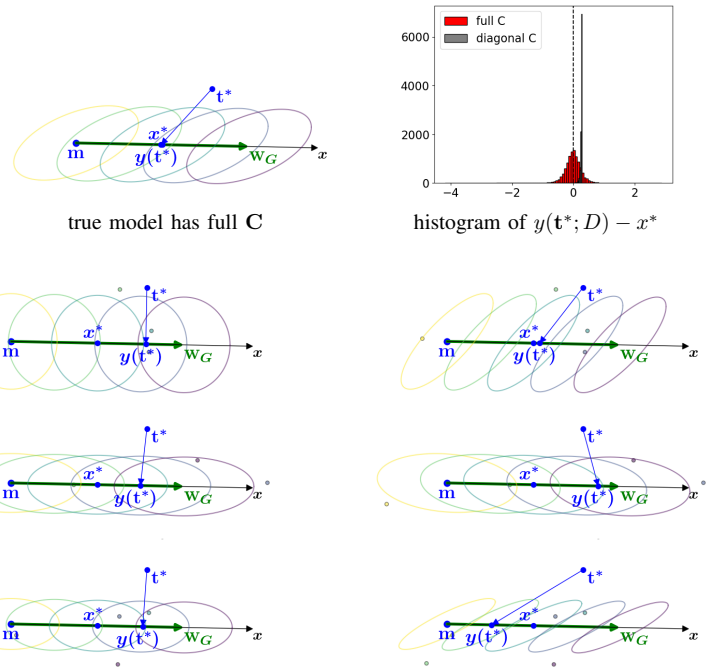
$$\underbrace{\mathbb{E}_{\mathbf{t}^*,D}\left[(x^* - y(\mathbf{t}^*;D))^2\right]}_{MSE} = \underbrace{\mathbb{E}_{\mathbf{t}^*}\left[(x^* - \mathbb{E}_D\left[y(\mathbf{t}^*;D)\right])^2\right]}_{bias} + \underbrace{\mathbb{E}_{\mathbf{t}^*,D}\left[(y(\mathbf{t}^*;D) - \mathbb{E}_D\left[y(\mathbf{t}^*;D)\right])^2\right]}_{variance} \quad (13)$$

where $\mathbb{E}_D[\cdot]$ denotes the expected value over all training sets D of a fixed size, and $\mathbb{E}_{\mathbf{t}^*}[\cdot]$ denotes the expected value over all possible inputs $\mathbf{t}^*$. In practice, if we consider M test pairs $\{\mathbf{t}^*_m, x^*_m\}_{m=1}^M$, and B different training sets $\{D_b\}_{b=1}^B$ of a given size, we can write for test subject m:

$$\frac{\sum_{b=1}^B \left(x^*_m - y(\mathbf{t}^*_m; D_b)\right)^2}{B} = \left(x^*_m - \bar{y}(\mathbf{t}^*_m)\right)^2 + \frac{\sum_{b=1}^B \left(y(\mathbf{t}^*_m; D_b) - \bar{y}(\mathbf{t}^*_m)\right)^2}{B} \quad (14)$$

where we have defined the mean prediction for test subject m as $\bar{y}(\mathbf{t}^*_m) = \sum_{b=1}^B y(\mathbf{t}^*_m; D_b)/B$. We can than average the decomposition in (14) over all test subjects.

Typically, a very flexible model will have a large variance and a low bias, reflecting an *overfitting* of the training data, while a strongly constrained method will have the opposite behaviour, resulting in *underfitting* of the training data [70], [71]. Finding the right balance in the bias-variance trade-off is a key point to achieve good results in a given setting, and



true model has full $\mathbf{C}$ histogram of $y(\mathbf{t}^*; D) - x^*$

fit with diagonal $\mathbf{C}$ fit with full $\mathbf{C}$

Fig. 14: Visualization of the bias-variance principle: Fitting a model with diagonal $\mathbf{C}$ (wrong model) yields predictions that are consistent across training runs, but systematically wrong (low variance, high bias), while fitting a model with full $\mathbf{C}$ (correct model) yields predictions that are more variable but on average correct. In the end, though, both models reach almost identical prediction performance. This principle is illustrated by the histogram of prediction errors, and by the cartoon examples, which display model inversion: The test data point $\mathbf{t}^*$ is projected orthogonally onto the direction of $\mathbf{w}_D$ to obtain predictions $y(\mathbf{t}^*)$, while x^* indicates the real target. These models are fitted using 5 data points for training, which are displayed as dots in the shown examples.

there is no method that can be *in absolute* better than others ("no free lunch" theorem) [69], [70].

The bias-variance decomposition principle is illustrated in Fig. 14, in a 2D toy example from the proposed method. We generate 5 data points for several training sets using a full covariance matrix, and we consider one specific test subject $(\mathbf{t}^*, x^*)$ drawn from the same distribution. We then predict the target variable for the test subject, fitting both a diagonal (*wrong* model) and a full (*correct* model) covariance matrix to the training sets. The histogram shows the distribution of the signed prediction error $y(\mathbf{t}^*; D) - x^*$ for the two models, over 10.000 training runs. The overall MSE is similar in the two cases (MSE $=$ 0.068 vs MSE $=$ 0.072), but the error distribution is very different: the model with diagonal $\mathbf{C}$ yields predictions that are very similar across training runs but systematically wrong (low variance, high bias), while predictions obtained by the more flexible model with full $\mathbf{C}$ vary more across training runs but they are on average correct (high variance, low bias). This is also illustrated by the three shown examples.

In order to gain a similar insight in the real-data experiment, we computed MSE, bias and variance with (14), for proposed method, RVoxM and SFCN, using the same training runs as described in section III-A and III-B for age prediction. The training sets in (14) are therefore the same that we used for assessing prediction performances (10 for sizes up to $N = 1000$ and 3 for larger sizes). We then averaged the decomposition in (14) over the 1000 subjects of our usual test set. The computed decomposition is shown in Fig. 15 (left). Let us first analyze the decomposition of our method (blue lines). We observe that the bias is reduced as the training set size increases, and the variance slightly decreases as well. This behaviour is achieved through the method's regularization hyperparameter K: for small training sizes, the hyperparameter constrains the models in order to control the variance, and this results in a larger bias. As the training size increases, the variance is naturally reduced thanks to the larger number of training subjects [70]. This allows the method to be less regularized - as we saw in Table I, the value of K increases as N becomes larger - and thus to achieve a smaller bias.

If we now consider the decomposition of RVoxM and

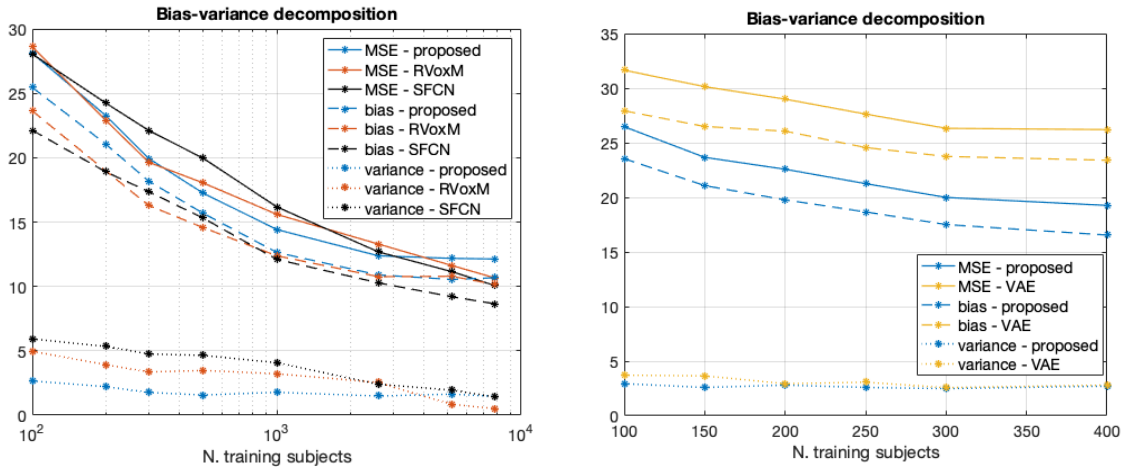


Fig. 15: Left: Bias-variance decomposition for the proposed method, RVoxM and SFCN. Right: Bias-variance decomposition for the proposed method and VAE. We used the same training runs as described in section III-A and III-B.

SFCN (red and black lines, respectively), we observe a similar behaviour as in our method, with decreasing bias and variance as N increases.

If we compare the decomposition of the three methods, we observe that the our method's variance is smaller than the others, except for very large N, where the benchmarks' variances reach (and become smaller than, in the RVoxM case) the proposed method's one. Instead, the proposed method's bias is larger than the other methods' counterpart, with some training sizes where they are comparable, especially for the RVoxM. This behaviour is in general expected since the proposed method is less flexible than other two, and therefore has a higher bias and a smaller variance, while for larger N, all methods can achieve a small variance thanks to the large number of subjects. The bias-variance trade-off is therefore a tool for interpreting prediction performances: a simpler model like the proposed method is competitive or it even outperforms the much more powerful SFCN with training sizes up to a few thousand subjects, because, although its strong assumptions make it on average incorrect (large bias), they also prevent it to overfit (small variance), and this compensates and possibly overcomes the large bias. Conversely, for larger training sizes, there is less risk of overfitting even for a flexible model such as the SFCN, and thus its smaller bias becomes decisive to obtain better prediction errors. These findings are in line with previous studies showing that a more powerful method is not necessarily better than a simpler one, and that, when the training size is limited, models with stronger assumptions - even if incorrect - may yield better performances than more flexible methods, because the latter overfit more [69].

We also computed the bias-variance decomposition of age prediction errors for the VAE. We again used the same training runs as in section III-B, i.e. deformable T1s cropped around the ventricular area as input data, and 10 training sets for each size, in a reduced range (from $N = 100$ to $N = 400$). We also computed the decomposition of our method, trained in the same setting. The computed MSE, bias and variances and displayed in Fig. 15 (right). We observe that the VAE has a slightly larger variance and a much larger bias then the proposed method. Therefore the VAE's worse performances than our method's for age prediction reported in section III-C are explained mostly by the VAE's higher bias.

IV. Extensions

The generative model proposed in (1) expresses a linear dependency of the voxels' intensities on the target variable. However, in some applications it can be of interest to also model nonlinear effects that characterize the images. Furthermore, in some cases additional information are available about the subjects, and it might be beneficial for predictions to take them into account. Therefore, in this section we show how the proposed model can be easily extended to include nonlinear effects of the variable of interest, and to incorporate subjects-specific covariates that are possibly known.

A. Known covariates

In some cases, subject-specific covariates are known and can be taken into account to build a more accurate model. Assuming each subject has L known covariates $y^1, \dots, y^L$, the model (1) can be extended to

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \sum_{l=1}^{L} y^l \mathbf{w}_y^l + \boldsymbol{\eta}, \tag{15}$$

where $\{\mathbf{w}_y^l\}_{l=1}^L$ are L extra spatial weight maps that also need to be estimated from training data. During training, the corresponding $\mathbf{W} = (\mathbf{m}, \mathbf{w}_G, \mathbf{w}_y^1, \dots, \mathbf{w}_y^L)$ can be estimated using (8), provided that $\boldsymbol{\phi}_n = (1, x_n, y_n^1, \dots, y_n^L)^T$ is used instead of $\boldsymbol{\phi}_n = (1, x_n)^T$. The updates of $\mathbf{V}$ and $\boldsymbol{\Delta}$ are still given by (11) and (12), where the estimated $\mathbf{W}$ and

$\boldsymbol{\phi}_n = (1, x_n, y_n^1, \ldots, y_n^L)^T$ are used. To predict an unknown variable of interest x^* from a subject with image $\mathbf{t}^*$ and known covariates $y^{*1}, \ldots, y^{*L}$, (4) and (6) remain valid, but with $\left(\mathbf{t}^* - \sum_l y^{*l}\mathbf{w}_y^l\right)$ replacing $\mathbf{t}^*$.

We explored the impact of including known variables into the model by adding age and gender as covariates in a classification experiment of multiple sclerosis (MS) patiens vs healthy controls. We performed this experiment using a private dataset from Klinikum rechts der Isar (Munich, Germany), from which we extracted T1-weighted scans of 131 MS subjects and 131 healthy controls, age- and sex-matched, obtaining a dataset of 262 subjects. We pre-processed all the scans with SPM12, in order to produce gray matter segmentations, modulated and warped to a standard template space. Since the occurrence of white matter lesions in MS patients is known to yield tissues misclassifications, as input to the segmentation pipeline we used a lesion-filled version of the T1-weighted scans that was available in the dataset. The obtained gray matter images were then used to discriminate between MS subjects and healthy controls. In order to perform a comparison in an unbiased way, we implemented the possible inclusion of age and gender in the model as an additional binary hyperparameter, which is estimated together with the latent space size. This model is then compared to the version where no covariates are used. In both cases, two-nested cross-validation loops are performed: On one hand, since the small training set size (262 subjects) prevents us to split the data into training and test set, we need to perform cross-validation to assess predictions performances in an unbiased way. On the other hand, an additional cross-validation loop is needed to estimate the model's hyperparameter(s), (since there are not enough data to create a separate validation set for hyperparameter selection.)

Results are shown in Table III. In all the folds, the cross-validation procedure selected the model with age and gender as covariates, which yields a slight improvement in performances as compared to the baseline model.

We also tested the possible inclusion of covariates on UK Biobank data, by adding age as known variable in gender prediction experiments based on deformable T1s. In this case, we found that the model with the extra hyperparameter regulating the inclusion of age achieved similar performances as the baseline model for every tested training size (from $N = 100$ to $N = 9800$), possibly because, even if age is not included in the model, the method automatically models its variability in the noise component.

It is worth noting how easy it is for the proposed method to incorporate known variables into the model, as opposed to discriminative neural networks, that would require to select in which network layer to insert them, as done for instance in [72] for inclusion of gender.

B. Nonlinearities in the causal model

For regression problems, the model (1) can also be extended to include nonlinear dependencies on the variable of interest:

$$\mathbf{t} = \mathbf{m} + x\mathbf{w}_G + \sum_{q=1}^{Q} f_q(x)\mathbf{w}_f^q + \boldsymbol{\eta}, \tag{16}$$

where $\{f_q(\cdot)\}_{q=1}^Q$ are Q nonlinear functions, and $\{\mathbf{w}_f^q\}_{q=1}^Q$ are their corresponding spatial weight maps. During training, each $f_q(x)$ can be treated as a known covariate, and therefore the same training procedure as in section IV-A applies. Once trained, estimating x^* from an image $\mathbf{t}^*$ is no longer governed by the linear equation (6), however inverting the model can still proceed by finely discretizing x^* into P possible values $x_p, p = 1, \ldots, P$, and evaluating the posterior probability of each:

$$p(x^* = x_p|\mathbf{t}^*, \mathbf{W}, \mathbf{V}, \boldsymbol{\Delta}) = \frac{\mathcal{N}(\mathbf{t}|\mathbf{m} + x_p\mathbf{w}_G + \sum_{q=1}^{Q} f_q(x_p)\mathbf{w}_f^q, \mathbf{C})}{\sum_{p'=1}^{P}\mathcal{N}(\mathbf{t}|\mathbf{m} + x_{p'}\mathbf{w}_G + \sum_{q=1}^{Q} f_q(x_{p'})\mathbf{w}_f^q, \mathbf{C})}, \tag{17}$$

where (20), (21), and (22) (Cf. Appendix III) can be used to evaluate the Gaussian distributions in (17). The prediction can then be obtained as the expected value: $\sum_{p=1}^{P} x_p p(x^* = x_p|\mathbf{t}^*, \mathbf{W}, \mathbf{V}, \boldsymbol{\Delta})$.

In order to investigate the effect of a nonlinear causal model, we performed age prediction on the IXI dataset, a publicly available collection of around 600 T1-weighted MRI scans from healthy subjects, aged 20-86 years. Since aging is known to have an approximately quadratic effect across adulthood on some brain structures [73] [74], and the IXI dataset covers an age span that is large enough to possibly show this behaviour, we used IXI data to test a causal model with quadratic dependency on age. As in the previous experiment, we used SPM12 to compute grey matter segmentations, modulated and warped to a standard template space. After performing segmentations quality control and implementing some exclusion criteria (e.g., removing subjects with unknown age), we obtained a dataset of 562 grey matter images of healthy subjects. As in the previous experiment, we used two-nested cross-validation loops to estimate a binary hyperparameter regulating the possible inclusion of a quadratic term into the model, together with the number of latent variables. This model is then compared to the linear version of the proposed method, without the additional hyperparmeter, trained on the same input data. In the quadratic model, predictions are made using a discretization of the age range in 20 bins. Results are shown in Table IV. In all the CV folds, a quadratic causal model was selected, yielding better performances as compared to the linear version of the model.

Additionally, we note that MAEs obtained on IXI data are larger than on UK Biobank, for similar training set sizes. In this regard, apart from the different type of input data (GM vs T1), we should take into consideration that the IXI dataset has a larger age range, and this intrinsically yields larger prediction errors [1].

We also tested the quadratic model for age prediction on deformable T1s from the UK Biobank, for several training sizes. In this case, including a quadratic dependency did not affect results except for very large training sets, where it yielded a slight improvement in test performances (MAE = 2.62 years with extra binary hyperparameter, which selected the quadratic version, vs MAE = 2.71 years with linear model, for $N = 7800$). A possible explanation for these results lies in the limited age span of the UK Biobank, which therefore

	CV accuracy	CV AUC	CV sensitivity	CV specificity
Without covariates	0.7023	0.7645	0.6718	0.7328
With extra hyperparameter (regulating covariates)	0.7214	0.7669	0.6794	0.7634

TABLE III: Performances achieved on the MS vs healthy classification task on 262 subjects from the Munich dataset, with two nested 5-fold CV loops. With the additional hyper-parameter regulating the use of covariates, age and gender are added in all the folds. AUC denotes the area under the ROC curve.

	CV MAE	CV RMSE	CV correlation
Linear model	4.7335	5.9283	0.9330
With extra hyperparameter (for quadratic vs linear model)	4.3627	5.4322	0.9445

TABLE IV: Performances for age prediction from GM images, on the IXI dataset, with two nested 5-fold CV loops. When using the additional hyper-parameter encoding a linear or quadratic causal model, the quadratic model is selected in all the folds.

requires very large sample sizes to observe a quadratic effect in the data.

V. Discussion and Conclusion

In this paper, we proposed a generative method for image-based predictions, which directly models how the variable of interest affects image intensities. It also includes a linear noise model, and it possibly incorporates nonlinearities in the target variable and/or additional covariates that may be known about the subjects. The model is subsequently "inverted" to make predictions using Bayes' rule, which is numerically possible even in case of added nonlinearities.

In the experiments performed for age prediction based on brain MRI scans, we proved that the proposed method for *regression* achieves competitive performances as compared to discriminative state-of-the-art models, for training set sizes up to a few thousands of subjects, which is the typical scenario in many neuroimaging applications. Furthermore, in the *classification* task of gender prediction, we showed that it is competitive with state-of-the-art methods for every tested training set size (up to 9800 training subjects). We also gave insight into the different methods' performances for age prediction in terms of bias-variance decomposition. We demonstrated that the proposed method, since it makes stronger assumptions, has in general larger bias and smaller variance than the discriminative benchmarks, with its small variance being the key feature behind its competitive performances. Additionally, we showed that, as compared to discriminative benchmarks, the proposed method has the advantage of being easier to use, less opaque and faster to train.

Other generative models were proposed in the literature for brain age prediction [59], [75], but, unlike our method, they include deep nonlinearities. The VAE proposed in [59] can be regarded as a deep nonlinear version of our method, where latent variables and the variable of interest are expanded nonlinearly through a neural network. The generative model proposed in [75] instead utilizes normalizing flows to model the bidirectional functional relationship between age and brain morphology. Adding deep nonlinearities in a generative method does not seem beneficial for brain age prediction. We showed this for VAE in section III-C, and, although we did not explicitly compare our method against [75], their reported MAE as percentage of age range (6.3% with N=4281) is similar to ours (7.4% with N=2600 and 7.3% with N=5200). Furthermore, including nonlinearities makes more difficult to visualize age-related morphological changes captured by the model. In fact, nonlinear generative methods can still provide interpretable maps, for instance by generating age-conditioned templates, from which age-related changes can be extracted using the jacobian determinant, as in [59], [75]. However, this visualization is age-gap dependent and it requires a lot of sampling and computations. Additionally, [75] produces an interpretable attribution map for its method, by computing the partial derivative of the inverse map with respect to age, but this is again computationally heavy and it results in a less intuitive spatial pattern than our spatial map. Furthermore, like the proposed method, nonlinear generative models can create counterfactual images, which allow to illustrate aging patterns encoded by the model, in a what-if scenario on a subject-specific level [76]. And yet, despite all these visualization techniques, the simplicity wherewith our method provides a single template expressing target-related anatomical changes at a group level cannot be attained by nonlinear models. Besides this advantage, the proposed method is also easier to use than its deep nonlinear counterparts, with less time and resources needed for training.

Among drawbacks of the proposed model, it should be mentioned that, while the method is well suited for scenarios with up to a few thousands of training subjects, for bigger sizes training the model becomes quite slow (on a CPU) and its performances are less competitive in regression tasks.

Possible future extensions of the proposed work include re-using part of a trained model in a separate task. In particular, we could train the model on a huge cohort of healthy subjects, and simply re-use the estimated noise model in a task involving a small cohort, e.g. classification of a certain disease vs healthy controls, where small sample sizes are the typical scenario. In the small cohort, only the causal part of the model, e.g the disease effect, would be estimated, resulting in a sub-second speed training. In fact, while learning the noise model is the time-consuming part of training, estimating the causal part is almost immediate. This technique has also the advantage of

providing a supposedly more accurate noise model, since this would be estimated on a large cohort.

Another possible future extension consists of modifying the proposed model to work in a longitudinal setting, where several scans per subject are available. Our method is particularly suited for a longitudinal scenario, since, unlike discriminative models, it can easily deal with inconsistent number of images per subject and time intervals between follow-up scans, which is the typical scenario in longitudinal studies. The longitudinal extension of our method can be obtained through mixed-effect models, where the temporal correlation between scans of the same subject is explicitly modelled.

Finally, in this paper the number of latent variables in the noise model was set using cross-validation, while it could be automatically estimated from training data using variational methods [61], without the need of re-training the model many times in a grid search procedure.

APPENDIX I
MAKING PREDICTIONS

Here we derive the expressions for making predictions about the variable of interest. For a binary target variable x^* with prior $p(x^*=0)=p(x^*=1)=0.5$, we have

$$\begin{aligned} & p(x^*{=}1|\mathbf{t}^*,\mathbf{W},\mathbf{C}) \\ = & \frac{p(\mathbf{t}^*|x^*{=}1,\mathbf{W},\mathbf{C})p(x^*{=}1)}{p(\mathbf{t}^*|x^*{=}1,\mathbf{W},\mathbf{C})p(x^*{=}1)+p(\mathbf{t}^*|x^*{=}0,\mathbf{W},\mathbf{C})p(x^*{=}0)} \\ = & \frac{1}{1+\frac{p(\boldsymbol{t}^*|x^*=0,\boldsymbol{W},\boldsymbol{C})}{p(\boldsymbol{t}^*|x^*=1,\boldsymbol{W},\boldsymbol{C})}} \\ = & \sigma\Big[\log p(\mathbf{t}^*|x^*{=}1,\mathbf{W},\mathbf{C})-\log p(\mathbf{t}^*|x^*{=}0,\mathbf{W},\mathbf{C})\Big] \end{aligned}$$

where

$$\begin{aligned} & \log p(\mathbf{t}^*|x^*{=}1,\mathbf{W},\mathbf{C})-\log p(\mathbf{t}^*|x^*{=}0,\mathbf{W},\mathbf{C}) \\ = & -\frac{1}{2}(\mathbf{t}^*-\mathbf{m}-\mathbf{w}_G)^T\mathbf{C}^{-1}(\mathbf{t}^*-\mathbf{m}-\mathbf{w}_G) \\ & +\frac{1}{2}(\mathbf{t}^*-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{t}^*-\mathbf{m}) \\ = & \mathbf{w}_G^T\mathbf{C}^{-1}(\mathbf{t}^*-\mathbf{m})-\frac{1}{2}\mathbf{w}_G^T\mathbf{C}^{-1}\mathbf{w}_G, \end{aligned}$$

which explains (2).

For a continuous target variable with flat prior, the log-posterior

$$\begin{aligned} & \log p(x^*|\mathbf{t}^*,\mathbf{W},\mathbf{C}) \\ = & -\frac{1}{2}(\mathbf{t}^*-\mathbf{m}-x^*\mathbf{w}_G)^T\mathbf{C}^{-1}(\mathbf{t}^*-\mathbf{m}-x^*\mathbf{w}_G)+\text{const} \end{aligned}$$

is quadratic with derivate

$$\frac{d\log p(x^*|\mathbf{t}^*,\mathbf{W},\mathbf{C})}{dx^*}=\mathbf{w}_G^T\mathbf{C}^{-1}(\mathbf{t}^*-\mathbf{m}-x^*\mathbf{w}_G) \quad (18)$$

and curvature

$$\frac{d^2\log p(x^*|\mathbf{t}^*,\mathbf{W},\mathbf{C})}{dx^{*2}}=-\mathbf{w}_G^T\mathbf{C}^{-1}\mathbf{w}_G.$$

Therefore, the posterior is Gaussian, with variance given by (5). The mean is obtained by setting (18) to zero, which yields (6).

APPENDIX II
ESTIMATE OF W

For training, the log marginal likelihood is given by

$$\begin{aligned} & \log p\left(\{\mathbf{t}_n\}_{n=1}^N|\{x_n\}_{n=1}^N,\mathbf{W},\mathbf{C}\right) \\ = & \sum_{n=1}^N -\frac{1}{2}(\mathbf{t}_n-\mathbf{W}\boldsymbol{\phi}_n)^T\mathbf{C}^{-1}(\mathbf{t}_n-\mathbf{W}\boldsymbol{\phi}_n)+\text{const}, \end{aligned}$$

which has as gradient with respect to $\mathbf{W}$

$$\sum_{n=1}^N \mathbf{C}^{-1}(\mathbf{t}_n-\mathbf{W}\boldsymbol{\phi}_n)\boldsymbol{\phi}_n^T.$$

Setting to zero and re-arranging yields (8).

APPENDIX III
EFFICIENT IMPLEMENTATION

Using Woodbury's identity, we obtain

$$\begin{aligned} \mathbf{C}^{-1} &= \boldsymbol{\Delta}^{-1}-\boldsymbol{\Delta}^{-1}\mathbf{V}\left(\mathbb{I}_K+\mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V}\right)^{-1}\mathbf{V}^T\boldsymbol{\Delta}^{-1} \\ &= \boldsymbol{\Delta}^{-1}-\boldsymbol{\Delta}^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}, \end{aligned} \quad (19)$$

and therefore (3) can be computed as

$$\mathbf{w}_D=\boldsymbol{\Delta}^{-1}\mathbf{w}_G-\boldsymbol{\Delta}^{-1}\mathbf{V}\boldsymbol{\Sigma}\left(\mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{w}_G\right).$$

Using this result, (5) is given by $\sigma_x^2=1/(\mathbf{w}_D^T\mathbf{w}_G)$.

Both computing the marginal likelihood (7), needed to monitor convergence of the EM algorithm for model training, and the inversion equation (17) for a forward model with nonlinearities involve numerical evaluations of the form

$$\log\mathcal{N}\left(\tilde{\mathbf{t}}|\,\mathbf{0},\mathbf{C}\right)\propto\tilde{\mathbf{t}}^T\mathbf{C}^{-1}\tilde{\mathbf{t}}+\log|\mathbf{C}|+\text{const}. \quad (20)$$

Using (19), the first term can be computed as

$$\begin{aligned} \tilde{\mathbf{t}}^T\mathbf{C}^{-1}\tilde{\mathbf{t}} &= \tilde{\mathbf{t}}^T\boldsymbol{\Delta}^{-1}(\tilde{\mathbf{t}}-\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\tilde{\mathbf{t}}) \\ &= \tilde{\mathbf{t}}^T\boldsymbol{\Delta}^{-1}(\tilde{\mathbf{t}}-\mathbf{V}\boldsymbol{\mu}), \end{aligned} \quad (21)$$

with $\boldsymbol{\mu}=\boldsymbol{\Sigma}\mathbf{V}^T\boldsymbol{\Delta}^{-1}\tilde{\mathbf{t}}$ being an estimate of the latent variables. The second term can be computed using Sylvester's determinant identity [77]:

$$\begin{aligned} |\mathbf{V}\mathbf{V}^T\boldsymbol{\Delta}^{-1}+\mathbb{I}_J| &= |\mathbf{V}^T\boldsymbol{\Delta}^{-1}\mathbf{V}+\mathbb{I}_K| \\ &= |\boldsymbol{\Sigma}|^{-1}, \end{aligned}$$

so that

$$\log|\mathbf{C}|=\log|\boldsymbol{\Delta}|-\log|\boldsymbol{\Sigma}|. \quad (22)$$

Finally, the EM update (12) of the diagonal matrix $\boldsymbol{\Delta}$ can be computed one element at a time: the j^{th} diagonal element is given by

$$\boldsymbol{\Delta}_{jj}=\frac{\sum_{n=1}^N\left(\tilde{t}_n^{\,j}-\mathbf{v}_j^T\boldsymbol{\mu}_n\right)\tilde{t}_n^{\,j}}{N},$$

where $\tilde{t}_n^{\,j}$ and $\mathbf{v}_j^T$ denote the j^{th} element (row) of $\tilde{\mathbf{t}}_n$ and $\mathbf{V}$, respectively.

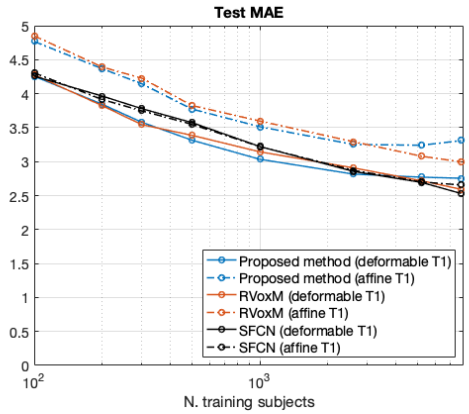


Fig. 16: Comparison of the proposed method, RVoxM and SFCN on the age prediction task. For each method, results are shown for both affine and deformable T1 input data.

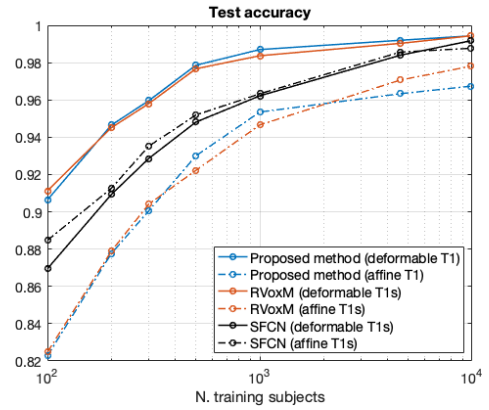


Fig. 17: Comparison of the proposed method, RVoxM and SFCN on the gender classification task. For each method, results are shown for both affine and deformable T1 input data.

APPENDIX IV
COMPUTATION OF HIGH-DIMENSIONAL EIGENVECTORS

In order to compute the eigenvectors of $\mathbf{V}\mathbf{V}^T$ in section III-A, we use the SVD decomposition of $\mathbf{V}$, given by $\mathbf{V} = \mathbf{U}\mathbf{S}\mathbf{R}^T$, with $\mathbf{U}$ $J \times J$ orthogonal matrix, $\mathbf{S}$ $J \times K$ diagonal matrix, and $\mathbf{R}$ $K \times K$ orthogonal matrix. This yields:

$$\mathbf{V}\mathbf{V}^T = \mathbf{U}\mathbf{S}(\mathbf{R}^T\mathbf{R})\mathbf{S}^T\mathbf{U}^T = \mathbf{U}(\mathbf{S}\mathbf{S}^T)\mathbf{U}^T, \qquad (23)$$

where $\mathbf{U}$ contains eigenvectors of $\mathbf{V}\mathbf{V}^T$ and $\mathbf{S}\mathbf{S}^T$ is a diagonal matrix with the eigenvalues.

APPENDIX V
AFFINE VS. DEFORMABLE T1S

Fig. 16 and Fig. 17 show prediction performances obtained by our method, RVoxM and SFCN for age and gender prediction respectively, on both affine and deformable T1s (with our own implementation). We obtain that performances of SFCN based on affine and deformable T1s are very similar, consistently with the finding reported in [5]. Instead, for our method and RVoxM, performances are clearly worse when using affine T1s as compared to deformable T1s, which is expected since they are linear methods, and therefore they do not have the capability of modeling nonlinear deformations in the data that have not been removed by the the affine registration.

APPENDIX VI
CODES

Regarding the SFCN for age prediction, we adapted the implementation[8] provided by [63] to more closely resemble the architecture and setting described in [5]. In particular we implemented the following changes:

- We added a convolutional block
- We used the same L2 weight decay coefficient reported in [5] (0.001)
- We used the MAE as metric that is evaluated on the validation set
- We changed the augmentation tecnique
- We added a pre-processing step (division by the images' means)

There are two aspects of the implementation provided by [63] that differ from the setting described in [5] which we did not change:

- the learning rate, because we empirically found it was better the one reported in [5]
- age is predicted as a continuous variable, with mean squared error as loss, instead of using a discretization in 40 classes and the KL-divergence as loss. [We kept this setting since it was reported to give the same results as with age binning.]

We then adapted the code to perform gender prediction, by using a sigmoid as activation function in the final layer of the network, and cross-entropy as loss function.

Regarding the RVoxM, we adapted the publicly available implementation[9] with the following changes:

- For age prediction, we parallelized part of the training loop to obtain a more efficient implementation
- For gender prediction, the provided code was not robust towards high dimensional input data, and we modified it to handle whole brain scans as input.

Finally, the code for the proposed model will be released at [address].

[8]https://github.com/pmouches/Multi-modal-biological-brain-age-prediction/blob/main/sfcn_model.py

[9]https://sabuncu.engineering.cornell.edu/software-projects/relevance-voxel-machine-rvoxm-code-release/

ACKNOWLEDGMENT

This research was conducted using the UK Biobank Resource under Application Number 65657, and it was made possible in part by the computational hardware generously provided by the Massachusetts Life Sciences Center (https://www.masslifesciences.com/).

REFERENCES

[1] James H Cole et al. Quantification of the biological age of the brain using neuroimaging. In *Biomarkers of human aging*, pages 293–328. Springer, 2019.

[2] Tobias Kaufmann et al. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623, 2019.

[3] Tong He et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276, 2020.

[4] Marc-Andre Schulz et al. Deep learning for brains?: Different linear and nonlinear scaling in uk biobank brain images vs. machine-learning datasets. *BioRxiv*, page 757054, 2019.

[5] Han Peng et al. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871, 2021.

[6] Fidel Alfaro-Almagro et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

[7] German National Cohort (GNC) Consortium geschaeftsstelle@ nationale-kohorte. de. The german national cohort: aims, study design and organization. *European journal of epidemiology*, 29(5):371–382, 2014.

[8] MM Breteler, T Stöcker, E Pracht, D Brenner, and R Stirnberg. Mri in the rhineland study: a novel protocol for population neuroimaging. alzheimer's dement. 10, p92, 2014.

[9] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald Henry, and Coen DA Stehouwer. The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451, 2014.

[10] Mohammad R Arbabshirani et al. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145:137–165, 2017.

[11] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016.

[12] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[13] Clifford R Jack Jr et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[14] Adriana Di Martino et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

[15] Kathryn A Ellis et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009.

[16] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1(1):1–13, 2014.

[17] Matthew F Glasser et al. The human connectome project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–1187, 2016.

[18] Terry L Jernigan, Timothy T Brown, Donald J Hagler Jr, Natacha Akshoomoff, Hauke Bartsch, Erik Newman, Wesley K Thompson, Cinnamon S Bloss, Sarah S Murray, Nicholas Schork, et al. The pediatric imaging, neurocognition, and genetics (ping) data repository. *Neuroimage*, 124:1149–1154, 2016.

[19] Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, James Loughead, Karthik Prabhakaran, Monica E Calkins, Ryan Hopson, Chad Jackson, Jack Keefe, Marisa Riley, et al. Neuroimaging of the philadelphia neurodevelopmental cohort. *Neuroimage*, 86:544–553, 2014.

[20] Norbert Hosten, Robin Bülow, Henry Völzke, Martin Domin, Carsten Oliver Schmidt, Alexander Teumer, Till Ittermann, Matthias Nauck, Stephan Felix, Marcus Dörr, et al. Ship-mr and radiology: 12 years of whole-body magnetic resonance imaging in a single center. In *Healthcare*, volume 10, page 33. MDPI, 2021.

[21] Stefan Haufe et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[22] Nishanth Arun et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[23] Marzyeh Ghassemi et al. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[24] Julius Adebayo et al. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[25] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[26] Rick Wilming, Céline Budding, Klaus-Robert Müller, and Stefan Haufe. Scrutinizing xai using linear ground-truth data with suppressor variables. *Machine learning*, pages 1–21, 2022.

[27] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[28] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR, 2020.

[29] Jindong Gu and Volker Tresp. Saliency methods for explaining adversarial attacks. *arXiv preprint arXiv:1908.08413*, 2019.

[30] Gabrielle Ras et al. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, 2022.

[31] Karen Simonyan et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*, 2014.

[32] David Baehrens et al. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[33] Dumitru Erhan et al. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[34] Avanti Shrikumar et al. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[35] Mukund Sundararajan et al. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[36] Jost Tobias Springenberg et al. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[37] Ramprasaath R Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[38] Daniel Smilkov et al. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[39] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[40] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[42] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[44] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*, 2017.

[45] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

[48] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The feature importance ranking measure. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 694–709. Springer, 2009.

[49] Chiara Mauri, Stefano Cerri, Oula Puonti, Mark Mühlau, and Koen Van Leemput. Accurate and explainable image-based prediction using a lightweight generative model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 448–458. Springer, 2022.

[50] J. Ashburner et al. Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821, 2000.

[51] C. Davatzikos et al. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage*, 14(6):1361–1369, 2001.

[52] MK Chung et al. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3):595–606, 2001.

[53] B. Fischl et al. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 97(20):11050, 2000.

[54] Nematollah K Batmanghelich et al. Generative-discriminative basis learning for medical imaging. *IEEE transactions on medical imaging*, 31(1):51–69, 2011.

[55] Erdem Varol et al. Generative discriminative models for multivariate inference and statistical mapping in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer, 2018.

[56] Pedro Domingos et al. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130, 1997.

[57] Andrew Y Ng et al. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[58] M. R. Sabuncu et al. The Relevance Voxel Machine (RVoxM): A Self-Tuning Bayesian Model for Informative Image-based Prediction. *IEEE transactions on medical imaging*, 31(12):2290–2306, 2012.

[59] Qingyu Zhao et al. Variational autoencoder for regression: Application to brain aging analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 823–831. Springer, 2019.

[60] Peter E Hart, David G Stork, and Richard O Duda. *Pattern classification*. Wiley Hoboken, 2000.

[61] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4, chapter 12. Springer, 2006.

[62] Donald B Rubin et al. Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[63] Pauline Mouches, Matthias Wilms, Deepthi Rajashekar, Sönke Langner, and Nils D Forkert. Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions. *Human brain mapping*, 43(8):2554–2566, 2022.

[64] Diederik P Kingma et al. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[65] Anders M. Fjell et al. High Consistency of Regional Cortical Thinning in Aging across Multiple Samples. *Cerebral Cortex*, 19(9):2001–2012, 2009.

[66] Anders M Fjell and Kristine B Walhovd. Structural brain changes in aging: courses, causes and cognitive consequences. *Reviews in the Neurosciences*, 21(3):187–222, 2010.

[67] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[68] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[69] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[70] Peter E Hart, David G Stork, and Richard O Duda. *Pattern classification*, chapter 9. Wiley Hoboken, 2000.

[71] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4, chapter 3. Springer, 2006.

[72] Karim Armanious, Sherif Abdulatif, Wenbin Shi, Shashank Salian, Thomas Küstner, Daniel Weiskopf, Tobias Hepp, Sergios Gatidis, and Bin Yang. Age-net: An mri-based iterative framework for brain biological age estimation. *IEEE Transactions on Medical Imaging*, 40(7):1778–1791, 2021.

[73] Kristine B Walhovd, Anders M Fjell, Ivar Reinvang, Arvid Lundervold, Anders M Dale, Dag E Eilertsen, Brian T Quinn, David Salat, Nikos Makris, and Bruce Fischl. Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of aging*, 26(9):1261–1270, 2005.

[74] Anders M Fjell, Lars T Westlye, Håkon Grydeland, Inge Amlien, Thomas Espeseth, Ivar Reinvang, Naftali Raz, Dominic Holland, Anders M Dale, Kristine B Walhovd, et al. Critical ages in the life course of the adult brain: nonlinear subcortical aging. *Neurobiology of aging*, 34(10):2239–2247, 2013.

[75] Matthias Wilms, Jordan J Bannister, Pauline Mouches, M Ethan MacDonald, Deepthi Rajashekar, Sönke Langner, and Nils D Forkert. Bidirectional modeling and analysis of brain aging with normalizing flows. In *Machine learning in clinical neuroimaging and radiogenomics in neuro-oncology*, pages 23–33. Springer, 2020.

[76] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

[77] A.G. Akritas, E.K. Akritas, and G.I. Malaschonok. Various proofs of Sylvester's (determinant) identity. *Mathematics and Computers in Simulation*, 42:585–593, 1996.

Bibliography

Adebayo, J. et al. (2018). Sanity checks for saliency maps. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Alfaro-Almagro, F. et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424.

Arbabshirani, M. R. et al. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*, 145:137–165.

Armanious, K., Abdulatif, S., Shi, W., Salian, S., Küstner, T., Weiskopf, D., Hepp, T., Gatidis, S., and Yang, B. (2021). Age-net: An mri-based iterative framework for brain biological age estimation. *IEEE Transactions on Medical Imaging*, 40(7):1778–1791.

Arun, N. et al. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267.

Ashburner, J. et al. (2000). Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baehrens, D. et al. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831.

Batmanghelich, N. K. et al. (2011). Generative-discriminative basis learning for medical imaging. *IEEE transactions on medical imaging*, 31(1):51–69.

Bishop, C. M. and Nasrabadi, N. M. (2006a). *Pattern recognition and machine learning*, volume 4, chapter 12. Springer.

Bishop, C. M. and Nasrabadi, N. M. (2006b). *Pattern recognition and machine learning*, volume 4, chapter 3. Springer.

Breteler, M., Stöcker, T., Pracht, E., Brenner, D., and Stirnberg, R. (2014). Mri in the rhineland study: a novel protocol for population neuroimaging. alzheimer's dement. 10, p92.

Chung, M. et al. (2001). A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3):595–606.

Cole, J. H. et al. (2019). Quantification of the biological age of the brain using neuroimaging. In *Biomarkers of human aging*, pages 293–328. Springer.

Davatzikos, C. et al. (2001). Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage*, 14(6):1361–1369.

Di Martino, A. et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Domingos, P. et al. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130.

Ellis, K. A. et al. (2009). The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687.

Erhan, D. et al. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

Fischl, B. et al. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 97(20):11050.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Fjell, A. M. et al. (2009). High Consistency of Regional Cortical Thinning in Aging across Multiple Samples. *Cerebral Cortex*, 19(9):2001–2012.

Fjell, A. M. and Walhovd, K. B. (2010). Structural brain changes in aging: courses, causes and cognitive consequences. *Reviews in the Neurosciences*, 21(3):187–222.

Fjell, A. M., Westlye, L. T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., Holland, D., Dale, A. M., Walhovd, K. B., et al. (2013). Critical ages in the life course of the adult brain: nonlinear subcortical aging. *Neurobiology of aging*, 34(10):2239–2247.

Friston, K. J., Frith, C., Liddle, P., and Frackowiak, R. (1991). Comparing functional (pet) images: the assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism*, 11(4):690–699.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210.

German National Cohort Consortium (2014). The german national cohort: aims, study design and organization. *European journal of epidemiology*, 29(5):371–382.

Ghassemi, M. et al. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.

Glasser, M. F. et al. (2016). The human connectome project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–1187.

Gu, J. and Tresp, V. (2019). Saliency methods for explaining adversarial attacks. *arXiv preprint arXiv:1908.08413*.

Hart, P. E., Stork, D. G., and Duda, R. O. (2000). *Pattern classification*, chapter 9. Wiley Hoboken.

Haufe, S. et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110.

He, T. et al. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276.

Hosten, N., Bülow, R., Völzke, H., Domin, M., Schmidt, C. O., Teumer, A., Ittermann, T., Nauck, M., Felix, S., Dörr, M., et al. (2021). Ship-mr and radiology: 12 years of whole-body magnetic resonance imaging in a single center. In *Healthcare*, volume 10, page 33. MDPI.

Jack Jr, C. R. et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691.

Jernigan, T. L., Brown, T. T., Hagler Jr, D. J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., et al. (2016). The pediatric imaging, neurocognition, and genetics (ping) data repository. *Neuroimage*, 124:1149–1154.

Kaufmann, T. et al. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature neuroscience*, 22(10):1617–1623.

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., and Dähne, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., et al. (2016). Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222.

Ng, A. Y. et al. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peng, H. et al. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871.

Ras, G. et al. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rubin, D. B. et al. (1982). Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sabuncu, M. R. et al. (2012). The Relevance Voxel Machine (RVoxM): A Self-Tuning Bayesian Model for Informative Image-based Prediction. *IEEE transactions on medical imaging*, 31(12):2290–2306.

Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., et al. (2014). Neuroimaging of the philadelphia neurodevelopmental cohort. *Neuroimage*, 86:544–553.

Schram, M. T., Sep, S. J., van der Kallen, C. J., Dagnelie, P. C., Koster, A., Schaper, N., Henry, R., and Stehouwer, C. D. (2014). The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451.

Schulz, M.-A. et al. (2019). Deep learning for brains?: Different linear and nonlinear scaling in uk biobank brain images vs. machine-learning datasets. *BioRxiv*, page 757054.

Selvaraju, R. R. et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shrikumar, A. et al. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Simonyan, K. et al. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sixt, L., Granz, M., and Landgraf, T. (2020). When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR.

Smilkov, D. et al. (2017a). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017b). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Snook, L., Plewes, C., and Beaulieu, C. (2007). Voxel based versus region of interest analysis in diffusion tensor imaging of neurodevelopment. *Neuroimage*, 34(1):243–252.

Springenberg, J. T. et al. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779.

Sundararajan, M. et al. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Varol, E. et al. (2018). Generative discriminative models for multivariate inference and statistical mapping in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–548. Springer.

Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., Quinn, B. T., Salat, D., Makris, N., and Fischl, B. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of aging*, 26(9):1261–1270.

Wilming, R., Budding, C., Müller, K.-R., and Haufe, S. (2022). Scrutinizing xai using linear ground-truth data with suppressor variables. *Machine learning*, pages 1–21.

Wilms, M., Bannister, J. J., Mouches, P., MacDonald, M. E., Rajashekar, D., Langner, S., and Forkert, N. D. (2020). Bidirectional modeling and analysis of brain aging with normalizing flows. In *Machine learning in clinical neuroimaging and radiogenomics in neuro-oncology*, pages 23–33. Springer.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–918.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fmri time-series revisited—again. *Neuroimage*, 2(3):173–181.

Wright, I., McGuire, P., Poline, J.-B., Travere, J., Murray, R., Frith, C., Frackowiak, R., and Friston, K. (1995). A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *Neuroimage*, 2(4):244–252.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhao, Q. et al. (2019). Variational autoencoder for regression: Application to brain aging analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 823–831. Springer.

Zien, A., Krämer, N., Sonnenburg, S., and Rätsch, G. (2009). The feature importance ranking measure. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 694–709. Springer.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., Breitner, J., Buckner, R. L., Calhoun, V. D., Castellanos, F. X., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1(1):1–13.