Bridging Simulation-based Inference and Hierarchical Modeling : Applications in Neuroscience

Louis Rouillard

université PARIS-SACLAY

Bridging Simulation-Based Inference And Hierarchical Modeling : Applications In Neuroscience

Combiner Inférence Basée sur Simulation et Modélisation Hiérarchique : Applications en Neurosciences

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Sciences du traitement du signal et des images
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Inria Saclay-Île-de-France (Université Paris-Saclay, Inria)**, sous la direction de **Demian WASSERMANN**, directeur de recherche

Thèse soutenue à Paris-Saclay, le 03 mai 2024, par

Louis ROUILLARD

Composition du jury

Membres du jury avec voix délibérative

Sylvain LE CORFF Professeur, Université Sorbonne	Président & Rapporteur
Gilles LOUPPE Professeur, Université de Liège	Rapporteur & Examinateur
Ruby KONG Chargée de recherche, National University of Singapore	Examinatrice
Pedro L. C. RODRIGUES Chargé de recherche, Inria Grenoble	Examinateur

THESE DE DOCTORAT

NNT : 2024UPASG024

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Combiner Inférence Basée sur Simulation et Modélisation Hiérarchique : Applications en Neurosciences

Mots clés : Inférence Variationelle, Neuroscience, Apprentissage Machine, Modélisation Bayesienne Hiérarchique, Inférence Basée sur Simulation

Résumé : La neuroimagerie étudie l'architecture et le fonctionnement du cerveau à l'aide de la résonance magnétique (IRM). Pour comprendre le signal complexe observé, les neuroscientifiques émettent des hypothèses sous la forme de modèles explicatifs, régis par des paramètres interprétables. Cette thèse étudie l'inférence statistique : deviner quels paramètres auraient pu produire le signal à travers le modèle.

L'inférence en neuroimagerie est complexifiée par au moins trois obstacles : une grande dimensionnalité, une grande incertitude et la structure hiérarchique des données. Pour s'attaquer à ce régime, nous utlisons l'inférence variationnelle (VI), une méthode basée sur l'optimisation.

Plus précisément, nous combinons l'inférence variationnelle stochastique structurée et les flux de normalisation (NF) pour concevoir des familles variationnelles expressives et adaptées à la large dimensionnalité. Nous appliquons ces techniques à l'IRM de diffusion et l'IRM fonctionnelle, sur des tâches telles que la parcellation individuelle, l'inférence de la microstructure et l'estimation du couplage directionnel. Via ces applications, nous soulignons l'interaction entre les divergences de Kullback-Leibler (KL) *forward* et *reverse* comme outils complémentaires pour l'inférence. Nous démontrons également les capacité de l'inférence variationelle automatique (AVI) comme méthode d'inférence robuste et adaptée à la large dimensionnalité, apte à relever les défis de la modélisation en neuroscience.

Title : Bridging Simulation-based Inference and Hierarchical Modeling : Applications in Neuroscience

Keywords : Variational Inference, Neuroscience, Machine Learning, Hierarchical Bayesian Modeling, Simulation-Based Inference

Abstract : Neuroimaging investigates the brain's architecture and function using magnetic resonance (MRI). To make sense of the complex observed signal, Neuroscientists posit explanatory models, governed by interpretable parameters. This thesis tackles statistical inference : guessing which parameters could have yielded the signal through the model.

Inference in Neuroimaging is complexified by at least three hurdles : a large dimensionality, a large uncertainty, and the hierarchcial structure of data. We look into variational inference (VI) as an optimization-based method to tackle this regime.

Specifically, we combine structured stochastic VI and normalizing flows (NFs) to design expressive yet scalable variational families. We apply those techniques in diffusion and functional MRI, on tasks including individual parcellation, microstructure inference and directional coupling estimation. Through these applications, we underline the interplay between the forward and reverse Kullback-Leibler (KL) divergences as complementary tools for inference. We also demonstrate the ability of automatic VI (AVI) as a reliable and scalable inference method to tackle the challenges of model-driven Neuroscience.

Synthèse détaillée en français

Cette thèse s'intéresse à la cartographie cérébrale à partir de données d'imagerie par résonance magnétique (IRM). En partant des images IRM du cerveau, on cherche à deviner quelle est la composition cellulaire de ses tissus, ou bien quelles régions du cortex sont connectées les unes aux autres pour assurer ses fonctions cognitives. Pour ce faire, on s'appuie sur des modèles explicatifs simples, dont on essaie d'inférer les paramètres à partir du signal magnétique. Cependant, en expliquant simplement un signal complexe, on crée de l'incertitude: plusieurs jeux de paramètres pourraient expliquer le signal observé. Cette thèse s'appuie sur l'inférence statistique pour deviner l'ensemble des paramètres explicatifs possibles pour générer un signal donné, et leur probabilités respectives, sous la forme de distributions de paramètres. L'organisation de la thèse est la suivante:

1. La partie I replace la thèse dans son contexte général. On présente d'abord les applications cibles en Neuroscience, en particulier deux modalités d'IRM utilisées dans les travaux présentés. L'IRM fonctionnelle mesure les variations de l'oxygénation du sang, un proxy de l'activité neuronale. L'IRM de diffusion mesure la diffusion de l'eau dans les tissus, afin d'en inférer la composition cellulaire. On présente ensuite la modélisation hiérarchique Bayésienne, qui permet de construire les modèles paramétriques avec lesquels on va chercher à expliquer le signal IRM. Enfin, on présente les méthodes statistiques qui permettent d'inférer les paramètres de ces modèles explicatifs. En particulier, cette thèse étudie l'inférence variationnelle (VI), qui traite les problèmes inverses comme des problèmes d'optimisation.

2. La partie II présente des technologies d'inférence statistique modernes, que cette thèse cherche à mettre à profit. La première technologie est l'inférence basée sur simulation (SBI). La SBI emploie de très expressifs approximateurs de distribution: les flots normalisants. Ces flots, puissants en faible dimensionnalité, peuvent servir de substitut aux distributions implicitement définies par un simulateur. Mais les flots normalisants ne sont pas utilisables dans le contexte de large dimensionnalité de la neuroscience. La seconde technologie est l'inférence variationnelle structurée et/ou stochastique (SVI). Elle recouvre des techniques d'optimisation qui permettent de traiter de larges problèmes en exploitant leur structure répétée. L'objectif de cette thèse est de combiner les outils de la SBI avec ceux de la SVI pour traiter des problèmes inverse en Neuroimagerie, à la fois très larges, et dont la complexité nécessite une grande expressivité.

3. La partie III met en avant les contributions de la thèse. Y sont d'abord présentées des avancées méthodologiques permettant le réemploi de flots normalisants à travers la structure répétée de modèles hiérarchiques. Ces technologies sont ensuite appliquées à trois problèmes ouverts en neuroscience. D'abord, la cartographie fonctionnelle individualisée du cerveau, localisant des fonctions majeures comme la vision ou le contrôle moteur. Ensuite, la cartographie tissulaire du cerveau à partir d'IRM de diffusion. Enfin, l'étude de la connectivité fonctionnelle du cerveau, ou comment identifier des réseaux de régions interconnectées participant aux des fonctions cognitives.

4. La partie IV fait office d'ouverture. Elle souligne les perspectives de recherche suggérées par cette thèse et ses connexions avec la modélisation générative.

La conclusion de cette thèse met en avant ses contributions vers une neuroscience interprétable et computationnellement efficace, via la modélisation paramétrique.

I believe that scientific knowledge has fractal properties, that no matter how much we learn, whatever is left, however small it may seem, is just as infinitely complex as the whole was to start with.

— Isaac Asimov

Remerciements

Je voudrais tout d'abord remercier le jury, Sylvain et Gilles, Pedro et Ruby, Alexandre et Luca d'avoir accepté d'être présents pour me challenger avec bienveillance. Luca AMBROGIONI et Alexandre GRAMFORT ont en particulier accepté de faire partie du jury uniquement en tant qu'invités. Merci de m'avoir donné la chance de vous présenter avec fierté le résultat de 4 années de travail.

Ensuite, je voudrais remercier Demian. La relation entre un superviseur et son étudiant est toujours complexe. Beaucoup de personnes ont déjà eu un manager, mais la relation au superviseur est plus subtile. Il faut trouver une difficile balance entre chef, mentor et collaborateur. Une forme de compagnonnage intellectuel. J'ai vécu un vrai tournant dans notre relation sur la troisième année de thèse, avec la participation à cetaines de tes collaborations de longue date, que je sais précieuses. Je voulais te remercier pour la confiance accordée, et ta supervision qui m'a fait beaucoup gagner en maturité, tant scientifique qu'humaine. Grâce à toi j'ai pris conscience dans cette thèse de la force du compromis, d'à quel point le temps est précieux, et de la dure nécessité de converger, même si c'est pas parfait, au terme de périodes d'exploration. Et tu sais mieux que personne que ça a été une longue route tortueuse, donc merci, merci beaucoup.

Plus généralement je voudrais remercier Parietal, puis MIND, deux équipes dans lesquelles je me suis senti comme à la maison. Je mesure le temps passé à tout organiser, tout financer, pour que les jeunes puissent se concentrer sur la technique. Temps qui peut parfois être invisible depuis l'intérieur du cocon. Et je sais qu'une équipe qui combine une exigence intellectuelle et une vraie bienveillance est quelque chose de rare, de difficile à construire, et à maintenir. Donc merci aux PIs: Demian, Bertrand, Alexandre, Philippe et Thomas. Et aussi merci aux générations de doctorants que j'ai vu passer, pre et post-covid. Une ambiance aussi joyeuse, c'est une vraie chance.

Merci également à mes collaborateurs sans qui les travaux présentés ici n'auraient jamais vu le jour. Le Stanford Cognitive & Systems Neuroscience Laboratory dirigé par Pr Vinod Menon, et son équipe: Byeongwook Lee —avec qui j'ai eu la chance de discuter longuement chaque semaine !— Srikanth Ryali, Nicholas Branigan, et Percy Mistry. Merci aussi à Maeliss Jallais et Marco Palombo du Cardiff University Brain Research Imaging Centre. Marco et Demian nous ont laissé la chance avec Maeliss de mener de front notre propre projet de recherche —en tant que jeune chercheurs— et je suis fier du résultat auquel nous sommes arrivés!

Je voudrais aussi remercier ma famille et mes amis d'avoir été présents lors de ma soutenance. Quatre ans, c'est long, et plus qu'une thèse j'ai l'impression que c'est un chapitre de ma vie qui s'est achevé. Et je suis très heureux d'avoir pu fêter ça avec vous. C'est assez émouvant de rassembler des gens d'horizons et de circonstances variés, on mesure le chemin parcouru.

Enfin, et avant toute chose, je voudrais remercier celle sans qui rien de tout ça n'aurait été possible.

A Sophie.

Contents

Introduction

The human brain is arguably one of the most challenging substrates of modern science. A circuitry of neurons is connected to form macroscopic tissues. From the electrophysiological activity of those cells, emerge primary functions, memory and behavior (Stangor and Walinga, 2014). How to make sense of this immensely complex system? Should we start at the biological basis, and investigate tracks of axons connecting distributed brain regions (F. Zhang et al., 2022)? Or should we rather investigate processes closer to cognition, for instance: where do groups of neurons activate when performing certain mental tasks (Poldrack et al., 2011)? Whatever the lens used, unveiling the brain's inner workings requires observing it in vivo, processing and producing information. However, measuring the brain non-invasively is a challenge on its own.

Thus we must rely on complex measurements of a complex object.

In this thesis, we attempt to make sense of this complexity via modeling (Koller and Friedman, 2009). Yet, providing simple explanations for complicated phenomena creates uncertainty: possibly several model parameters could yield the observed signal. We look into statistical inference as a way to capture, quantify, and ultimately reduce this uncertainty (Bishop, 2006; Koller and Friedman, 2009; Gelman, Carlin, et al., 2004).

In the rest of this introduction, we present the general organization of this dissertation.

Part I Chapter 1 introduces two non-invasive Neuroimaging techniques: functional magnetic resonance imaging (fMRI) and diffusion magnetic resonance imaging (dMRI). fMRI is a proxy to the brain's neuronal activity (Poldrack et al., 2011), and dMRI to its cellular composition (Bihan and Iima, 2015). The two techniques do not measure directly the brain's microarchitecture or function, but the magnetic **signal** that results from it. To make sense of this complex signal, Neuroscientists oftentimes posit explanatory **models**, governed by a few latent **parameters** (Thomas Yeo et al., 2011; Kong, J. Li, et al., 2019; Novikov et al., 2019). The goal of this thesis is to guess which parameters could have yielded the observed signal through the model. This objective is called **statistical inference** (Bishop, 2006; Koller and Friedman, 2009; Gelman, Carlin, et al., 2004).

Chapter 2 dives deeper into the explanatory models' definition. We show how parameters and signals can be linked via nested distributions, possibly across multiple measurements and subjects. Those interpretable distributions are called **hierarchical Bayesian models (HBMs)** (Koller and Friedman, 2009). Chapter 3 describes

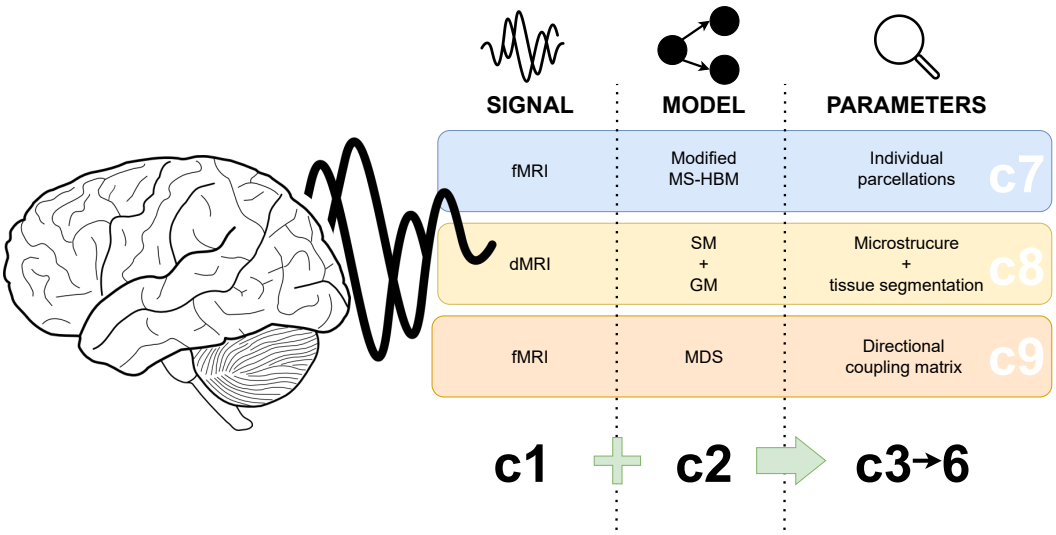

Fig. 1.: **Thesis organization** We organize our thesis around the tryptic signal/model/parameters. Neuroimaging signals are described in Chapter 1, models in Chapter 2, and inference methods to go from the signal to the model parameters in Chapters 3 to 6. This thesis presents three applications, each one an instance of the signal/model/parameters tryptic. Chapter 7 uses fMRI signal, a variation of the multi-session HBM (Kong, J. Li, et al., 2019), and recovers individual parcellations. Chapter 8 uses dMRI signal, a combination of the SM (Novikov et al., 2019) with a GM, and recovers microstructure parameters adjoined to a tissue segmentation (Jelescu, Palombo, et al., 2020). Chapter 9 uses fMRI signal, the MDS model (Ryali et al., 2011), and recovers directional coupling matrices.

in greater detail statistical inference to tackle those models. We introduce various techniques to go from an observed signal to a distribution of latent parameters. In particular, we present **variational inference (VI)**, an optimization-based method (David M. Blei et al., 2017). This thesis investigates the ability of VI to automatically tackle large-scale, hierarchical Neuroimaging problems.

Part II Chapters 4 and 5 present modern trends in inference. Chapter 4 introduces universal density approximators: **normalizing flows (NFs)** (Papamakarios, Nalisnick, et al., 2019). Using datasets of signal/parameter pairs, NFs can reliably approximate nearly any distribution, with immense potential in inference (Papamakarios and Murray, 2016; Cranmer et al., 2020). Yet NFs do not scale to the large dimensionality of our Neuroimaging problems. Chapter 5 introduces techniques in VI to tackle the large-scale. The key insight is to reflect the **model's structure** into parsimonious parameterizations and stochastic training schemes (Matthew D Hoffman and David M Blei, 2015; Ambrogioni, Lin, et al., 2021; Matthew D. Hoffman, David M. Blei, et al., 2013). This thesis' goal is to leverage the combined advantages of those modern techniques.

Part III Chapter 6 presents methodological contributions. By aggregating NFs into a causal structure, we show that we can increase VI's expressivity without compromising its scalability (Rouillard and Wassermann, 2022; Rouillard, Bris, et al., 2023). We then apply these techniques to a variety of Neuroscience problems. In Chapter 7, we map the individual brains of a thousand subjects, using fMRI scans (Kong, J. Li, et al., 2019). We segment the brain into macroscopic parcels, associating every point in the cortex with cognitive functions such as vision or motor control. In Chapter 4, we tackle microstructure estimation: inferring the cellular composition of the brain from a dMRI scan (Jallais and Palombo, 2023). To reduce the uncertainty in this estimation, we design composite models, combining learned density approximators with a hierarchical structure (Glöckler et al., 2022; Powell et al., 2021). In Chapter 9, we go back to fMRI to estimate directional coupling across the full brain (Ryali et al., 2011; Frässle, Lomakina, et al., 2017). In doing so, we unveil the individual links connecting distributed brain regions, building up the networks from which brain function emerges.

Finally, **Part IV** discusses open questions in inference, leading to this thesis' conclusion.

A graphical summary of the thesis organization is visible in Figure 1.

Across our applications, we investigate the **applicability of general VI methods to tackle large-scale, complex problems**. By removing the need for pen-and-paper derivations, while ensuring the reliability and scalability of inference, we push towards a simplification of the research cycle (Ambrogioni, Lin, et al., 2021; Ambrogioni, Silvestri, et al., 2021). *In the future, could model-driven Neuroscience be as simple as hypothesizing a HBM to explain the observed data, while automatic inference would yield the model's interpretable parameters?*

Part I

Background

1 Neuroscience and Neuroimaging

This chapter provides some background information on Neuroimaging. We start with a general motivation for non-invasive measurements of the human brain. We then introduce two imaging techniques used in our applications: fMRI and dMRI. We finish by motivating the methods developed in this thesis.

1.1 Non-invasive measurements: study the in-vivo brain

How to study the human brain?

Though earliest studies of the brain anatomy can be traced back to ancient Egypt —around 1600 BC, through the Edwin Smith papyrus— the theory of the functional specialization of the brain emerged in the 19th century. Franz Joseph Gall (1758-1828) proposed the association between the brain and the mind, further hypothesizing that specific mental faculties could be associated with specific brain regions. Famous autopsies supported this theory. For instance, Paul Broca (1824-1880) dissected the brain of an aphasic patient, Victor Leborgne, and discovered the area associated with speech production. Though highly informative in localizing function inside the brain, those "ablation studies" were obviously of limited reproducibility.

Another more modern example is histology. Staining microscopic slices of tissues to observe those under a microscope is a gold standard for characterizing brain microstructure. Yet, histology studies can only be performed post-mortem. Another limitation is that cell bodies can shrink through the processing of the tissue slices and do not necessarily reflect the microscopic structure of the in-vivo brain (Amunts et al., 2020; Howard et al., 2022).

It is thus essential to investigate the in-vivo brain in a non-invasive manner.

To study the brain non-invasively, one can measure the signal resulting from its electrical and metabolic activity, a field broadly described as Neuroimaging.

Electroencephalography (EEG), for instance, measures the brain's spontaneous electrical activity through postsynaptic neuron potentials. This thesis focuses on another imaging technique: magnetic resonance imaging (MRI). MRI uses strong magnetic fields to investigate brain anatomy, function and microstructure through diverse acquisition sequences. This thesis focuses on fMRI and dMRI in our applications, which we will introduce in the following two sections.

1.2 Functional MRI (fMRI): a proxy to neural activity

The first Neuroimaging modality we focus on is functional magnetic resonance imaging (fMRI) (Ogawa and T.-M. Lee, 1990; Poldrack et al., 2011). fMRI aims at measuring the activity of neurons inside the brain. When neurons activate —propagating action potentials— they require glucose. This leads to an increased blood flow towards regions of high neuronal activity. This increased inflow, in turn, leads to a higher concentration of oxygenated hemoglobin in the active regions. This locally changes the magnetic properties of the blood. This so-called blood-oxygen-level dependent (BOLD) contrast can be measured via MRI (Ogawa and T.-M. Lee, 1990).

We mention a few trends in fMRI research to put this thesis in perspective. First, fMRI is not a direct measure of neural activity but rather an indirect proxy. We measure the delayed variation in oxygen concentration resulting from neural activity. In mathematical terms, a widespread hypothesis models the observed fMRI time series as the convolution of some underlying neuronal activity by the hemodynamic response function (HRF) —visible in Figure 1.1 (Poldrack et al., 2011; Ryali et al., 2011). Moreover, this HRF can vary across different brain regions, meaning that the underlying neuronal activity is —to a certain extent— unknown (Devonshire et al., 2012; Handwerker et al., 2004; Taylor et al., 2018). In Chapter 9, we'll try to go beyond this uncertainty to investigate the actual neural activity in the brain.

Another trend in fMRI is the mapping of the human brain through parcellations. Many criteria can group different brain areas into "parcels" of distinct connectivity, microarchitecture, topography and function (Simon B. Eickhoff et al., 2018b). The resulting parcellations are of a dual interest. First, parcels constitute intermediate specialized units that can be integrated into larger-scale networks, leading to a hierarchical organization of brain function (Van Den Heuvel and Pol, 2010). Second, parcellations reduce the dimensionality of brain signals through meaningful components (Dadi et al., 2020). From a statistical standpoint, this allows for instance

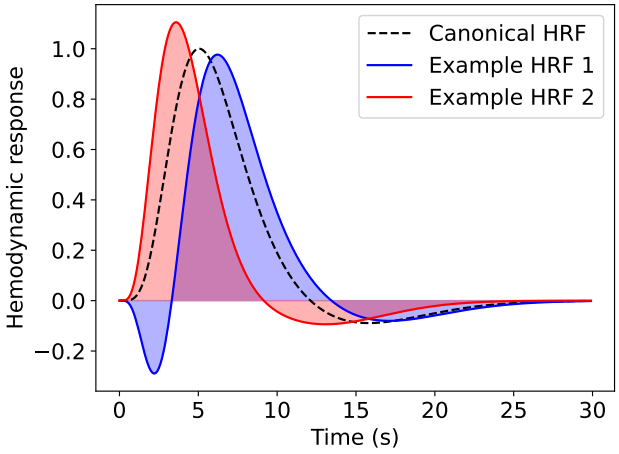

Fig. 1.1.: Hemodynamic Response Function (HRF) Glucose consumption at time 0 will lead to a complex response that spans over dozens of seconds. The concentration in oxygenated hemoglobin quickly rises to a peak before stabilizing to its original value. The canonical HRF is represented in dotted black. Two examples of HRF —corresponding to two different regions in the brain— are represented in color. Both HRFs differ from the canonical HRF in their "time-to-peak" or with the presence of an "initial dip" before the peak. Those differences can affect the measured precedence of the underlying activations. If the BOLD peak for one region happens before the other, is it because the region activated sooner or because its HRF has a smaller time-to-peak?

experimenters to detect contrasts in the brain without being hampered by conservative multiple comparison corrections. In this thesis, we focus on the parcellation emerging from functional connectivity (Van Den Heuvel and Pol, 2010).

A first point of note in connectivity-based parcellations is their variability across individuals. Even after mapping the different anatomies of two subjects, they could still differ in the localization of precise brain functions. Those differences bode the question of the existence of one "universal" brain map (Simon B. Eickhoff et al., 2018b). From a more practical standpoint, precisely mapping the brain of a *given* individual has medical applications. In the context of brain tumor removal, localizing precise brain functions from an fMRI scan could help select areas that would minimally impact a subject's recovery after a brain resection (Mandonnet, 2011). We endeavor to obtain such individual parcellations in Chapter 7.

Another point of note in brain connectivity lies in its very definition. A dominant standard in the fMRI community is using correlation to define brain connectivity. The underlying assumption is that if two areas have correlated signals, they should be connected at some level and partake in the same brain functions. Correlation has a massive advantage in its simplicity, both to compute and to interpret. Yet correlation lacks the finer granularity needed to unveil the individual role of parcels

as part of larger networks. If two parcels, A and B, have correlated signals, is it because A activates B? Or because B activates A? Or because yet another region, C, activates both A and B? Many methods attempt to refine the notion of connectivity into so-called "causal" or "effective" connections (Ryali et al., 2011; K. J. Friston et al., 2003). We study in greater depth this problem in Chapter 9.

1.3 Diffusion MRI (dMRI): a proxy to brain microstructure

The second Neuroimaging modality we focus on is dMRI (Bihan and Iima, 2015). dMRI is a key modality for biological tissue structure imaging. Applied to the brain, dMRI helps quantify neuron soma sizes or their axon orientation (see Figure 1.2). To do so, dMRI quantifies how water molecules diffuse in a tissue. By controlling the direction and strength of the magnetic field, dMRI can assess how easily water can diffuse in a given direction, at a given scale. When their movement is not impeded by any obstacle, the diffusion of water molecules is isotropic. In contrast, in tissues, this movement is blocked by cell membranes and other molecules crowding the environment —which is denoted by the local *diffusivity*. By quantifying the deviation from free diffusion, dMRI informs about the local (micro)structure of biological tissues. For instance, inside neurons, water can only diffuse parallel to the axons. One application of dMRI is thus tractography, which investigates the white matter bundles connecting different brain parts (F. Zhang et al., 2022). This thesis focuses on another application of dMRI: microstructure estimation, which investigates the *statistical* cellular composition of brain voxels (Alexander et al., 2019; Jelescu, Palombo, et al., 2020).

We mention a few trends in microstructure estimation research to put this thesis in perspective. Similar to fMRI, dMRI is only a —very— indirect proxy of its target of interest: microstructure. To guess which microstructure could have produced the observed water diffusion, neuroscientists first hypothesize biophysical models of tissue. Some models are realistic and necessitate to then simulate water diffusion in the synthetic tissue (for instance via Monte Carlo) (Callaghan et al., 2020; Ginsburger et al., 2019). Other models are simpler geometrical approximations that do not require to simulate water diffusion. Those biophysical models first approximate neurons through tubes, balls, and different compartments of different diffusivities (Hui Zhang et al., 2012; Novikov et al., 2019; Palombo et al., 2020). Second, they compute the magnetic signal resulting from the diffusion of water

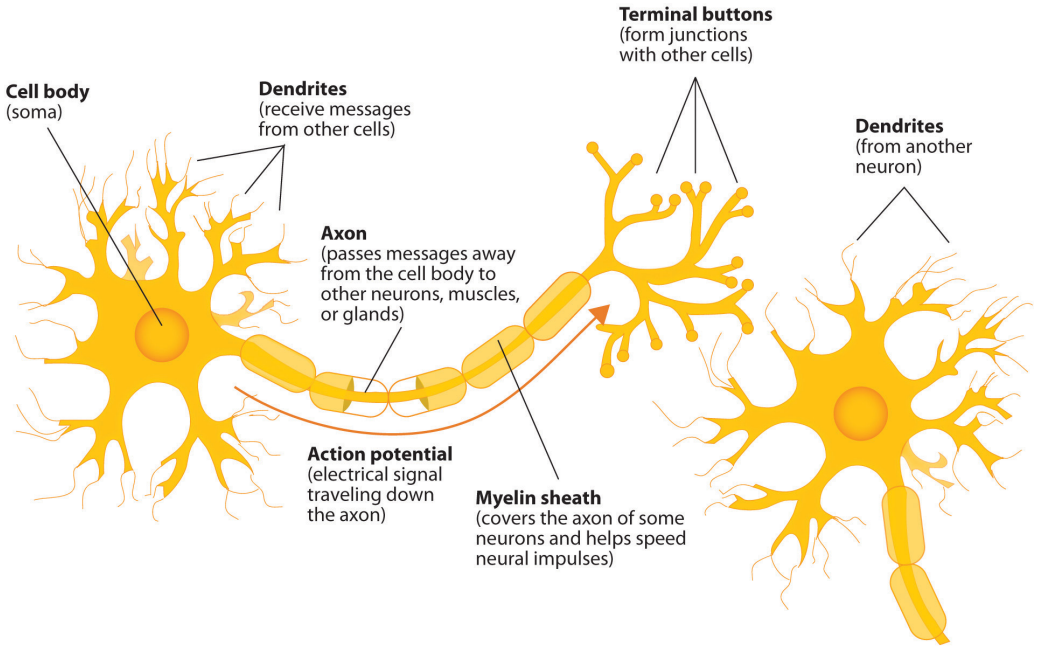

Fig. 1.2.: **Neurons** Schematic representation of the different compartments of a neuron. Some dMRI models simplify those compartments through simple geometric shapes (Hui Zhang et al., 2012; Novikov et al., 2019; Palombo et al., 2020). The soma can be approximated via a ball, the axon via a straight tube, etc... *Figure adapted from Stangor and Walinga (2014)*.

inside those simple geometries. This produces a "forward" model that must then be fitted on the observed signal to go in the "backward" direction: from the signal to the ball sizes and tube orientations.

At its core, dMRI has to deal with strong uncertainty. Through the "forward" model, different geometries can produce the same magnetic signal. If water diffusion is impeded in a given direction, is it due to many orthogonal cell membranes or because many metabolites reduce the diffusivity? The more complex the model —for instance, adding more compartments— the more difficult it becomes to identify its parameters. Uncertainty is further amplified by the low spatial resolution of dMRI. dMRI typically measures the magnetic signal coming from 1mm-large voxels (volumetric pixels). Such large volumes are ill-adapted to measuring microscopic physical phenomena, which can create degeneracies when inverting biophysical models (Jelescu, Palombo, et al., 2020). A first trend in dMRI research lies in developing models that are good approximations of the microstructure while being governed by parameters easily identifiable from the signal (Hui Zhang et al., 2012; Novikov et al., 2019; Palombo et al., 2020).

Yet, no matter the model's simplicity, parameters can never be perfectly identified. One reason for this is the noise that corrupts the observed signal. This is especially true in clinical setups with weaker MRI fields and less controlled experimental conditions. As a result, another trend in dMRI research does not output single parameter estimates, but distributions (Jallais, Rodrigues, et al., 2021; Jallais and Palombo, 2023; Powell et al., 2021). Distributions encapsulate multiple possible explanatory microstructures, along with their respective probabilities. This can provide experimenters insights about parameter estimates robustness and reliability, which is crucial for interpreting the results.

In Chapter 8, we endeavor to identify microstructure model parameters and reduce their estimation uncertainty.

1.4 Major hurdles in Neuroimaging: uncertainty and large scale

In this section, we mention a few hurdles in Neuroimaging that motivate this thesis. Our goal is not to make an exhaustive list but to provide some context.

Throughout Sections 1.2 and 1.3, we touched upon one major hurdle in Neuroimaging: uncertainty. The human brain's connectivity, microarchitecture and function

are immensely complex. To make sense of the observed MRI **signal**, Neuroscientists posit explanatory **models**. Those models are governed by a few interpretable **parameters**. This "forward" relationship —from the parameters to the signal— can be complex, random at its very core, and corrupted by noise. Thus, going in the "backward" direction —from the signal to the parameters— is even more complex. Many a method can output a possible set of parameters to explain the observed signal. But could other, equally explanatory sets of parameters also exist? **The first objective of this thesis is the development of methods that can capture this uncertainty.**

Another hurdle in Neuroimaging lies in the massive dimensionality it entails. The MRI signal is typically measured across hundreds of thousands of voxels —volumetric pixels. The voxel-wise signal itself can be high-dimensional. In fMRI, voxel times series can span over a thousand timesteps. In dMRI, hundreds of magnetic field directions and strengths are collected at each voxel. This means that a single brain measurement contains dozens of millions of values. **The second objective of this thesis is the development of methods that can tackle this massive dimensionality.**

A third focus of this thesis is intertwined with the two aforementioned hurdles: hierarchical modeling. To gain statistical power, and capture population trends, data in Neuroimaging is collected across dozens or even hundreds of subjects (Van Essen et al., 2012; Sudlow et al., 2015). By combining the information across subjects, we can reduce the uncertainty in a given subject's parameters. At the same time, those population studies further multiply the already massive dimensionality. Hierarchical modeling thus constitutes both a blessing and a curse. **The third objective of this thesis is to develop methods capable of hierarchical modeling.**

The tryptic parameters/model/signal lies at the core of this thesis. Each of our applications will be an instantiation of this tryptic. **How can we infer all the possible parameters susceptible to generating the observed, massive MRI signal through the hypothesized hierarchical model?**

2 Hierarchical modeling: combining the information

In Section 1.4, we described how neuroscientists posit explanatory **models** to investigate the human brain. This chapter opens the box of those models. We first introduce hierarchical Bayesian modeling as a way to link the observed **signal** to hypothesized latent **parameters**. Next, we review how this type of modeling is applied in medicine and Neuroimaging. This section illustrates basic principles of inference detailed more formally in Chapter 3

2.1 Hierarchical Bayesian Modeling (HBM)

2.1.1 Random Variables and distributions

How to represent unknown latent parameters?

One representation that innately incorporates the notion of uncertainty is **random variable (RV)s**. This thesis focuses on statistical inference over RVs through the **Bayesian** framework (Gelman, Carlin, et al., 2004; Bishop, 2006; Koller and Friedman, 2009). A more detailed treatment of the Bayesian theory (including Bayes theorem) will be given in Section 3.1.

As a running example, consider measuring the weight of salmons inside a river. Each salmon i has an unknown weight denoted θ_i^{salmon}. Let us assume we have little information about the distribution of salmon weights in the world. Our prior for each salmon's weight is thus modeled as an uninformative uniform distribution with a large support. We assume that the weight of each salmon i follows this distribution, which is denoted:

$$\begin{aligned} \forall i = 1..N \quad \theta_i^{\text{salmon}} &\sim \mathcal{U}(w_{\text{min}}, w_{\text{max}}) \\ w_{\text{max}} - w_{\text{min}} &\gg 1 \end{aligned} \tag{2.1}$$

such a **distribution** encapsulates a —potentially infinite— set of values for a RV, along with the probability associated with each value. See Figure 2.1 for a complex example of distribution.

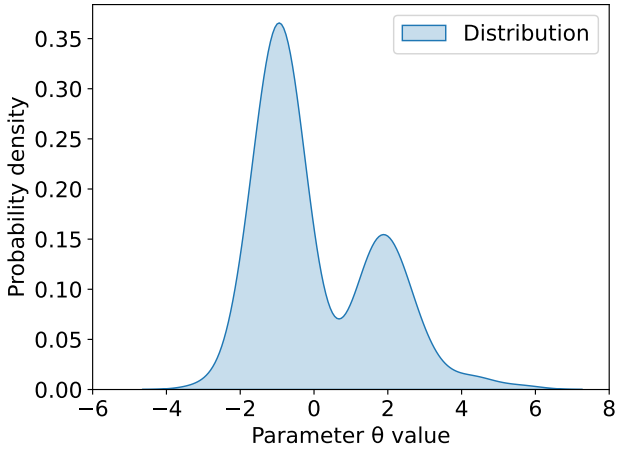

Fig. 2.1.: Example 1D distribution The distribution of one one-dimensional parameter θ. In this continuous case, θ can take any value between -4 and 7. The probability of falling into a given interval can be computed by integrating the probability density. This distribution is rather complex. It features two **modes**, centered on -1 and 2, corresponding to high probability regions. It also features some long **tail** spanning from 4 to 7, where the density does not vanish completely.

For scientific purposes, we wish to know the value of θ_i^{salmon}. But due to budget cuts, all we have access to is a rusty scale. This scale measures the weight of the salmon with a large known error σ^2_{scale}. We can denote this measured weight X_i^{salmon}. X_i^{salmon} follows the **conditional distribution**:

$$\forall i = 1..N \quad X_i^{\text{salmon}} | \theta_i^{\text{salmon}} \sim \mathcal{N}(\theta_i^{\text{salmon}}, \sigma^2_{\text{scale}}) \tag{2.2}$$

that is to say, the measured weight is equal to the true weight, plus some random white noise.

To use the terminology from Section 1.4, the true salmon weight is the **parameter** we wish to retrieve. The measured weight is the observed **signal**. Finally, the **model** is the joint distribution between the parameters and the signal:

$$\underbrace{P(X_i^{\text{salmon}}, \theta_i^{\text{salmon}})}_{\text{Joint}} = \overbrace{P(X_i^{\text{salmon}} | \theta_i^{\text{salmon}})}^{\text{Likelihood}} \times \underbrace{P(\theta_i^{\text{salmon}})}_{\text{Prior}} \tag{2.3}$$

where $P(\theta_i^{\text{salmon}})$ corresponds to Equation (2.1), and $P(X_i^{\text{salmon}} | \theta_i^{\text{salmon}})$ corresponds to Equation (2.2). The Bayesian terminology in Equation (2.3) will be detailed more thoroughly in Section 3.1. This model P can be represented using the directed acyclic graph (DAG) in Figure 2.2 (Koller and Friedman, 2009).

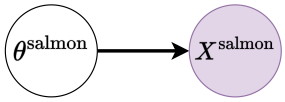

Fig. 2.2.: Salmon example graphical model Random Variables are represented using nodes. Conditional dependency is represented via directed edges. Observed RVs —the signal— are represented as grayed nodes. White nodes correspond to the inferred parameters.

Observing the noisy X_i^{salmon}, we cannot infer θ_i^{salmon} unequivocally. In this simple Gaussian case, we can mathematically derive the *distribution* of potential θ_i^{salmon} values, *given* the measured weight:

$$\begin{aligned} \forall i = 1..N \quad \theta_i^{\text{salmon}} | X_i^{\text{salmon}} &\sim \mathcal{N}(\mu_{\text{post,i}}, \sigma^2_{\text{post}}) \\ \mu_{\text{post,i}} &\simeq X_i^{\text{salmon}} \\ \sigma^2_{\text{post}} &\simeq \sigma^2_{\text{scale}} \end{aligned} \tag{2.4}$$

In Bayesian terms, the distribution described in Equation (2.4) is called the *posterior*, the distribution of the parameters given the signal.

The precision $\sigma^{-2}_{\text{post}}$ in this posterior is roughly equal to the rusty scale's precision, which is low. As a result, we have little information about each salmon's weight. In the next section, we show how hierarchical modeling can increase this precision.

2.1.2 Hierarchical models and plates

How to infer salmon weights with greater precision?

One way is to combine the information across salmons. Salmon specialists argue that salmon weights tend to be geographically homogeneous. Taking into account this information, we can define a new *hierarchical* prior to replace the prior described in Equation (2.1):

$$\begin{aligned} \theta^{\text{river}} &\sim \mathcal{U}(w_{\text{min}}, w_{\text{max}}) \\ \forall i = 1..N \quad \theta_i^{\text{salmon}} | \theta^{\text{river}} &\sim \mathcal{N}(\theta^{\text{river}}, \sigma^2_{\text{river}}) \\ \forall i = 1..N \quad X_i^{\text{salmon}} | \theta_i^{\text{salmon}} &\sim \mathcal{N}(\theta_i^{\text{salmon}}, \sigma^2_{\text{scale}}) \end{aligned} \tag{2.5}$$

In this new prior, salmons' weights are concentrated around a mean river weight. This implies that salmons' weights are not considered independent anymore. In particular, observing one salmon's weight is informative about another salmon's, since both weights are perturbations of the same river mean weight.

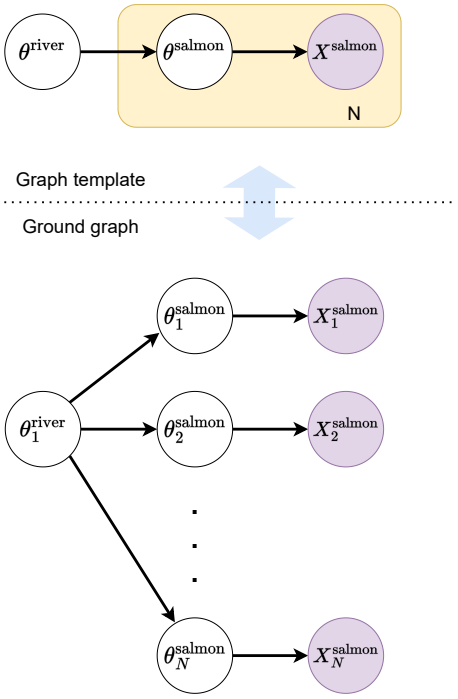

Fig. 2.3.: Salmon example hierarchical graphical model On top, we represent a graph template, with the corresponding ground graph at the bottom. The plate N symbolizes many repeated RVs instantiating the RV templates. In this case, as many salmon's weights θ_i^{salmon} as there are salmons in the river.

Switching to a hierarchical setup, we significantly increase the number of RVs in the model (as many RVs as salmons). Such large models can be compactly represented via plate-enriched directed acyclic graph (DAG) templates, as in Figure 2.3 (Koller and Friedman, 2009). The resulting graphical models feature RV *templates* that symbolize multiple similar *ground* RVs. As an example, the RV template θ^{salmon} in Figure 2.3 represents a generic salmon's weight, of which there are as many instances as the cardinality of the plate. The plate structure also denotes that the multiple ground RVs corresponding to the same RV template are conditionally independent and identically distributed (i.i.d) That is to say, all the salmon's weights are independent given the river's mean weight. In addition, all the salmon's weights follow the same Gaussian distribution described in Equation (2.5).

Because we have changed the prior for θ^{salmon}, the posterior is also affected:

$$\begin{aligned}
\forall i = 1..N \quad \theta_i^{\text{salmon}} | X^{\text{salmon}} &\sim \mathcal{N}(\mu_{\text{hier_post,i}}, \sigma^2_{\text{hier_post}}) \\
\mu_{\text{hier_post,i}} &\simeq \sigma^2_{\text{hier_post}}(\frac{\hat{\mu}_{\text{river}}}{\sigma^2_{\text{river}}} + \frac{X_i^{\text{salmon}}}{\sigma^2_{\text{scale}}}) \\
\hat{\mu}_{\text{river}} &= \frac{1}{N}\sum_{i=1}^{N} X_i^{\text{salmon}} \\
\sigma^2_{\text{hier_post}} &\simeq \frac{1}{\sigma^{-2}_{\text{river}} + \sigma^{-2}_{\text{scale}}}
\end{aligned} \tag{2.6}$$

We can compare this posterior to the one in Equation (2.4):

- We have added the river's mean precision to the scale's: $\sigma^{-2}_{\text{scale}} \leftarrow \sigma^{-2}_{\text{river}} + \sigma^{-2}_{\text{scale}}$. In doing so, we significantly increased the precision at which we can infer each salmon's weight;

- At the same time, we bias each salmon's weight towards the river's empyrical mean $\hat{\mu}_{\text{river}}$. This means that the inferred salmon weight no longer solely depends on the measured signal, but also on some prior hypothesis —that salmon's weights should be similar inside the river.

This example illustrates the **bias versus variance** trade-off at the core of the Bayesian prior design. We reduced the uncertainty in the posterior (the variance) at the cost of a now-biased result. In Section 3.1 we detail more this trade-off.

In the next section, we move on from our simple salmon example to medical and neuroscientific examples.

2.2 Hierarchical modeling in medicine and neuroscience

In this section, we review instances of hierarchical modeling in medicine and neuroimaging. We then delineate the type of modeling and inference we perform. This section thus positions our methods in the Neuroscience literature.

2.2.1 Hierarchical modeling in medicine and neuroimaging

Hierarchical modeling has a long history in medical research (Gelman and Hill, 2006; McGlothlin and Viele, 2018). The latter is typically interested in recovering treatment effects over subjects. Scientists perform group-level analyses to test differences between randomized populations. Applied over larger groups, so-called *population studies* gain statistical power to overcome spurious findings (Fayaz et al., 2016). General statistical frameworks to perform such analysis include analysis of variance (ANOVA), or multilevel regression models (Gelman and Hill, 2006). Those frameworks are similar to Bayesian hierarchical modeling, considering multiple subjects and measurements per subjects as *plates* —as described in Section 2.1.2.

Hierarchical modeling is also almost systematic in Neuroimaging research. In particular, a broad standard in fMRI is the use of generalized linear models (GLMs) to measure contrasts between experimental conditions (Poldrack et al., 2011). For instance: which regions in the brain are more active when seeing a human face versus a scrambled image? Those contrasts are collected at the subject level —via so-called *first-level* models. To test for statistical significance, the contrasts are then aggregated at the group level —via *second-level* models. Similar to medical research, such analysis can be performed at very large scales, leading to population studies (Towsley et al., 2011; Sudlow et al., 2015).

Interpreting data hierarchically is thus ubiquitous in our application field. This bodes the question: what are the specificities of the applications enabled by our methods? In Section 2.2.2, we provide some context under the lens of statistics. In Section 2.2.3, we focus on the modeling aspect.

2.2.2 Frequentist and Bayesian statistics: the right tool for each job

Taking a step back from the examples described in the section above, we can (roughly) formalize a standard frequentist hypothesis testing framework. First, *features* are collected at the subject level, *independently*. Second, in a posthoc analysis, features are *aggregated* using some statistical test, controlling *statistics* such as the false discovery rate (FDR), or the p-value.

In what ways does our Bayesian hierarchical estimation framework differ?

The first difference is at the *feature collection* level. Frequentist methods oftentimes resort to maximum likelihood (ML) estimators, which provide unbiased point estimates for the parameters. This means that the obtained parameters Θ are a

single value maximizing $P(X|\Theta)$. In contrast, we obtain full parameter posterior distributions —which are computationally much more difficult to achieve. One advantage of posterior distributions is that they incorporate a notion of statistical confidence *at the subject level*.

A second difference is at the *feature aggregation* level. Frequentist methods rely on features being collected independently to perform statistical tests. In contrast, hierarchical modeling breaks this independence. As in our salmon example in Section 2.1.2, subject-level features will be less noisy, but will typically be biased towards each other.

A third difference lies in the type of *statistics* used. Frequentist methods often focus on p-values as a measure of hypothesis acceptability. P-values can be harder to interpret for non-statistician audiences. For instance, lower p-values do not necessarily imply a larger effect size. Nor can p-values be interpreted directly as hypothesis probabilities. In contrast, we focus on group-level effect estimation, using actual posterior densities. Posterior distributions have been argued as easier to interpret for non-statistician audiences (Kruschke and Liddell, 2018).

Compared to frequentist hypothesis testing, this thesis thus focuses on a delineated class of problems:

> Parameter **estimation** applications, where significant **noise** is expected at the subject level. First, this noise can be **quantified and interpreted** via distributions. Second, this noise can be **reduced** by sharing the information across local features, at the cost of statistically entangling the latter.

As an opening, we sustain the growing opinion that frequentist and Bayesian viewpoints shouldn't be opposed, but rather combined (Bzdok and Yeo, 2017). In particular, the difference between the frequentist hypothesis testing and the Bayesian hierarchical estimation frameworks is not clear-cut.

For instance, on the *feature collection* side: when adding regularization over a frequentist optimization, one implicitly injects priors. The only difference with the Bayesian maximum a posteriori (MAP) viewpoint is that the latter is explicit and conceptualizes those priors as distributions. Conversely, using weakly informative priors brings Bayesian estimation closer to the ML framework.

Similarly, there is a range of estimates spanning between point estimates and full distributions, including Empyrical Bayes or MAP-II estimation (Murphy, 2012). On

the *statistics* side, frequentist methods are amenable to estimation with uncertainty — via ML estimates with confidence intervals (Kruschke and Liddell, 2018). Conversely, the Bayesian framework is amenable to hypothesis testing via Bayes factors and model comparison (Kruschke and Liddell, 2018). A whole continuum of methods thus combines to a diverse degree the frequentist and Bayesian viewpoints, and the appropriate method should be selected for each considered problem.

To conclude, as far as the Neuroscience community is concerned, we can cite the opinion paper from Bzdok and Yeo (2017):

> *As a general tendency, the more one adheres to frequentist instead of Bayesian ideology, the less computationally expensive and the less technically involved are the statistical analyses. It is a widespread opinion that Bayesian models do not scale well to the data-rich setting, although there is currently insufficient work on the behavior of Bayesian methods in high-dimensional input data. [...] In sum, the scalability of model estimation in the data-rich scenario is calibrated between frequentist numerical optimization and Bayesian numerical integration.*

Through this thesis, we strive to enable the transparency and richness of the Bayesian viewpoint in such "*data-rich settings*". By repurposing the "*frequentist optimization*" methods into a "*Bayesian integrative*" framework, we further bridge the gap between both viewpoints. The scalability of our methods will be detailed in greater length in Chapter 3.

2.2.3 Hierarchical Bayesian Models as generative, non-parametric, large-scale models

The previous section focused on the statistical aspect of our method: Bayesian estimation with uncertainty. In this section, we review modeling aspects: what are the types of models enabled by our methods?

To offer some context from the Neuroimaging community, we briefly review highlights from the Bzdok and Yeo (2017) opinion paper. Bzdok and Yeo (2017) explore the question:

> *How will the unprecedented data richness shape data analysis practices [inside the Neuroscience community]?*

They encourage a shift towards:

- *non-parametric* models, defined as adaptive models where "*the number of parameters increases explicitly or implicitly with the number of available data points*". Those are opposed to parametric models with a fixed number of parameters, prone to under-fitting in data-rich scenarios;

- *generative* models, that explicitly model the distribution of the signal $P(X)$ using some interpretable, hidden representation of the brain. Those are opposed to discriminative models, that only model the distribution of some target variable y (e.g. a cognitive score) conditional to the signal: $P(y|X)$.

- methods reconciling Bayesian interpretability and frequentist computational efficiency;

- out-of-sample generalization as statistics in data-rich, high-dimensional settings. Those are opposed to p-values and other in-sample estimates more adapted to small samples and parametric settings. On that point, the authors actually argue for the combination of both statistics, depending on the context.

In the next paragraphs, we analyze the fit of HBMs to this taxonomy.

Non-parametric models Under the most restrictive definition of non-parametric models, HBMs *are* parametric models. That is to say, HBMs *do* assume certain (conditional) distributions form for the signal and the latent parameters (e.g. Gaussians). Thus, HBMs are prone to model misspecification and to underfitting the observed signal. Nonetheless, by nesting parametric models, HBMs gain in expressivity and can eventually fit adequately the observed signal. As an example, consider Gaussian mixtures with non-parametric variance, which combine simple building blocks to represent multi-modal, heavy-tailed distributions (see Section 3.4 for a definition of those terms). In a similar vein, HBMs can jointly model and marginalize hyper-parameters such as the number of components in an independent component analysis (ICA), clusters in a clustering, or topics in an author-topic model (Goodfellow et al., 2016; Rosen-Zvi et al., 2010). Plate-enriched HBMs can also dynamically adapt their number of parameters with the available data, by distinguishing population from subject-level parameters. HBMs can thus adapt their internal representations dynamically as more and more data becomes available. As such **HBMs are —according to Bzdok and Yeo (2017)'s definition— non-parametric models**.

Generative models By design, HBMs link latent parameters with the observed signal via a joint distribution and aim at interpreting the observed data. HBMs are thus a good fit with the definition from Bzdok and Yeo (2017). HBMs can also be used as generative models able to generate synthetic signals similar to biological signals. Indeed, after inference, latent parameters posteriors are "fitted" to the observed data, and synthetic samples can be drawn from the posterior predictive distribution. **HBMs thus are interpretable generative models**.

As detailed in Section 2.2.2, our methods combine Bayesian modeling and frequentist optimization. Finally, in our applications, we use out-of-sample generalizations as a main validation strategy.

In conclusion, our approach follows closely the recommendations from Bzdok and Yeo (2017):

> We focus on models hierarchically **nesting parametric distributions** to augment their expressivity. Via plates, those models can dynamically **adapt their number of parameters** with the amount of available data. Through inference, those models **learn the distribution** of the observed signal, and link it to **interpretable** latent parameters.

As an opening, we acknowledge that different researchers than Bzdok and Yeo (2017) could have drastically different opinions over the future of Neuroscience methods. In particular, fully non-parametric, deep neural network (DNN)s approaches have shown impressive performance over the past decades (Eickenberg et al., 2017; Jang et al., 2017; Plis et al., 2014). Those approaches completely bypass the model specification needed for HBMs, approach big data simply via a massive parameterization, and often thwart the trade-offs of traditional statistics (Goodfellow et al., 2016). As a counterpoint, Efron and Hastie (2021) argue that DNNs are ill-suited for parameter estimation since they do not ensure the unicity of their trained weights, nor incorporate confidence intervals.

In contrast to purely non-parametric DNN approaches, **this thesis applies tools from the machine learning community to modernize traditional Bayesian modeling.**

3 Statistical inference: dealing with uncertainty

In Chapter 2, we described the **models** posited by Neuroscientists to explain the observed brain **signal**. This chapter abstracts back model specification and describes *inference*: how to guess which **parameters** generated the signal through the model. We focus on statistical inference as a framework to tackle the uncertainty hurdle described in Section 1.4. We start with a general definition of Bayesian inference. We then describe methods to perform approximate inference in practice. Next, we present VI, the statistical inference method we focus on. We finish by presenting some practical obstacles for inference, which serve as additional motivations for this thesis.

3.1 The Bayesian inference formalism

In Section 1.4, we introduced the tryptic parameters/model/signal. Guessing the model's parameters susceptible to generating the signal can be broadly described as *statistical inference*. As presented in Section 2.1.1, this thesis frames statistical inference using the Bayesian formalism (Bishop, 2006; Gelman, Carlin, et al., 2004). In this section, we more formally generalize the salmon weights example from Chapter 2.

We denote model parameters using the symbol Θ. The observed signal is denoted using the letter X. **This thesis focuses on the case of continuous RVs.** We explicitly mention parts of the thesis that relate to discrete RVs. Probability distributions are denoted using the uppercase letters P or Q, and their densities using the lowercase letters p or q. We consider interchangeably a model and the joint distribution it defines over parameters and signal: $P(X, \Theta)$.

In the context of inference, the Bayesian theory can be summarized as:

> "My new assumption about the values of the parameters is the combination of my prior assumptions and the evidence that has been presented to me".

Formally, the model incorporates some **prior** distribution for its parameters: $P(\Theta)$. Along with this prior, the model defines the **likelihood** of the signal given the parameters: $P(X|\Theta)$. The model can be fully described by the **joint** distribution: $P(X, \Theta) = P(X|\Theta) \times P(\Theta)$. Inference aims at finding the distribution of the parameters *given* the signal, called the **posterior** distribution. Using Baye's theorem, the posterior can be decomposed as:

$$\underbrace{P(\Theta|X)}_{\text{Posterior}} = \frac{\overbrace{P(X|\Theta)}^{\text{Likelihood}} \times \overbrace{P(\Theta)}^{\text{Prior}}}{\underbrace{P(X)}_{\text{Evidence}}} \tag{3.1}$$

We can further detail each factor in the above equation.

The *likelihood* $P(X|\Theta)$ is the statistical link between the parameters and the signal. In this thesis, we will both consider cases in which this distribution is *explicit* or *implicit*. The *explicit* scenario corresponds to a modeling choice from the experimenter: for instance, X can be hypothesized as the realization of a normal distribution whose mean and variance depend on the parameters Θ: $X|\Theta \sim \mathcal{N}(\mu(\Theta), \sigma(\Theta))$. The *implicit* scenario will be further detailed in Chapter 4, and corresponds to cases where we can *sample* from $P(X|\Theta)$ via a simulator, but do not have access to an exact formula for the likelihood.

The *prior* $P(\Theta)$ encapsulates the assumptions of the experimenter about the values of Θ before even seeing the observed data. This includes the experimenter's *domain knowledge*, informed by years of research on the topic. The prior is the central tool in Bayesian theory to deal with the bias/variance trade-off (Bishop, 2006). A strong prior —high density over small support— can reduce the uncertainty in the parameter estimation: the posterior precision will typically add the prior's precision to the precision due to the evidence (as in our salmon example in Section 2.1.2). In doing so, the inferred solutions are *biased* towards the solutions assumed *a priori* (independent of the evidence brought by the data). On the contrary, a weak prior —low density over a large support— will only consider the evidence due to the data at the cost of a larger uncertainty. In Chapter 8, we engineer meaningful priors to reduce the variance of the posterior $P(\Theta|X)$.

The *evidence* $P(X)$ is, sadly, unknown for most non-trivial models P. To compute $P(X)$, one would need to *marginalize* the likelihood over all the possible latent parameter values:

$$P(X) = \int \cdots \int_{\Theta} P(X|\Theta)P(\Theta)d\Theta \tag{3.2}$$

Marginalization in the evidence amounts to considering an infinite number of parameter configurations. For each parameter configuration, we can compute the probability of the observed signal. Integrating over all configurations provides a single term —the evidence— that translates how likely the observed signal is according to the model P. Expect in simple —called *conjugate*— cases in which the above integral can be analytically computed, the evidence is unknown. As the number of parameters composing Θ increases, even estimating $P(X)$ via Monte Carlo becomes infeasible. With $P(X)$ unknown, $P(\Theta|X)$ can only be evaluated up to a normalization constant —preventing the use of Bayes theorem "as is" to perform inference.

The intractability of the evidence is the central computational problem of the Bayesian theory. In the following sections, we illustrate how this hurdle can be circumvented.

3.2 Approximate inference and Markov Chain Monte Carlo (MCMC)

In this section, we give a brief overview of methods allowing to perform inference in practice. A simplified illustration of the available methods is given in Figure 3.1.

In its most general form, the inference problem is highly complex. Koller and Friedman (2009) illustrate the non-deterministic polynomial-time (NP)-hard complexity of the *exact* inference problem. They show that even *approximate* inference —computing probabilities up to an ϵ error— is also NP-hard! In practice, the structure of the hypothesized model —and the associated factorization of $P(X, \Theta)$— is critical to performing inference effectively. In low-dimensional cases, factorizing P permits local operations on "factors" composing the full distribution. This avoids needing to treat the entire joint distribution $P(X, \Theta)$ at once. Exploiting the conditional independence structure of P thus gives a basis for *exact* inference methods: variable elimination or clique-trees (Koller and Friedman, 2009). Sadly, those algorithms' computational and memory complexity is roughly exponential in the number of RVs composing Θ (a more accurate diagnostic based on the treewidth of P can be found in Koller and Friedman (2009)). This renders *exact* inference methods infeasible for many real-life problems.

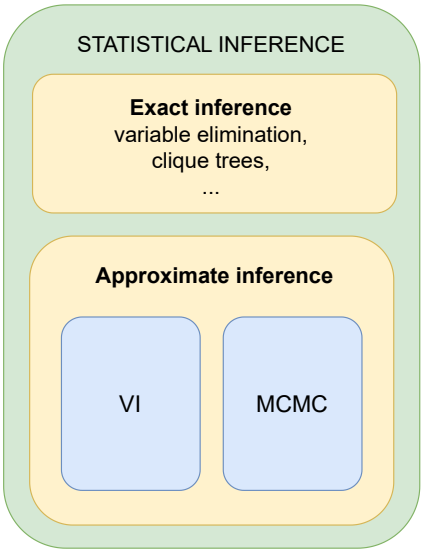

Fig. 3.1.: Simplified Venn diagram of statistical inference This thesis focuses on variational inference (VI), which is an instance of approximate inference. Approximate inference is needed to tackle many real-world inference problems, due notably to their size (number of RVs).

Computational constraints thus drive the need for *approximate* inference methods. Given the dimensionality of our target applications —illustrated in Section 1.4— **this thesis focuses on approximate inference**.

The first well-studied family of approximate inference methods is Markov chain Monte Carlo (MCMC) (Andrieu et al., 2003). MCMC infers through sampling a Markov chain whose equilibrium distribution is the posterior distribution $P(\Theta|X)$. To set $P(\Theta|X)$ as its equilibrium, MCMC relies on posterior *ratios*, which circumvents the intractability of the evidence introduced in Section 3.1:

$$\frac{P(\Theta^{\text{proposal}}|X)}{P(\Theta^{\text{current}}|X)} = \frac{P(X|\Theta^{\text{proposal}}) \times P(\Theta^{\text{proposal}}) \times \cancel{P(X)}}{P(X|\Theta^{\text{current}}) \times P(\Theta^{\text{current}}) \times \cancel{P(X)}} \tag{3.3}$$

where Θ^{current} is the current state of the chain, and Θ^{proposal} is proposed by a *kernel* as the next state of the chain. The chain's probability of jumping to Θ^{proposal} depends on this ratio, which ensures it samples from $P(\Theta|X)$. MCMC has the advantage of sampling from the *true* posterior *asymptotically*, if sufficiently many samples are drawn from the chain. Though it does not provide an *explicit* posterior, in a sense, MCMC is not so much an "approximate" method. MCMC research is, to this day, a flourishing field of research (Andrieu et al., 2003).

We mention a few MCMC research trends linked to this thesis. At the core of MCMC is the design of efficient kernels that propose the following sample in the chain. A first

improvement over random walk kernels is the integration of Hamiltonian dynamics through Hamiltonian Monte Carlo (HMC) (Duane et al., 1987). This allows for "longer jumps" between states, meaning that successive samples are less correlated, and fewer samples are needed to approximate $P(\Theta|X)$. Recently, Matthew D. Hoffman, Sountsov, et al. (2019) combined HMC with the *reparametrization trick* —which we will detail in greater length in Section 3.3.4— to yield even more efficient kernels. This research is interesting in combining elements from two fields, generative modeling —a focus of this thesis— and MCMC. We believe such combinations will be more and more fruitful in the future, breaking barriers between sometimes hermetic research fields.

Another interesting trend in MCMC research is the emergence of sequential Monte Carlo (SMC) (Moral et al., 2007). SMC uses importance (re)-sampling to evolve a population of particles into a target distribution —for instance, $P(\Theta|X)$. Contrary to traditional MCMC, SMC can leverage the parallelization capabilities of modern graphics processing unit (GPU)-accelerated software. This unified framework paves the way for GPU-accelerated hybrid methods.

In terms of limitations, MCMC is a sampling-based method. As such, MCMC can struggle in high-dimensional scenarios: this can be seen as an instance of the *curse of dimensionality* (Donoho, 2000). In general, compared to variational methods, MCMC is considered slower and more computationally intensive (David M. Blei et al., 2017). Given the large dimensionality of our target applications, we thus focus on another branch of approximate inference: variational inference (VI) (presented in Section 3.3). Yet, we underline that our omission of MCMC methods stems from a lack of time rather than a partisan mindset, and we propose in Section 10.1 some combinations between MCMC and VI. We believe in the potential of such combinations to bridge the asymptotic exactness of MCMC with the computational efficiency of VI.

3.3 Variational Inference (VI): an optimization take on approximate inference

This section details the branch of approximate inference we focus on: **variational inference (VI)**. We start with a general definition that introduces the concept of a variational family. Next, we define inference *amortization*, a key concept in this thesis. We then expand on different types of losses that can be used to train variational densities. We finish with practical considerations regarding VI's implementation.

We voluntarily present a modern, machine learning-oriented take on VI. Note that this section is not meant to be exhaustive, as modern trends of VI will also be presented in Chapter 5.

3.3.1 The variational family and the approximation gap

VI frames inference as an optimization problem.

VI starts by defining a variational family, denoted $\mathcal{Q}$, of distributions over the parameters Θ. Then, VI finds inside $\mathcal{Q}$ the distribution $Q(\Theta)$ *closest* to the target distribution $P(\Theta|X)$. How we measure the *closeness* of Q to its target will determine the *loss* minimized during optimization. We detail different usable losses in Section 3.3.3.

The optimized *weights* parameterize the different distributions inside the variational family. For instance, a simple variational family corresponds to parametric Gaussians $\mathcal{Q} = \{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma \in \mathbb{R}^{+*}\}$. Using this example family, VI would reduce to finding the values for the mean and variance that would yield the Gaussian closest to $P(\Theta|X)$. In general, VI amounts to the following equation:

$$\begin{aligned} \mathcal{Q} &= \{Q(\Theta; \phi); \phi \in \text{Dom}(\phi)\} \\ \phi^* &= \underset{\phi \in \text{Dom}(\phi)}{\arg\min}\ \text{Div}(Q(\Theta; \phi)||P(\Theta|X)) \end{aligned} \tag{3.4}$$

where $\text{Dom}(\phi)$ denotes the domain of the weights ϕ, and Div denotes a divergence between distributions, null if and only if $Q(\Theta; \phi) = P(\Theta|X)$.

An important caveat about VI is that, after the optimization, VI does not necessarily yield $P(\Theta|X)$. VI yields the distribution best approximating $P(\Theta|X)$ *inside the variational family*. A trivial example of VI's failure is if one tries to fit the multimodal distribution in Figure 3.4 (top) using a simple Gaussian (bottom right). The actual form of the distribution is very far from a Gaussian: no matter how effective the optimization, the resulting approximation will always be poor. This asymptotic limit on VI performance has been coined the *approximation gap* by Cremer et al. (2018).

The crux of VI can be understood from Figure 3.2. We must design a variational family $\mathcal{Q}$ as expressive as possible to minimize the approximation gap. In Figure 3.2, this amounts to making the $\mathcal{Q}$-cloud larger and larger so that it covers $P(\Theta|X)$. Yet, the most expressive variational families are not necessarily the most amenable to optimization. In Figure 3.2, the red optimization arrow can be so convoluted that the optimal solution Q^* is never found in a reasonable time (Bottou and Bousquet,

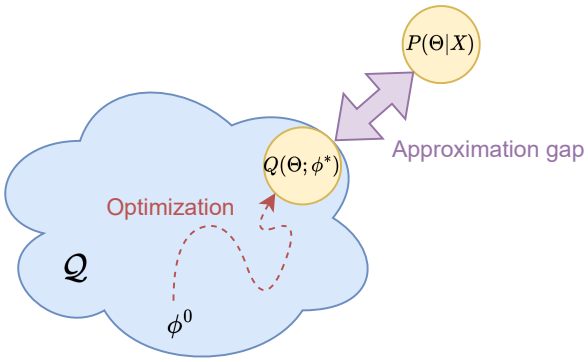

Fig. 3.2.: Variational Inference (VI) The $\mathcal{Q}$-cloud represents the variational family, inside which we optimize to find the distribution $Q(\Theta;\phi^*)$ closest to $P(\Theta|X)$. VI optimizes the weight ϕ from an initial value ϕ^0 to the optimal values ϕ^*. The ability of Q^* to approximate its target well depends on the expressivity of $\mathcal{Q}$, as measured by the approximation gap.

2007). Finding expressive yet computationally effective variational families is thus an active research area (Weilbach et al., 2020; Titsias and Ruiz, 2019; Ambrogioni, Silvestri, et al., 2021). In this thesis, we strive to strike a particular balance between **expressivity, computational efficiency, and scalability** to large dimensions.

3.3.2 Inference amortization, and the amortization gap

Here, we define inference amortization, a key concept in this thesis.

As explained in Section 3.3.1, VI searches for the distribution Q that best approximates the posterior of Θ for a given observed value $\mathbf{X}_0$ for the signal X. Denoting the corresponding optimized weights ϕ_0 yields the approximation: $Q(\Theta;\phi_0) \simeq P(\Theta|X=\mathbf{X}_0)$. When presented with a new data point $\mathbf{X}_1$, optimization has to be performed again to search for the weights ϕ_1, such that $Q(\Theta;\phi_1) \simeq P(\Theta|X=\mathbf{X}_1)$. Focusing on a given value of the signal is the default paradigm of inference. For instance, MCMC —described in Section 3.2— needs to run a new chain when presented with a new value for the signal.

Instead of inferring "from scratch", *sample amortized* VI (C. Zhang et al., 2019; Cremer et al., 2018) regresses the weights ϕ using an *encoder* f of the observed signal $\mathbf{X}$: $q(\Theta;\phi = f(\mathbf{X}_i)) \simeq p(\Theta|X=\mathbf{X}_i)$. The cost of learning the encoder weights is *amortized* since inference over any new signal $\mathbf{X}_i$ requires no additional optimization. To underline that we now learn the posterior of Θ for *any* value of X, we denote the amortized variational family $Q(\Theta|X)$ instead of the non-amortized $Q(\Theta)$.

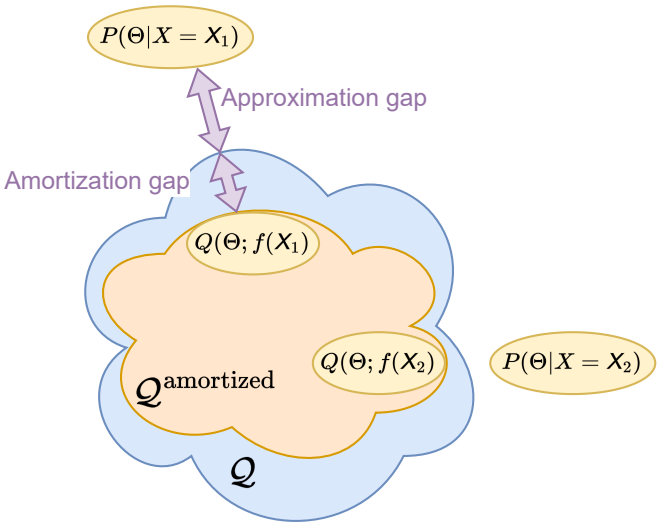

Fig. 3.3.: The amortization gap (Cremer et al., 2018) On top of finding the optimal weights ϕ_0 and ϕ_1 corresponding to the signals x_0 and x_1, amortized inference requires to learn an encoder f such that $\phi_1 = f(x_1)$ and $\phi_2 = f(x_2)$. This both computationally complicates inference and reduces the expressivity of the amortized family $\mathcal{Q}^{\text{amortized}}$ —due to the limited expressivity of the encoder f.

Amortization can massively speed up inference in high-dimensional contexts. For instance, consider microstructure estimation as described in Section 1.3. Running inference from scratch on hundreds of thousands of voxels can take a massive time. In contrast, it is much faster to train an amortized estimator once and reuse it on every voxel by feeding its signal to the encoder (Jallais and Palombo, 2023).

Yet, if amortization is attractive from a computational point of view, it also complicates inference.

From a theoretical point of view, amortization can reduce the expressivity of the variational family, introducing on top of the approximation gap an *amortization gap* (Cremer et al., 2018). Indeed, in addition to finding the optimal weights ϕ_0 and ϕ_1 corresponding to the signals $\mathbf{X}_0$ and $\mathbf{X}_1$, amortization requires learning an encoder that perfectly maps $\mathbf{X}_0 \mapsto \phi_0$ and $\mathbf{X}_1 \mapsto \phi_1$. Our experiments in Chapter 6 illustrate this amortization gap, which becomes more apparent as the complexity of the model P (and thus its posterior) increases. We illustrate the amortization gap in Figure 3.3.

We'll repurpose the concept of amortization in Chapter 6 to improve inference's computational efficiency.

3.3.3 Distribution divergences as optimization losses

This section details how the closeness of $Q(\Theta)$ to $P(\Theta|X)$ can be measured in practice. This closeness plays the role of a minimized loss during optimization.

The most prominent way to measure closeness is based on the **Kullback Leibler divergence (KL)**, which takes its roots in information theory. Considering two distributions P_1 and P_2, the KL divergence from P_2 to P_1 is defined as:

$$\begin{aligned} \mathrm{KL}(P_1||P_2) &= \mathbb{E}_{P_1}\left[\log \frac{p_1}{p_2}\right] \\ &= \int_{x\in\mathcal{X}} p_1(x) \times \log \frac{p_1(x)}{p_2(x)} dx \end{aligned} \tag{3.5}$$

Where $\mathcal{X}$ denotes the space over which P_1 and P_2 are defined. An important feature of the KL is that it is *not* symmetric: $\mathrm{KL}(P_1||P_2) \neq \mathrm{KL}(P_2||P_1)$ —that is why the KL is called a *divergence* and not a *distance*. From a computational perspective, this asymmetry is important. To compute $\mathrm{KL}(P_1||P_2)$ using Monte Carlo —using discrete samples to evaluate the integral in Equation (3.5)— requires sampling from P_1 *and* evaluating p_1. In contrast, it requires *only* evaluating p_2. This difference has important consequences in terms of implementation, as detailed in Section 3.3.4.

Since the KL is asymmetric, in the context of VI it yields two different losses.

The reverse KL divergence (r-KL) The r-KL corresponds to computing the divergence from $P(\Theta|X)$ to $Q(\Theta)$:

$$\begin{aligned} \mathrm{Div}_{\text{r-KL}}(Q(\Theta)||P(\Theta|X)) &= \mathrm{KL}(Q(\Theta)||P(\Theta|X)) \\ &= \mathbb{E}_Q\left[\log \frac{q(\Theta)}{p(\Theta|X)}\right] \end{aligned} \tag{3.6}$$

In the context of inference, the r-KL cannot be computed because we do not know the posterior $P(\Theta|X)$ and cannot evaluate its density. However, rewriting the r-KL yields a usable loss:

$$\begin{aligned}
\mathrm{KL}(Q(\Theta)||P(\Theta|X)) &= \mathbb{E}_Q\left[\log \frac{q(\Theta)}{p(\Theta|X)}\right] \\
&= \mathbb{E}_Q\left[\log q(\Theta) - \log p(\Theta|X)\right] \\
&= \mathbb{E}_Q\left[\log q(\Theta) - \log p(\Theta, X)\right] + \mathbb{E}_Q\left[\log p(X)\right] \\
&= \mathbb{E}_Q\left[\log q(\Theta) - \log p(\Theta, X)\right] + \log p(X) \\
&\propto -\mathrm{ELBO}(Q) \\
\mathrm{ELBO}(Q) &= \mathbb{E}_Q\left[\log p(\Theta, X) - \log q(\Theta)\right] \\
\log P(X) &= \mathrm{KL}(Q(\Theta)||P(\Theta|X)) + \mathrm{ELBO}(Q)
\end{aligned} \tag{3.7}$$

Evidence — Divergence ≥ 0 — Lower bound

Equation (3.7) reveals the evidence lower bound (ELBO) term. As the last line in Equation (3.7) underlines, since the KL term is always positive, the ELBO is always lower than the evidence term $\log P(X)$ (hence its name). Contrary to the KL, the ELBO only depends on computable terms: the variational and joint densities. Maximizing the ELBO is equivalent to minimizing the KL since the evidence term, though of unknown value, is fixed with respect to Q (it only depends on P).

The ELBO is the historical centerpiece of VI (David M. Blei et al., 2017). It combines multiple advantages:

- computing the ELBO using Monte Carlo is cheap, only requiring sampling from the variational distribution and evaluating the variational and joint densities;
- comparing different variational distributions Q —possibly from different variational families— over the same model P, differences in ELBO directly translate differences in KL divergence;
- provided perfect approximation, $\mathrm{KL}(Q(\Theta)||P(\Theta|X)) = 0$ and $\mathrm{ELBO}(Q) = \log P(X)$, meaning the ELBO provides an efficient estimator for the evidence, usable in contexts such as Bayesian model comparison (Gelman, Carlin, et al., 2004).

Nonetheless, the r-KL has notable drawbacks that will be detailed in Section 3.4.

The forward KL divergence (f-KL) Opposite to the r-KL, the f-KL corresponds to computing the divergence from $Q(\Theta)$ to $P(\Theta|X)$:

$$\begin{aligned}\text{Div}_{\text{f-KL}}(Q(\Theta)||P(\Theta|X)) &= \text{KL}(P(\Theta|X)||Q(\Theta)) \\ &= \mathbb{E}_{P(\Theta|X)}\left[\log \frac{p(\Theta|X)}{q(\Theta)}\right]\end{aligned} \tag{3.8}$$

In this form, the f-KL cannot be evaluated, because we do not know the posterior $P(\Theta|X)$ and cannot evaluate its density nor sample from it. Contrary to the r-KL, no rewriting allows using the f-KL in the non-amortized context.

However, switching to the amortized setup —described in Section 3.3.2— we can sample from the joint distribution $P(X, \Theta)$. Since we now learn the posterior of Θ for *any* value of X, we can replace $Q(\Theta)$ by $Q(\Theta|X)$. The following *amortized* f-KL can be evaluated:

$$\begin{aligned}\text{Div}^{\text{amo}}_{\text{f-KL}}(Q(\Theta|X)||P(\Theta|X)) &= \mathbb{E}_{P(X)}\left[\mathbb{E}_{P(\Theta|X)}\left[\log \frac{p(\Theta|X)}{q(\Theta|X)}\right]\right] \\ &= \mathbb{E}_{P(X)}\left[\mathbb{E}_{P(\Theta|X)}\left[\log p(\Theta, X) - \log p(X) - \log q(\Theta|X)\right]\right] \\ &= \mathbb{E}_{P(X)}\left[\mathbb{E}_{P(\Theta|X)}\left[\log p(\Theta, X) - \log q(\Theta|X)\right] - \log p(X)\right] \\ &\propto \mathbb{E}_{P(X,\Theta)}\left[\log p(\Theta, X) - \log q(\Theta|X)\right] \\ &\propto \mathbb{E}_{P(X,\Theta)}\left[-\log q(\Theta|X)\right]\end{aligned} \tag{3.9}$$

amortization thus yields a computable loss, yet comes at a cost as explained in Section 3.3.2. Strategies to mitigate that cost are presented in Sections 4.2.2 and 10.1.

Marginal inference Contrary to the r-KL, the f-KL is easily amenable to **marginalizing** unwanted parameters. Splitting the parameters $\Theta = \Theta^{\text{target}} \cup \Theta^{\text{other}}$, it is possible to focus solely on the inference over Θ^{target} by defining a variational distribution $Q(\Theta^{\text{target}}; \phi)$ and training over the loss:

$$\text{Div}^{\text{amo}}_{\text{f-KL}}(Q(\Theta^{\text{target}}|X)||P(\Theta^{\text{target}}|X)) \propto \mathbb{E}_{X,\Theta^{\text{target}},\Theta^{\text{other}}\sim P(X,\Theta)}\left[-\log q(\Theta^{\text{target}}|X)\right] \tag{3.10}$$

In practice, the variational distribution now targets the *marginal* posterior of Θ^{target}, integrating over all the possible realizations of Θ^{other}. In some cases, Θ^{other} corresponds to high-dimensional intermediate states (e.g. in Chapter 9 a latent time series) while Θ^{target} are low-dimensional parameters (the noise level in the series). Using the f-KL then bypasses the large dimensionality of the problem, to focus inference on the key parameters only. We exploit this feature in Chapters 8 and 9.

Alternative divergences This thesis primarily focuses on the r-KL and f-KL. As an opening, we mention a few other distribution divergences that can be used to measure the closeness of $Q(\Theta)$ to $P(\Theta|X)$. Those divergences can be used instead of the KL as an optimization loss.

Note: This sub-section is more mathematically involved and refers to terms defined in Section 3.4.

A generalization of the KL is Rényi's α-divergences (Y. Li and Turner, 2016), which enables a smooth interpolation from the ELBO to the log-likelihood. Different values for α yield the KL divergence, or divergences proportional to the Hellinger or χ^2. Typically, larger α values encourage mass-covering properties —useful in multi-modal contexts— but can create instability when $Q(\Theta)$ is much different from $P(\Theta|X)$. As a consequence, *α-adaptive* methods have also been developed (Dilin Wang et al., 2018). Divergences can also be defined using the more general framework of Csiszár f-divergences (Ali and Silvey, 1966). In doing so, closeness with the objective function can be measured using the Jensen-Shanon or total variation divergences. Different ways to compare distributions can also be envisioned. For example, Ambrogioni, Güçlü, Güçlütürk, et al. (2018) propose a loss based on the Wasserstein distance. Recently, Modi et al. (2023) propose score matching —matching the gradients of the log density between $Q(\Theta)$ and $P(\Theta|X)$.

The usage of alternatives to the KL was mostly omitted during this thesis. Nonetheless, we argue there is potential in integrating divergences adapted to the inference problem at hand as part of automatic variational inference (AVI) (Ranganath et al., 2013).

3.3.4 VI in the machine learning era

Here, we describe how VI can be implemented in practice, leveraging modern software and hardware.

Historically, VI required mathematical mastery. Starting from the model P and the variational family $\mathcal{Q}$, an optimization routine had to be derived using pen and paper. That is to say, a set of equations to update the weights of the variational family at each optimization step. This approach is similar in spirit to the expectation maximization (EM) algorithm (Bishop, 2006). To derive tractable equations, statisticians often resorted to *conjugate* distributions. This means that the choice of variational family and the ability to solve the optimization were intertwined. This mathematical legwork remains computationally attractive where applicable (Thomas Yeo et al.,

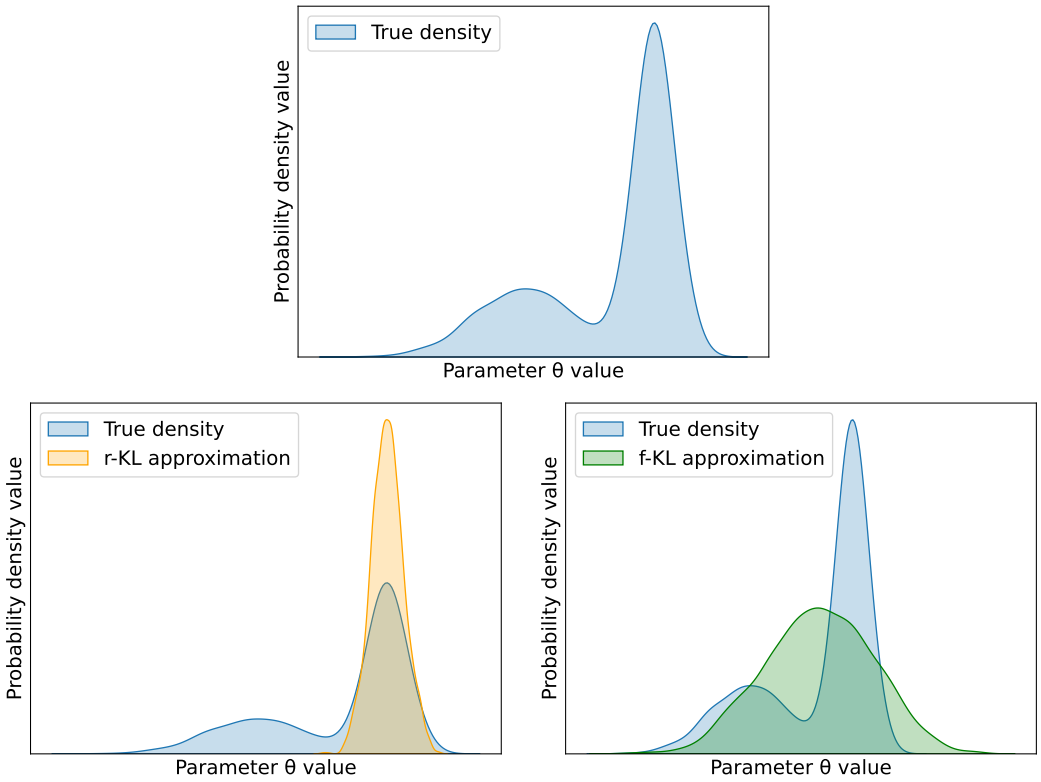

Fig. 3.4.: Illustration of the r-KL and f-KL behavior (Bishop, 2006) *On top*, a bi-modal target distribution. *On the left*, a reverse Kullback Leibler divergence (r-KL)-fitted Gaussian approximation. The r-KL is mode-seeking. Even a more expressive family than parametric Gaussians wouldn't necessarily cover the full target. *On the right*, a forward Kullback Leibler divergence (f-KL)-fitted Gaussian approximation. The f-KL is moment-matching and enforces a coverage of the full target's support. A more expressive family than parametric Gaussians would better match the target.

2011; Kong, J. Li, et al., 2019; Dao et al., 2021). However, it creates substantial barriers to entry for novice experimenters.

In contrast, this thesis focuses on **automatic differentiation variational inference (ADVI)** (Kucukelbir et al., 2016; Ranganath et al., 2013). Implementing P and Q using automatic differentiation software allows propagating $\mathrm{Div}(Q(\Theta;\phi)||P(\Theta|X))$ —the loss— back to the weights ϕ of the variational family (Dillon et al., 2017; Bingham et al., 2019). Contrary to the manual derivations above, ADVI does not constrain the experimenter in the choice of model or variational family and requires little mathematical mastery.

What's more, implementing VI using automatic differentiation software leverages the powerful tools developed in the machine learning community:

- optimizers and schedulers to update the variational family weights (Diederik P. Kingma and Ba, 2015);
- neural network architectures that can be intertwined with stochastic functions (Bishop et al., 1995);
- GPU acceleration, which can massively speed up parallel computation.

In the rest of this section, we mention some implementation details at the core of modern probabilistic software.

Monte Carlo integration The losses in Equations (3.7) and (3.9) include expectations over distributions —respectively over $Q(\Theta)$ and $P(X,\Theta)$. Calculating analytically the corresponding integrals is, in practice, computationally infeasible. Instead, we can rely on the Monte Carlo unbiased estimator:

$$\begin{aligned}\mathbb{E}_{P_1}\left[f(x)\right] &= \int_{x\in\mathcal{X}} f(x)p_1(x)dx \\ &\simeq \frac{1}{N}\sum_{x_i\sim P_1} f(x_i)\end{aligned} \tag{3.11}$$

Replacing the integral with a mean renders losses computationally tractable. As the number of samples N from P_1 increases, the estimator's variance decreases: the mean value is a more faithful estimate of the integral. In machine learning, this is linked to mini-batching. The $f(x_i)$ can also be estimated independently, which means that Monte Carlo integration is amenable to GPU parallelization.

Application to f-KL Applying Monte Carlo integration renders the f-KL loss computationally tractable:

$$\mathcal{L}^{\text{f-KL}}(\phi) = \frac{1}{N} \sum_{X_i, \Theta_i \sim P(X,\Theta)} -\log q(\Theta_i; f(X_i, \phi)) \tag{3.12}$$

where we assume that the density q is differentiable with respect to ϕ. Using Monte Carlo integration for the f-KL leads to the following steps:

1. sample from the joint distribution $p(X, \Theta)$, which is amenable to the *implicit* setup as described in Chapter 4;

2. for each sample, feed the value of X to the encoder of the amortized $Q(\Theta|X)$;

3. conditioned by X, maximize the density q over the value of Θ

This framework is similar to the supervised learning setup in machine learning: training over an i.i.d dataset, we feed to a (probabilistic) regressor a feature (X_i) and maximize a density over a target (Θ_i).

Application to r-KL Applying Monte Carlo integration yields:

$$\mathcal{L}^{\text{r-KL}}(\phi) = \frac{1}{N} \sum_{\Theta_i \sim Q(\Theta;\phi)} \log q(\Theta_i; \phi) - \log p(X, \Theta_i) \tag{3.13}$$

where we assume that the densities p and q are differentiable with respect to ϕ. Using Monte Carlo integration for the r-KL leads to the following steps:

1. sample from the variational distribution $Q(\Theta; \phi)$;

2. for each sample, evaluate both the variational density q over Θ and the joint density p between Θ and the observed signal X;

3. maximize the ELBO.

However, directly differentiating through those steps is not possible for subtle reasons. In the f-KL case, the expectation was taken over $P(X, \Theta)$, a distribution that does not depend on the weights ϕ of the variational family. In contrast, in the r-KL, the expectation is taken with respect to $Q(\Theta; \phi)$. As differentiation through an expectation is not well-defined, an additional implementation detail is necessary for ADVI to work.

The reparameterization trick How to differentiate through an expectation over a parametric distribution? A broadly used solution to this problem is the reparameterization of $Q(\Theta;\phi)$. Instead of a parametric distribution, we can rewrite Q as a parametric transformation of a fixed distribution. As an example, in the Gaussian case:

$$\Theta \sim \mathcal{N}(\mu,\sigma) \iff \Theta = \mu + \sigma U \quad \text{with} \quad U \sim \mathcal{N}(0,1) \tag{3.14}$$

a parametric Gaussian can be rewritten as a parametric affine transformation of a standard Gaussian. Many standard distributions are amenable to such a rewriting. In Section 4.1, we introduce a class of transformations called normalizing flows (NFs) that leverage reparameterization to produce powerful density approximators (Papamakarios, Nalisnick, et al., 2019).

To compute an expectation over a parametric distribution, we can leverage the reparameterization trick:

$$\begin{gathered}\Theta \sim Q(\Theta;\phi) \iff \Theta = T(U;\phi) \quad \text{where} \quad U \sim P_U \\ \frac{\partial}{\partial\phi}\mathbb{E}_{Q(\Theta;\phi)}\left[f(\Theta)\right] = \mathbb{E}_{P_u(U)}\left[\frac{\partial}{\partial\phi} f(T(U;\phi))\right]\end{gathered} \tag{3.15}$$

Applying this to the r-KL, we can reparameterize Q to differentiate through Equation (3.13) —both through the sampling *and* density evaluation.

Differentiation without reparameterization The reparametrization trick is often sufficient to implement ADVI. Yet, in some cases, Q cannot be fully reparameterized, and alternative strategies are required. A typical scenario where the reparameterization trick is not applicable is the presence of *discrete* RVs in Θ. In detail, it is possible to differentiate through the evaluation of the *probability* of discrete RVs. However, it is impossible to differentiate through their *sampling*. We present three strategies to circumvent this issue.

Any RV: REINFORCE gradient estimation Computing the expectation can be done using the log-derivative trick:

$$\frac{\partial}{\partial\phi}\mathbb{E}_{Q(\Theta;\phi)}\left[f(\Theta;\phi)\right] = \mathbb{E}_{Q(\Theta;\phi)}\left[\frac{\partial}{\partial\phi} f(\Theta;\phi) + f(\Theta)\frac{\partial}{\partial\phi}\log q(\Theta;\phi)\right] \tag{3.16}$$

Using this trick, the density q needs to be differentiable, but not the *sampling* from Q. This makes the REINFORCE estimator usable for a broad class of distributions, including discrete distributions. However, the REINFORCE estimator has a large variance, making it computationally impractical. Learnable control variates can help reduce this variance (Tucker et al., 2017; R. Liu et al., 2019).

Discrete RVs: continuous reparameterization Another solution to treat discrete RVs is to approximate those through continuous distributions. This can be done using the Gumbell-Softmax trick (Grathwohl et al., 2018). Note that at high temperatures, this approximation is imperfect and introduces a bias compared to the discrete case. Annealing from high to low temperatures can lead the approximation from a more numerically stable to a more exact regime. We apply this solution in Chapter 7.

Discrete RVs: enumeration Treating discrete RVs can also be done efficiently via enumeration. To prevent marginalization over an exponentially large number of worlds, variable elimination has been recently revisited on tensors by Aitchison (2019) and Obermeyer et al. (2019). We use this solution in Chapter 8.

Software implementation This section presented a list of essential building blocks to implement ADVI. Via Monte Carlo integration, reparameterization, and automatic differentiation, any variational family $\mathcal{Q}$ can be used to infer over any model P, in a fast and scalable way. We conclude this presentation by mentioning two software packages that we extensively used:

- `Tensorflow probability (TFP)` is a probabilistic library based on top of Tensorflow (Dillon et al., 2017; Martín Abadi et al., 2015). TFP is a relatively low-level application programming interface (API) that provides much control to the user. We found TFP particularly useful to prototype complex variational families. But in high dimensional cases, we often struggled with numerical instability.

- `Pyro` is built on top of Pytorch (Bingham et al., 2019; Paszke et al., 2019). Pyro specializes in stochastic variational inference (SVI) —which we present in Section 5.2.2. Pyro provides high-level APIs, making it easier to use for beginners. However, this level of automation also makes Pyro harder to customize. Pyro also provides powerful state-of-the-art optimization routines, making it computationally attractive in high-dimensional cases.

We would generally encourage users interested in SVI to use the Pyro library. We also underline that both libraries implement other probabilistic methods, including MCMC, SMC, or importance sampling.

3.4 Hurdles in statistical inference

This short section reviews a few common obstacles to inference encountered during this thesis and serves as a counterpoint to Section 1.4. Modeling in Neuroimaging requires dealing with both uncertainty and high dimensionality. As we have shown so far, statistical inference can deal with uncertainty. But, as detailed below, high dimensionality in itself complicates inference. From a technical standpoint, the hurdles detailed here serve as additional motivations for this thesis.

3.4.1 Distribution complexity: multi-modality and heavy tails

A common hurdle in inference is the complex shape of the unknown posterior distribution. Consider Figure 2.1 as the posterior distribution that inference should retrieve. It presents two complex features: multi-modality and heavy tails, which we define below, and illustrate in Figure 3.5.

Multi-modality and mode collapse Multi-modality corresponds to different regions in the Θ-space that may explain the observed signal X. For example, consider microstructure estimation described in Section 1.3. If the observed water diffusion in a given direction is low, is it because of a low diffusivity? Or because the direction is orthogonal to the direction of the neuron axons? Unlikely parameter configurations may separate those highly-explanatory regions. This creates distinct distribution **modes**, separated by low-density regions.

To explain why multi-modality may impede inference, recall that inference methods usually explore the parameter space locally. MCMC runs a Markov Chain where the next proposed state is a perturbation near the current state. Conversely, VI smoothly interpolates distributions between its start and end points during optimization. Because of this locally exploratory behavior, inference methods may get "stuck" in one high-explanatory region. Consequently, inference methods may ignore alternative —potentially equally explanatory— modes. This behavior is coined **mode collapse**, and is illustrated in Figure 3.5 (left).

Mode collapse using MC methods Using MCMC, a standard diagnostic to detect multi-modality is running multiple chains separately. The experimenter can then assess the mixing of the chains: chains stuck in different modes would not mix. In the same vein, ensembling multiple chains can help tackle multi-modality (Foreman-Mackey

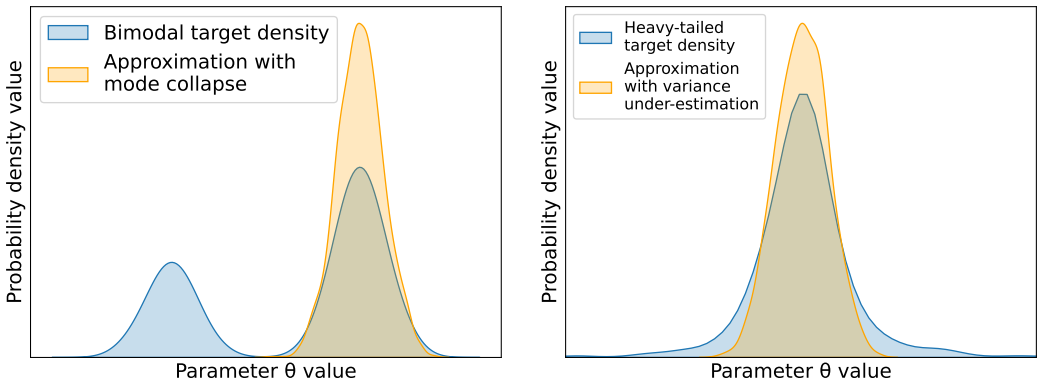

Fig. 3.5.: Pitfalls in inference *On the left*: an illustration of mode collapse. The variational approximation focuses on the distribution mode on the right and ignores the left one. *On the right*: variance under-estimation. The variational approximation ignores the low-density tails surrounding the distribution mode. As a result, it underestimates the uncertainty in the true distribution.

et al., 2013). SMC is another promising method in that direction since it is akin to running multiple importance-resampled "chains" in parallel.

Mode collapse using VI The r-KL —described in Section 3.3.3— is notably **mode-seeking** (David M. Blei et al., 2017). This feature is both a blessing and a curse: the r-KL provides quick convergence but is prone to getting stuck in one posterior mode. In contrast, the f-KL "forces" the variational density to consider all the modes in the distribution and prevents mode collapse. To understand why this is the case, we reproduce here the expression of both divergences:

$$\begin{aligned} \mathrm{Div}_{\text{r-KL}}(Q(\Theta)||P(\Theta|X)) &= \mathbb{E}_{Q}\left[\log \frac{q(\Theta)}{p(\Theta|X)}\right] \\ \mathrm{Div}_{\text{f-KL}}(Q(\Theta)||P(\Theta|X)) &= \mathbb{E}_{P(\Theta|X)}\left[\log \frac{p(\Theta|X)}{q(\Theta)}\right] \end{aligned} \tag{3.17}$$

both divergences evaluate the difference between the log densities q and p. In the r-KL however, this difference is integrated over the support of the variational distribution. If Q does not cover some modes of P, the difference will not be evaluated on that mode. In contrast, the f-KL integrates the same difference over the support of the target distribution P. This forces Q to consider all of the modes in P —as illustrated in Figure 3.4.

Heavy tails and variance under-estimation Distribution **tails** correspond to the low-density regions surrounding the high-density modes. Due to mechanisms similar to mode collapse, inference methods tend to ignore distribution tails, focusing on

highly explanatory regions. Yet those tails may be "heavy", meaning that the density slowly decreases away from the modes, creating a "flat" distribution. As a result, inference methods ignoring those heavy tails may (significantly) underestimate the posterior variance —as illustrated in Figure 3.5 (right). Due to its mode-seeking properties, variance underestimation is a known issue with r-KL (David M. Blei et al., 2017; Giordano et al., 2015; B. Wang and Titterington, 2005). Training with the f-KL can help prevent variance under-estimation.

3.4.2 High dimensionality

Here we briefly mention high dimensionality as a major hurdle with inference. Chapter 5 provides a more detailed treatment in the context of VI.

To illustrate the issue with dimensionality, we can draw a parallel between inference and generative modeling (Bond-Taylor et al., 2022). Inference is akin to estimating an unknown distribution: the posterior. To learn $P(\Theta|X)$, statistical methods use samples in the Θ-space via Monte Carlo. In MCMC, the samples are the states from the Markov Chain. In SMC, the samples are the particles. In VI, using the f-KL, we draw the samples from the joint $P(X, \Theta)$. Using the r-KL, we draw samples from the variational $Q(\Theta; \phi)$.

To estimate correctly a density from samples, a rule of thumb would be to have at least one sample in each ϵ-hypersphere in the hypercube of the distribution's support. As the dimensionality of the space augments, the volume of those spheres vanishes exponentially. Thus, exponentially many samples are required to estimate the density correctly —as per the curse of dimensionality (Donoho, 2000). Without special attention, this makes high dimensional density estimation ill-posed.

High dimensionality also creates hurtful synergies with other hurdles. For instance, consider mode collapse and log-density-based losses (such as the r-KL). Imagine that the target distribution is bi-modal, but the variational distribution focuses on one single mode. In terms of log density, this mistake corresponds numerically to a $\log(2) = 0.3$ error. In high-dimensional contexts, the number of dimensions roughly multiplies log-based losses, which easily reach thousands of units. On top of this, Monte Carlo estimators also get noisier as the dimensionality increases, with a variance reaching hundreds of units. All in all, the numerical cost of dropping a mode may get numerically buried.

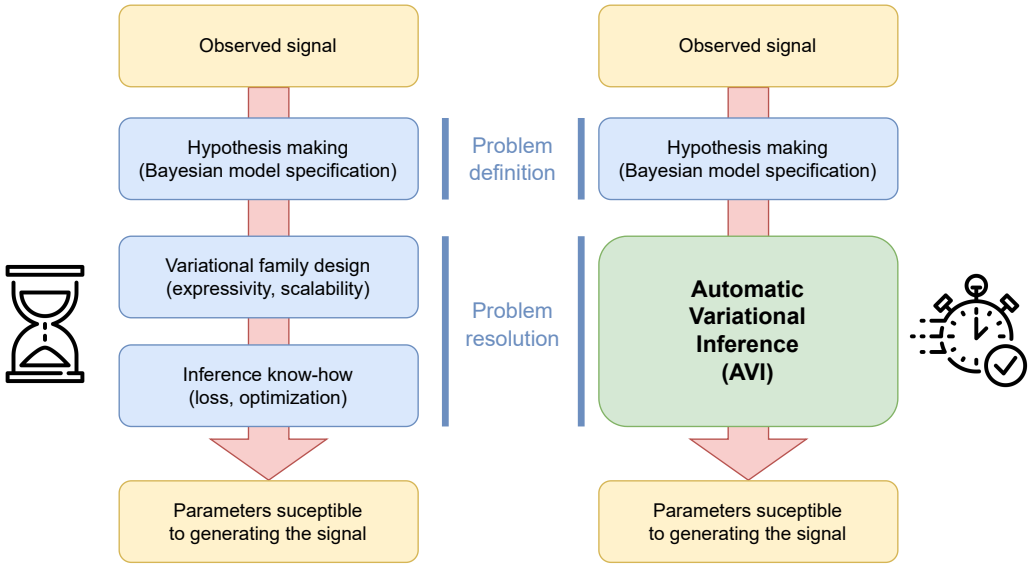

Fig. 3.6.: Automatic variational inference (AVI) On the left: a manual and lengthy research cycle. Time and effort are spent not only on the problem definition but also on its resolution. On the right: AVI automates the variational family design and the inference loop. This reduces methodological barriers to entry and speeds up the research cycle.

3.4.3 Technical mastery and automatic inference

As this thesis pursues translational research, we underline one element often omitted in statistics: the skillset of the experimenter.

Consider ourselves neuroscientists wanting to identify a latent parameter from an observed MRI signal. First, we need to link the latent parameter to the signal via a model P. This represents the first methodological skill set: Bayesian model specification. Second, we need to select an inference method, with various considerations. What is the dimensionality of the problem? Do we expect a complex posterior? This represents a second methodological skill set: statistical inference. Choosing VI as our inference method, next we need to choose a variational family $\mathcal{Q}$. Do we expect multiple modes and/or heavy tails, or can we default to Gaussians? What statistical dependencies do we want to model inside Θ? This represents a third methodological skill set: variational family design. Lastly, we need to fit the variational family over the signal. As detailed in Section 3.3.4, this step used to require involved mathematical derivations. Pen and paper can be avoided by relying on automatic differentiation, but using the latter requires altogether different skills. How to implement a training loop on GPU? Which optimizers should we use? How can we stabilize the training? This represents a fourth methodological skill set: optimization and machine learning.

In this thesis, we explore potential simplifications of this research loop, as illustrated in Figure 3.6. Similar to the *Neurolang* approach, we strive to design automatic inference methods (Zanitti et al., 2022). Following that design, the experimenter would focus on the problem *definition*, while the problem *resolution* would be automated. An example of this design is query languages such as structured query language (SQL): a user asks a question to an existing database, but does not specify *how* that question should be computationally answered. Can a similar design be envisioned for statistical inference?

In VI, **automatic variational inference (AVI)** pursues the automatic derivation of the variational family $\mathcal{Q}$ from the model P (Kucukelbir et al., 2016; Ambrogioni, Lin, et al., 2021; Ambrogioni, Silvestri, et al., 2021). Separately, the Machine Learning community has designed many wrappers with different levels of abstraction: see *AutoML* (Hutter et al., 2019), *Skorch* (Tietz et al., 2017) or *Keras* (Chollet et al., 2015). Combining those elements, end-to-end APIs could be designed and statistical inference would only require Bayesian modeling skills from the experimenter.

Under this automation angle, this thesis studies two questions. First, **what is the applicability of VI to solve complex, hierarchical and high-dimensional Neuroimaging problems?** And in particular, **can expressive yet scalable variational families be automatically derived from the model?**

Part conclusion

This first part (Part I) presented general background information motivating this thesis. We tackle inference in Neuroimaging: finding the parameters susceptible to yield the observed MRI signal through an experimenter-specified HBM. This task is complexified by the massive dimensionality inherent to Neuroscience. To perform inference, our methodology of choice is variational inference (VI), based on optimization. We leverage machine learning techniques to modernize traditional hierarchical modeling. Our objective is to design expressive variational families that would scale well to high-dimensional, hierarchical problems. In the next part (Part II), we review modern trends in inference, linked to machine learning, that we'll exploit throughout our contributions (Part III).

Part II

Modern trends in inference

4 Simulation-based inference (SBI)

This chapter reviews a modern trend in inference dubbed simulation-based inference (SBI) by Cranmer et al. (2020). Specifically, we focus on the training of surrogate densities when the link between the parameters and the signal is only *implicitly* defined by a simulator. We start by presenting normalizing flows (NFs), a family of powerful density approximators. We then move on to their application in likelihood-free inference. We finish with some opening on the fruitful combination of these techniques with VI.

4.1 Normalizing flows (NFs): powerful density approximators

This section introduces NFs, a family of density approximators recently developed in the machine learning community (Rezende and Mohamed, 2015). Our objective is to underline the potential of the technology and to provide some basics to understand the rest of this thesis. We refer interested readers to the review from Papamakarios, Nalisnick, et al. (2019).

4.1.1 General definition

NFs leverage the reparameterization trick we presented in Section 3.3.4. The term "normalizing" comes from their reparametrization of a Gaussian distribution:

$$\begin{aligned} V &= T(U;\phi) && \text{where} \quad U \sim \mathcal{N}(\vec{0},\vec{1}) \\ p_V(V) &= p_U(U) \times |\det J_T(U;\phi)|^{-1} && \text{where} \quad U = T^{-1}(V;\phi) \end{aligned} \tag{4.1}$$

$p_V(V)$: Push-forward density; $p_U(U)$: Base density; $|\det J_T(U;\phi)|^{-1}$: Change of volume

J_T designates the Jacobian of the transformation T, the matrix of its partial derivatives:

$$J_T(U) = \begin{bmatrix} \frac{\partial T_1}{\partial U_1} & \cdots & \frac{\partial T_1}{\partial U_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_D}{\partial U_1} & \cdots & \frac{\partial T_D}{\partial U_D} \end{bmatrix} \tag{4.2}$$

where D denotes the dimensionality of the space over which both densities are defined. In Equation (4.1), T is a **diffeomorphism**: an invertible function that maps the U-manifold to the V-manifold, such that both T and its inverse T^{-1} are continuously differentiable. This means that every point in the U-manifold is smoothly "pushed" into a single point in the V-manifold, and vice-versa. As a result, the density $p_V(V)$ at a point V is equal to the density of the corresponding point in the U-manifold, divided by the change of volume $|\det J_T(U;\phi)|$ induced by the transformation T. By contracting and expanding the space $\mathbb{R}^D$, T can mold the simple distribution $\mathcal{N}(\vec{0}, \vec{1})$ into a possibly very complex distribution, such as the distribution of human faces (Durk P Kingma and Dhariwal, 2018)!

The novelty of NFs does not come from new theory, but from the class of transformations T. A first insight is the development of efficient transforms T whose change of volume can be easily computed. We give such an example of transformation in Section 4.1.2. A second insight comes from the **composability** of transforms:

$$\begin{aligned} T &= T_N \circ \ldots \circ T_1 \\ T^{-1} &= T_1^{-1} \circ \ldots T_N^{-1} \\ \det J_T(U) &= \det J_{T_N}(T_{N-1} \circ \ldots \circ T_1(U)) \times \ldots \times \det J_{T_1}(U) \end{aligned} \tag{4.3}$$

Stacking even moderately expressive elementary transforms T_i can yield a very expressive transform T, whose change of volume can be easily computed from the elementary changes of volume. This same principle makes the expressivity of neural networks, which can approximate nearly any function by stacking multiple simple layers (Bishop et al., 1995). This composability is the origin of the term "flow": a simple distribution progressively "flows" into a complex one through successive transformations —as illustrated in Figure 4.1. Similar to fluid mechanics, the streamlines induced by T cannot cross, ensuring its invertibility.

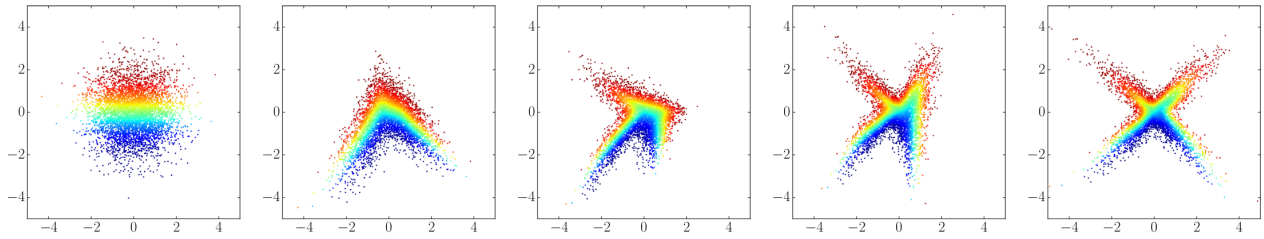

Fig. 4.1.: Illustration of a NF The simple Gaussian distribution (left) flows into a complex distribution (right) through 4 successive steps. *Figure adapted from Papamakarios, Nalisnick, et al. (2019).*

4.1.2 An example NF: the masked autoregressive flow (MAF)

This section introduces masked autoregressive flows (MAFs), a family of NF we use during this thesis (Papamakarios, Pavlakou, et al., 2017). It also exemplifies the design principles that have driven some of the recent fast development of NFs.

Following the taxonomy from Papamakarios, Nalisnick, et al. (2019), MAFs are **finite-composition, affine autoregressive flows**.

The term **autoregressive** relates to the possible decomposition of *any* density as the product of conditional densities:

$$V = \begin{bmatrix} V_1 \\ \vdots \\ V_D \end{bmatrix} \qquad p_V(V) = \prod_{i=1}^{D} p_V(V_i | V_{<i}) \tag{4.4}$$

where $V_{<i}$ denotes the values for all the dimensions of V up to i. Autoregressive flows exploit this insight to produce expressive invertible transformations:

$$\begin{aligned} V &= T(U) \\ &= \begin{bmatrix} \tau(U_1; \emptyset) \\ \tau(U_2; h_2(U_1)) \\ \vdots \\ \tau(U_D; h_D(U_1, ..., U_{D-1})) \end{bmatrix} \end{aligned} \tag{4.5}$$

Where τ denotes a unary invertible transform, whose parameters h_i at dimension i are regressed from the values of U up to i. This autoregressive pattern facilitates the inversion of the transformation T:

- to go from U to V, all the τ-transformations parameters h_i can be computed in parallel;
- to go from V to U, one can iteratively compute $U_1 = \tau^{-1}(V_1; \emptyset)$, then $U_2 = \tau^{-1}(V_2; h_2(U_1))$, up to $U_D = \tau^{-1}(V_D; h_D(U_1, ..., U_{D-1}))$.

In addition, the autoregressive structure facilitates the computation of the change of volume of T:

$$|\det J_T(U)| = \prod_{i=1}^{D} |\frac{\partial \tau}{\partial U_i}(U_i; h_i(U_{<i}))| \tag{4.6}$$

The term **affine** relates to the implementation of the unary transform τ:

$$\tau(U_i; h_i(U_{<i})) = \alpha_i(U_{<i}) \times U_i + \beta_i(U_{<i}) \quad \text{where} \quad h_i = (\alpha_i, \beta_i) \tag{4.7}$$

Importantly, because the parameters (α, β) are regressed from U, the transformation T is *not* an affine transform —it is only "locally" affine. Furthermore, (α, β) can be complex functions of U, typically implemented using neural networks. Using affine transformation further facilitates the computation of the change of volume:

$$|\det J_T(U)| = \prod_{i=1}^{D} |\alpha_i(U_{<i})| \tag{4.8}$$

Finally, the term **masked** comes from the implementation of the regressors for the parameters (α, β). In the MAF architecture, those are computed using a single pass of a multi-layer perceptron (MLP). To ensure an autoregressive structure, a triangular binary mask is applied over the weights of this network. The weights ϕ of the transform T are the weights of this neural network.

The architecture of MAFs gives several insights about NFs's computational efficiency:

- First, the time complexity of NFs depends on the "direction" in which they are applied. To go from U to V, all the parameters h_i can be computed using a single pass of a neural network, in a time $\mathcal{O}(1)$. However, to go from V to U, the h_i are computed recursively, meaning that the computation time becomes $\mathcal{O}(D)$. Attention thus has to be paid when using autoregressive flows in practice.
- Second, the parameterization of the flow scales poorly with the dimensionality D of the manifold it is applied on. In the case of the MAF, this parameterization corresponds to the weights of an MLP, which scale in $\mathcal{O}(D^2)$.

In Section 5.1.1, we explain how those tradeoffs negatively affect the scale-up of NFs to large dimensions.

4.1.3 Using normalizing flows (NFs) as conditional density approximators

In Section 4.1.1, we showed how NFs, by stacking multiple transformations, yield powerful density approximators. In the context of inference, we encounter *conditional* distributions, that depend on the realization of other RVs. For instance, in the context of *amortized* inference —described in Section 3.3.2— we condition the distribution of the parameters Θ based on the value of the signal X. In HBMs, distributions are conditioned by the realization of parent RVs in the graph.

NFs are easily amenable to model such conditional distributions. To condition the distribution of some reparameterized RV V on the value of some other RV W, it suffices to parameterize the transform T using the value of W:

$$\begin{aligned} &V \sim P_V(V|W) \\ &V = T(U; W, \phi) \qquad \text{where} \quad U \sim \mathcal{N}(\vec{0}, \vec{1}) \end{aligned} \tag{4.9}$$

To parameterize T based on W, we typically pass the value of W to an encoder. This encoder can be a neural network, with inductive biases adapted to W's geometry, such as a convolutional neural network (CNN) (Bishop et al., 1995). This thesis refers to the output of this encoder as an *encoding*. In Chapter 6, we exploit the **couple encoding/conditional density approximator** to design scalable variational families.

As an example, conditioning in a MAF can be done at the level of the masked MLP —taking the notations from Section 4.1.2:

$$\tau(U_i; h_i(U_{<i})) = \alpha_i(U_{<i}, W) \times U_i + \beta_i(U_{<i}, W) \tag{4.10}$$

where the value of W is simply concatenated on top of the value of U at the input of the MLP.

4.2 Inference using distribution surrogates

This section describes the use of NFs in the context of likelihood-free inference. We start with a general definition of simulation-based inference (SBI). We then move on to the type of SBI we focus on: inference using density surrogates.

On a technical note, implementations of the algorithms listed in this section are provided by the Python `sbi` library (Tejero-Cantero et al., 2020).

4.2.1 Simulation-based inference (SBI): a general definition

In Section 3.1, we introduced statistical inference: determining the parameters Θ susceptible to generating the signal X through a model. This statistical model defines a joint distribution $P(X, \Theta)$. For traditional inference methods to work, the joint density p of this model needs to be evaluated. For instance in MCMC, the joint density is used to compute posterior ratios —see Section 3.2. Conversely, in VI, the joint density is evaluated as part of computing the ELBO —see Section 3.3. As the model P is typically specified by the experimenter, its density is in general readily available. But what happens if it is not?

A typical scenario where the density is not available is in the presence of a **simulator**. Constructed based on decades of research, those simulators take as input parameters Θ, and output a realistic signal X (for instance in physics, genetics, and dMRI Justin et al., 2019; Beaumont et al., 2002; Ianuş et al., 2017). Producing this synthetic signal may involve non-deterministic control flow, non-differentiable operations, and even external pieces of software. This effectively makes the simulator a black box which only *implicitly* defines a joint distribution, from which it is possible to *sample*, but whose density we cannot *evaluate*. Thus, though ubiquitous in science, those simulators are not suited for traditional inference (Cranmer et al., 2020). The field of SBI aims at performing inference in this challenging context.

Following the taxonomy from Cranmer et al. (2020), SBI methods can be broadly decomposed into two groups.

The first type of method integrates the simulator directly into the inference loop. This includes methods that iteratively draw samples from the simulator and compare those to the observed signal, progressively refining confidence intervals (Rubin, 1984). Other methods leverage probabilistic and differentiable programming to (partially) open the black box of the simulator and improve their sample efficiency (Baydin et al., 2019; Dillon et al., 2017; Bingham et al., 2019). Tracing back the

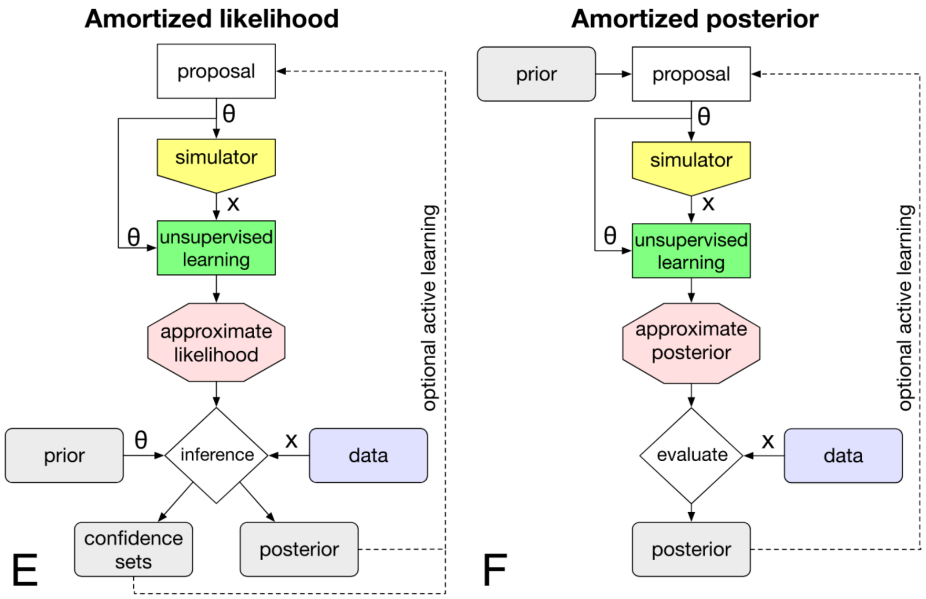

Fig. 4.2.: Flowcharts for neural likelihood estimation (NLE) (right) and neural posterior estimation (NPE) (right) Both methods rely on a synthetic dataset to learn a surrogate distribution. Active learning corresponds to the *sequential* methods described in Section 4.2.2. *Figure adapted from Cranmer et al. (2020).*

observed signal progressively to the latent parameters, those methods provide deeper insights into the mechanisms leading to the observed data. However, those methods cannot be amortized, and are likely to suffer in high-dimensional scenarios.

This thesis thus focuses on the second type of SBI methods: the usage of **surrogate models**. Using samples (Θ, X) from the simulator, neural networks can approximate either the likelihood $P(X|\Theta)$, the posterior $P(\Theta|X)$, or likelihood ratios $P(X|\Theta_1)/P(X|\Theta_2)$ (Papamakarios, Sterratt, et al., 2019; Lueckmann et al., 2017; Mohamed and Lakshminarayanan, 2016). Once trained, those surrogates can be plugged into traditional inference methods. As an example, likelihood ratios can be used as part of MCMC. The next section focuses on the other two possibilities: the learning of a surrogate posterior or likelihood distribution.

4.2.2 Neural posterior estimation (NPE) and neural likelihood estimation (NLE)

This section puts together elements presented in Section 3.3 and Section 4.1 to yield neural posterior estimation (NPE) (Papamakarios and Murray, 2016; Greenberg et al., 2019) and neural likelihood estimation (NLE) (Papamakarios, Sterratt, et al., 2019). Both methods use a synthetic dataset of couples $\{(X^i, \Theta^i)\}$, where X^i is the

output of the simulator run with parameters Θ^i. Similar to the supervised machine learning setup, methods use either X or Θ as an input, and maximize a probability over the other term:

- in the case of NPE, the density over Θ is maximized, conditioned by X. This amounts to minimizing the amortized f-KL between the variational distribution $Q(\Theta|X;\phi)$ and the posterior $P(\Theta|X)$. This is the setup we described in Section 3.3.3;
- in the case of NLE, the density over X is maximized, conditioned by Θ. Papamakarios, Sterratt, et al. (2019) show this amounts to minimizing the amortized KL between the variational distribution $Q(X|\Theta;\phi)$ and the likelihood $P(X|\Theta)$:

$$\mathrm{Div}_{\mathrm{KL}}^{\mathrm{amo}}(Q(X|\Theta)||P(X|\Theta)) \propto \mathbb{E}_{P(X,\Theta)}\left[-\log q(X|\Theta)\right] \tag{4.11}$$

Those methods rely on expressive NFs as variational distributions Q. As a result, provided sufficiently many couples $\{(X^i,\Theta^i)\}$, NPE and NLE can approximate perfectly the distribution implicitly defined by the simulator. Both methods are illustrated in Figure 4.2.

Comparing both methods, NLE learns the likelihood and thus is less dependent on the choice of the prior $P(\Theta)$. This allows for more flexibility, for instance, to plug NLE into a frequentist inference, or to change the prior during inference (Cranmer et al., 2020). As a drawback, NLE relies on a companion inference method —such as MCMC or VI— to derive the posterior $P(\Theta|X)$ using the learned likelihood. In contrast, NPE directly learns the posterior. As further detailed in Section 5.1.1, NFs struggle when applied over a large-dimensional space. In some cases, the dimensionality of Θ can be much smaller than the dimensionality of X, and NPE should be preferred (and vice-versa). More subtly, Papamakarios, Sterratt, et al. (2019) underline that the likelihood $P(X|\Theta)$ can sometimes be a much "simpler" distribution than the posterior $P(\Theta|X)$. In this case, NLE should be preferred... at the cost of moving the burden of inference to the companion inference method. All in all, experimenters should try out which method yields better results on their inference problem, as permitted by the `sbi` library (Tejero-Cantero et al., 2020).

A major feature of **surrogate-based SBI methods** is that they **are by default amortized**. As explained in Section 3.3.2, amortization complicates inference. NPE and NLE can thus suffer from poor sample efficiency. To circumvent this issue, **sequential** variants have been developed (Papamakarios, Sterratt, et al., 2019; Papamakarios and Murray, 2016; Greenberg et al., 2019). Sequential methods specialize over the

small part of the space containing a given signal $\mathbf{X}_0$ and the associated posterior $P(\Theta|X = \mathbf{X}_0)$. A fully amortized approximate proposal distribution is first learned —either a surrogate likelihood or posterior. The latter is used to get a rough estimate of the posterior $P(\Theta|X = \mathbf{X}_0)$. New samples Θ^i are sampled from this approximate posterior, and fed to the simulator to yield the corresponding X^i. The new "posterior predictive" X^i are more similar to the true observed signal $\mathbf{X}_0$ than if the Θ^i were sampled from the prior. It is thus more relevant to train over those samples for inference over $\mathbf{X}_0$. A more specialized proposal distribution can then be learned using this new augmented dataset, and so on. After a few simulation rounds, the rough proposal distribution is refined into a good *local* approximation of its target —either the likelihood or the posterior. Sequential methods are an instance of **active learning** (Cranmer et al., 2020). They constitute a trade-off: the approximation for the signal of interest is improved at the cost of true amortization (over the entirety of the Θ-space).

4.3 "VI is biased": revisiting a statistician's idiom

Through this chapter, we have reviewed some recent advances developed by the SBI community. This short opinion section underlines the potential of those techniques in the context of VI.

As oftentimes heard during conferences and workshops, a main drawback of VI is that it is *biased*. This argument is used for instance to put forward the comparative advantages of MCMC. Indeed, MCMC has asymptotic guarantees: provided sufficiently long runs, MCMC ensures to sample from the true posterior —see Section 3.2. VI does *not* benefit from such guarantees and is in fact ***doubly* biased**. The first bias comes from the **approximation gap** described in Section 3.3.1. After convergence, VI yields the best approximation to the posterior *inside the variational family*. In practice, misspecifying the variational family can be disastrous. This puts onto the experimenter the additional burden of choosing an appropriate variational family —as already hinted to in Section 3.4. A second bias comes from the prominent usage of the **r-KL** as training loss. Even if the variational family *could* capture the true posterior, training using the r-KL can lead to mode collapse and variance underestimation —as explained in Section 3.4. As a result, even using an expressive family does not guarantee convergence to the true posterior.

The techniques described in this section offer a counterpoint to this bias issue.

First, NFs asymptotically nullify the approximation gap. As an example, Durk P Kingma and Dhariwal (2018) have applied the *Glow* NF architecture to model distributions as complex as human faces. Used as variational distributions, NFs could approximate virtually any posterior. Second, as described in Section 3.4, training using the f-KL prevents degenerate behaviors such as mode collapse or variance under-estimation. Leveraging those techniques, traditional drawbacks of VI could thus be completely circumvented.

From a methodological point of view, this thesis attempts to **leverage the advantages of NFs and the f-KL in the context of large-scale VI**. What is the applicability of SBI methods in the explicit-likelihood context? How to use those techniques at scale?

Large-scale inference

5

This chapter dives deeper into large-scale inference, a hurdle we introduced in Section 3.4. We first underline the shortcomings in a high dimension of the promising methods introduced in Chapter 4. We expand on the necessity to exploit the causal structure of the model for efficient inference. Finally, we introduce methods in VI that allow exploiting this structure.

5.1 Shortcomings of SBI in high dimensions

Chapter 4 presented two promising techniques in the context of inference. The first is NFs. Much like neural networks are universal function approximators (Bishop et al., 1995), NFs are universal density approximators. The second technique is the use of the f-KL as training loss. Contrary to the r-KL, the f-KL does not incur a biased approximation of the target distribution. This section reviews the applicability of both techniques in high dimensions.

5.1.1 Shortcomings of NFs: the necessity to exploit structure in inference

In Section 4.1.2, we presented the MAF as an example of NF architecture. Consider the usage of a MAF as a variational distribution Q to fit the posterior over a space of dimension $D \gg 1$. In terms of time complexity, training using the r-KL necessitates both sampling from Q and evaluating its density. As shown in Section 4.1.2, this operation has a $\mathcal{O}(D)$ complexity. In terms of parameterization, the number of weights scales with $\mathcal{O}(D^2)$.

Note: training using the f-KL would only have a $\mathcal{O}(1)$ time complexity, but the f-KL has shortcomings of its own described in the next section.

As presented in Section 1.4, applications in Neuroscience can reach the dimensionality of the thousands or millions. Applying NFs over this dimensionality is therefore out of the question.

Making a parallel with machine learning, this result is not a surprise. A key ingredient in machine learning's success is **exploiting data's geometry through adapted inductive biases** (Bishop et al., 1995). For instance, applying a MLP over the entirety of a hundred thousand-pixel image is computationally infeasible. In contrast, it is much more efficient to exploit an image's translation invariance through a parsimoniously parameterized CNN. Similarly, applying a NF over the entirety of an image is wasteful, and successful applications of NFs over images have resorted to convolutional structure, or custom multi-scale architectures (Durk P Kingma and Dhariwal, 2018; Dinh et al., 2016). The architecture of NFs is thus "too general" to tackle the large-scale effectively.

How to meaningfully reduce the generality of NFs?

The generality of NFs is twofold. The first generality lies in the arbitrary *shapes* of the densities that can be modeled. This includes multiple modes, heavy tails, or complex geometries such as cross-shaped or banana-shaped distributions (Papamakarios, Nalisnick, et al., 2019). This is an important property to conserve. The second generality is in the arbitrary conditional *dependencies* modeled across the individual RVs that compose the D-space. As detailed in Section 4.1.2, autoregressive flows model arbitrary distributions as a succession of conditional distributions. Roughly, using permutations of the D-space, NFs "test out" all the possible conditional dependencies across dimensions in D (Papamakarios, Nalisnick, et al., 2019). Yet, in HBMs, a lot of individual RVs are (conditionally) independent, and it is wasteful to consider the corresponding dependencies as part of inference. Said differently, **the causal structure of HBMs is a strong geometry that should be reflected in inductive biases**. This same insight inspired exact inference methods such as variable elimination —as presented in Section 3.2.

Note: The distinction between a distribution's complex shape and the dependencies it entails is simplistic. Often, the difference between the two is not clear-cut. For instance, a 2D banana distribution's complex shape has more to do with a strong dependency between the two 1D RVs that compose it. The architectures proposed in Chapter 6 rather revolve around separating RVs into different "blocks", modeling the full dependencies *inside* blocks, but only a few *across* blocks.

The graphical causal structure can be injected in a NF via masking, as done by Wehenkel and Louppe (2021) or Weilbach et al. (2020). In Chapter 6, we aggregate flows in a graphical structure, which is an equivalent formulation, but opens the possibility for stochastic training, as detailed in Section 5.2.2. In essence, **we decouple NFs ability to model arbitrary distribution shapes from their ability to model arbitrary conditional dependencies**, keeping the former, and adapting the latter to the large scale.

5.1.2 Shortcomings of the f-KL: amortization and ease of optimization

Sections 3.3.3 and 4.2.2 already expanded upon the first shortcoming of the f-KL: the need for amortization. The fact that amortization complicates inference is a major issue of the f-KL, especially in the already challenging context of large-scale inference.

Another issue of the f-KL is the harder optimization it can entail. This thesis generally puts forward the comparative reliability of the f-KL compared to the r-KL. Yet, considering *both* divergences in the amortized setup, the f-KL is not always unequivocally more desirable than the r-KL. To our surprise, though the formulation of both amortized divergences is dual, we encountered in this thesis examples where their computational efficiency wildly differed (see Section 6.4). Our interpretation is that the "unforgiving" nature of the f-KL —described in Section 3.4— can in practice complicate optimization by forcing the variational family to consider the entirety of the support of its target at once. In contrast, the more locally exploratory behavior of the r-KL can lead to a smoother optimization. Another failure mode of the f-KL is mixture models (also mentioned in Section 6.4). Considering all the possible permutations of the components of a mixture into multi-modal posteriors is in practice wasteful —as per the so-called *label switching* problem (Jasra et al., 2005). In this context, mode collapse could be considered a desirable property.

The f-KL should thus not be applied indiscriminately over the entirety of the Θ parameter space at once. In Chapters 8 and 9, we rather leverage the f-KL over key sets of low-dimensional parameters, to disseminate its advantages into a r-KL training.

5.2 Leveraging causal structure in VI

In Section 5.1.1, we described how NFs, though powerful density approximators, could fail when applied over high-dimensional structured HBMs. In this section, we review how the VI community tackles inference over such models. We consider separately the treatment of conditional dependencies and the training over subparts of the full model. Throughout this section, we slightly redefine terms from the literature into a taxonomy we feel more apt to describe the trade-offs available to experimenters.

5.2.1 Conditional dependencies modeled in the variational family

This section uses as a running example the salmon example described in Section 2.1.2. We lay out the graph corresponding to Figure 2.3 in Figure 5.1 (left). On the right of the figure, from top to bottom, we detail in chronological order the dependency schemes used in VI.

Mean-field dependency scheme (MF) The simplest dependency scheme is to consider every RV in the graph independently. This corresponds to the graph on the top right in Figure 5.1. The MF was originally used to facilitate pen and paper inference, allowing for dedicated optimization routines (David M. Blei et al., 2017). In this thesis, we use a "**blockwise**" definition for the MF. MF denotes the independence *across* RVs (the nodes in the graph), but not necessarily *inside* RVs (inside a given node). This means that we still model the conditional dependencies across the different dimensions that constitute a given RV. Though computationally attractive, the MF increases the approximation gap, and can result in biased inference —as described in Section 3.3.1.

Structured VI Starting from the MF, experimenters can choose to model arbitrary dependencies across RVs in the graph. As an example, the dependencies from the model P can be replicated in the variational family $\mathcal{Q}$, as in the middle graph in Figure 5.1. This design was originally proposed by Matthew D Hoffman and David M Blei (2015), under the name *structured* VI (Ambrogioni, Lin, et al., 2021). In this thesis, **structured VI refers to the injection of a meaningful causal structure in the variational family**, not necessarily the model's structure. As an example, Ambrogioni, Silvestri, et al. (2021) add to the model's causal structure a "backward" dependency scheme, reversing the model's connections. Another example is the work

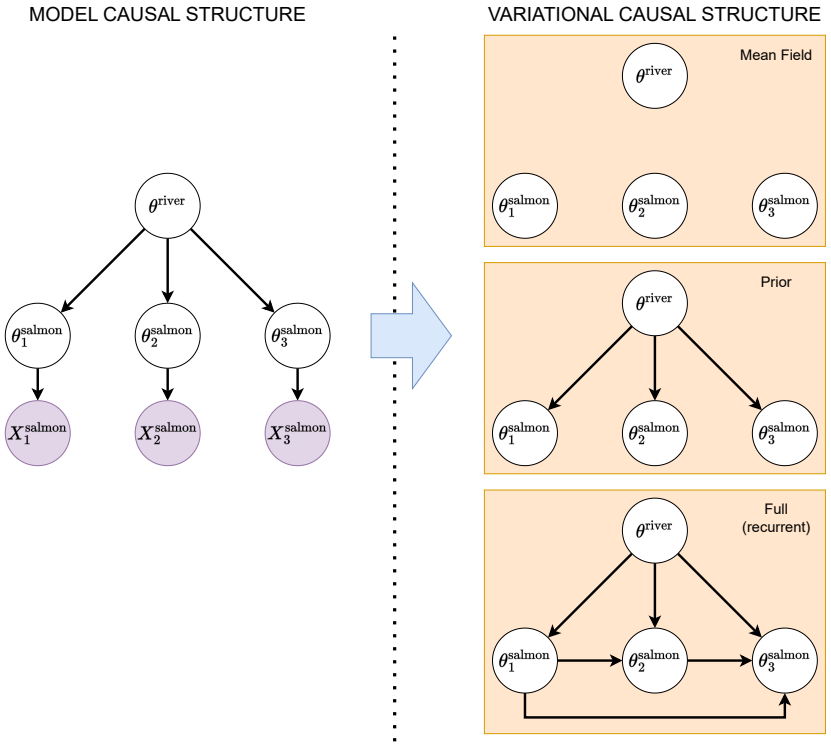

Fig. 5.1.: Model and variational dependency structure On the left, we represent the ground graph corresponding to the salmon example in Section 2.1.2. On the right, we show 3 possible dependency schemes modeled in the variational family. From top to bottom, no dependencies (mean-field); the same dependencies as the prior; all the possible dependencies. The last option is the most expressive, but the most computationally costly.

from Webb et al. (2018), who model only the necessary and sufficient dependencies. As a rule of thumb, modeling more dependencies in $\mathcal{Q}$ generally increases its expressivity, at the cost of a reduced computational efficiency.

Full dependencies At the other end of the spectrum from the MF, the experimenter can choose to model *every* possible dependency across RVs. In practice, this can be achieved via an autoregressive structure, as described in Section 4.1.2 and illustrated in the bottom graph from Figure 5.1. This is the design principle of NFs, as detailed in Section 5.1.1. Modeling all the possible dependencies ensures a minimal approximation gap. But this generality can seriously affect the computational efficiency of $\mathcal{Q}$, resulting in worse results in practice (Ambrogioni, Lin, et al., 2021; Ambrogioni, Silvestri, et al., 2021).

Choosing which dependencies to model in $\mathcal{Q}$, and deriving those automatically from the model's graph, is an integral part of automatic variational inference (AVI) —as described in Section 3.4. The other main design choice is the shapes of the parametric densities for each RV in the graph. Introducing a trade-off between expressivity and computational efficiency, **structured VI is a key ingredient to scale inference up to large dimensions**.

5.2.2 Training over a subsample of the model's graph

In Section 5.2.1 we showed how the graphical structure could be leveraged to adapt the variational family's parameterization. In some cases, however, even a parsimoniously parameterized family cannot be trained over the full model due to its sheer size. In Neuroimaging examples, evaluating gradients to update the weights of a family spanning over hundreds of thousands of voxels requires massive memory and compute. To control the computational resources necessary for inference, one would need to train over only a *subset* of those voxels at a time. Matthew D. Hoffman, David M. Blei, et al. (2013) originally introduced this design under the name **stochastic variational inference (SVI)**.

To take a concrete example, consider the salmon example illustrated in Figure 5.1. Training using the r-KL, and using a MF variational family, we minimize:

$$\begin{aligned} -\text{ELBO}(\phi) = \mathbb{E}_{\Theta\sim Q}[\log p(\theta^{\text{river}}) - \log q(\theta^{\text{river}};\phi^{\text{river}}) \\ + \sum_{i=1}^{3} \log p(\theta_i^{\text{salmon}}|\theta^{\text{river}}) - \log q(\theta_i^{\text{salmon}};\phi_i^{\text{salmon}})] \end{aligned} \tag{5.1}$$

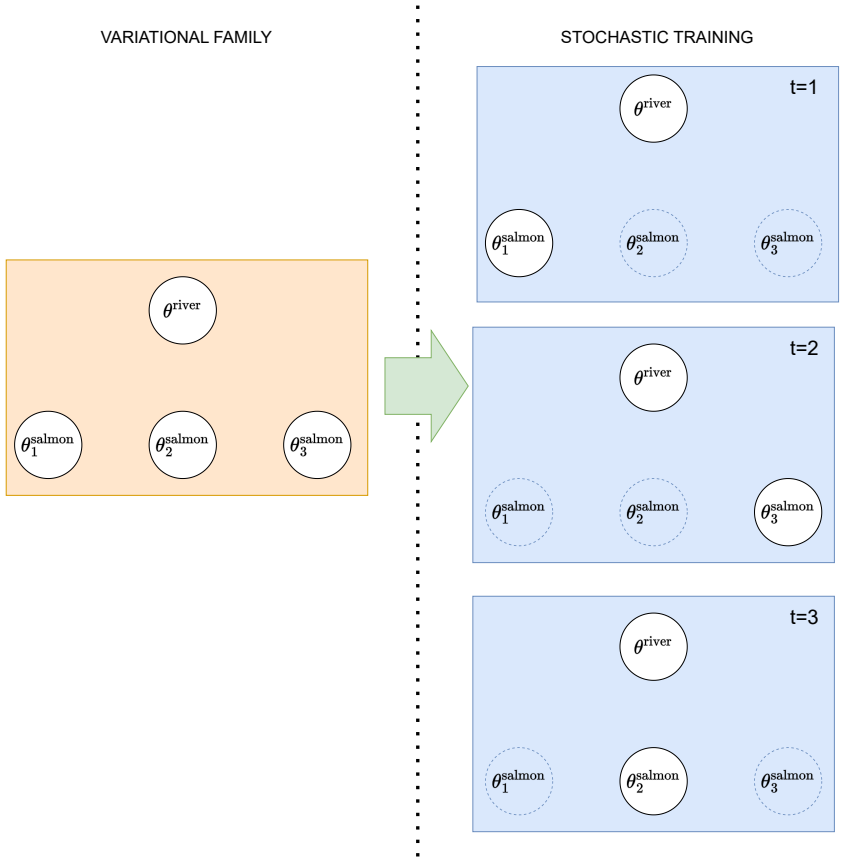

Fig. 5.2.: Stochastic variational inference (SVI): At each optimization step t, we can train over a random subset of the model's graph. As an example, training only over a single salmon in the river. We update the river mean parameters based on this salmon's observed weight. In expectation, cycling through all the salmons, this stochastic training yields the same result as observing all the salmons at once.

where we split the variational weights ϕ associated with each RV. At each optimization step t, we can randomly select a single salmon $choice[t]$, as illustrated in Figure 5.2.

Selecting a random salmon, we can minimize the *stochastic* loss (Matthew D. Hoffman, David M. Blei, et al., 2013):

$$\begin{aligned} -\,\text{ELBO}^{\text{sto}}(\phi)[t] = \mathbb{E}_{\Theta\sim Q}[\,&\log p(\theta^{\text{river}}) - \log q(\theta^{\text{river}};\phi^{\text{river}}) \\ &+3\times(\log p(\theta^{\text{salmon}}_{choice[t]}|\theta^{\text{river}}) - \log q(\theta^{\text{salmon}}_{choice[t]};\phi^{\text{salmon}}_{choice[t]}))] \end{aligned} \tag{5.2}$$

where we compute the density over only a third of the θ^{salmon} RVs, and only update the corresponding weights. This means that we can **control the amount of compute and memory required at each optimization step**.

Critically, Matthew D. Hoffman, David M. Blei, et al. (2013) show that training over random subsets of the graph yields the same result as training over the full graph. Mathematically, this amounts to showing that:

$$\mathbb{E}_{choice[t]}\left[\text{ELBO}^{\text{sto}}(\phi)[t]\right] = \text{ELBO}(\phi) \tag{5.3}$$

meaning that the expectation of the stochastic loss over random RV subsets is equal to the "full" loss. This result is not tied to the MF approximation. Matthew D Hoffman and David M Blei (2015) generalize this result to variational families replicating the prior's dependencies —as in the middle example in Figure 5.1. In Chapter 6, we apply the same idea to arbitrary plate-enriched graphs.

Stochastic training is commonplace in machine learning. As an example, in supervised learning, networks are trained over minibatches taken inside a dataset (Bishop et al., 1995). The same idea can be applied in SVI, considering several salmons at once. An important difference is that in machine learning, stochastic training is performed across i.i.d data points. In contrast, **the RVs in SVI are only *conditionally* i.i.d**. There is thus more to the result in Equation (5.3) than meets the eye. Importantly, the result from Equation (5.3) would not generalize to *any* subset of RVs in the graph. For instance, one could not "ignore" θ^{river} during a stochastic training step. Stochastic training is only possible across a model's plates —see Section 2.1.2 for a definition. This means that the model's causal structure is exploited to yield a stochastic training scheme. In essence, this is similar to deriving the variational dependencies based on the model's graph in Section 5.2.1.

Part conclusion

This part reviewed modern trends in inference. In Chapter 4, we put forward NFs and the f-KL as promising methods to circumvent VI's inherent bias. Yet, those methods are poorly suited to large-scale inference. In Chapter 5, we reviewed methods in VI to tackle this regime. The key idea is that a model's causal structure is a strong geometry that must be exploited to handle large-scale inference computationally. In Part III, we combine all these techniques to tackle the hierarchical inference problems described in Part I.

Part III

Contributions

6 Methodological contributions: expressive and scalable structured automatic VI

This chapter condenses our machine learning contributions. We leverage the methods from Part II in large-scale hierarchical problems. We voluntarily present a unified framework for both publications: ADAVI (Rouillard and Wassermann, 2022) and PAVI (Rouillard, Bris, et al., 2023). We specify where applicable the individual contributions of both papers.

We first present an architecture leveraging the expressivity of NFs into structured VI. We then present the fruitful combination of this design with SVI.

6.1 Leveraging NFs into structured VI

In Section 4.1.3, we presented NFs as powerful conditional density approximators. Specifically, we put forward the **encoding/conditional density estimator** couple. In this design:

- an expressive NF is amortized across data samples. The flow's role is to approximate a possibly complex distribution, based on an encoding of the observed signal;
- a lightweight encoding summarizes the statistics of the observed signal. This encoding is usually the low-dimensional output of an encoder applied to the possibly high-dimensional signal.

This section leverages this architecture in the context of structured VI.

Contributions in this section relate both to the automatic dual amortized variational inference (ADAVI) and plate amortized variational inference (PAVI) publications.

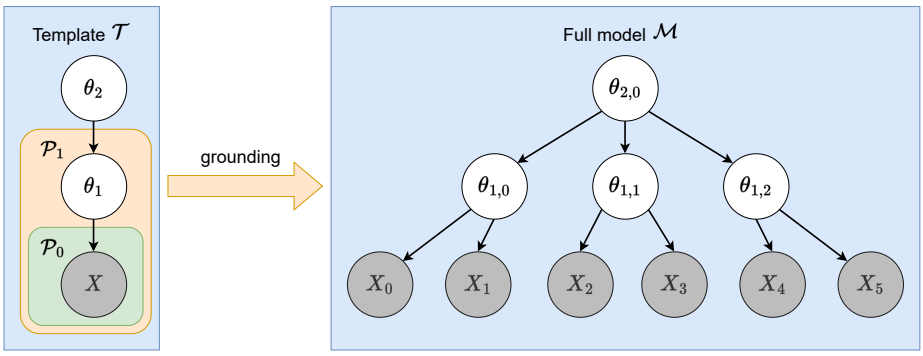

Fig. 6.1.: Generic plate-enriched HBM template The template $\mathcal{T}$ (left) can be *grounded* into the *full* model $\mathcal{M}$ (right). We aim to perform inference over $\mathcal{M}$. Yet, $\mathcal{M}$ can feature large cardinalities. As an example, instead of $\theta_{1,0}, ..., \theta_{1,2}$, $\mathcal{M}$ can feature $\theta_{1,0}, ..., \theta_{1,1000}$ —corresponding to a thousand different subjects. This can make inference over $\mathcal{M}$ computationally intractable.

6.1.1 Background notations: HBM templates

This section formalizes the notations associated with the plate-enriched DAG presented in Section 2.1.2. We denote as $\mathcal{T}$ those templates. $\mathcal{T}$ feature RV *templates* that symbolize multiple similar *ground* RVs. For instance, in Section 2.1.2, θ^{salmon} denotes a generic salmon's weight. In this section, we refer to the abstract template illustrated in Figure 6.1. $\mathcal{T}$ refers here to a generic population study, with a population parameter θ_2, several subject parameters θ_1, and several observations for each subject X. Similar to our salmon example, we will imagine that the parameters and signal represent inferred and observed weights.

We denote the template $\mathcal{T}$'s vertices, corresponding to RV templates, as X —the observed signal— and $\Theta = \{\theta_i\}_{i=1..I}$ —the inferred parameters. We denote $\mathcal{T}$'s plates as $\{\mathcal{P}_p\}_{p=0..P}$, and the plates θ_i belongs to as $\text{Plates}(\theta_i)$. I and P respectively denote the number of latent RV templates and plates in $\mathcal{T}$, which are in general not equal. In the toy example from Figure 6.1, there are two latent RV templates: θ_1 and θ_2, respectively the subject and population mean weights. $\mathcal{T}$ also features two plates $\mathcal{P}_1, \mathcal{P}_0$, which respectively denote subjects in the population and the measurements per subject. Graphically, we can see that $\text{Plates}(\theta_2) = \emptyset$, whereas $\text{Plates}(\theta_1) = \{\mathcal{P}_1\}$ and $\text{Plates}(X) = \{\mathcal{P}_0, \mathcal{P}_1\}$.

By instantiating the repeated structures symbolized by the plates $\mathcal{P}$ in $\mathcal{T}$, we obtain a heavier graph representation: the HBM $\mathcal{M}$. This instantiation is visible in Figure 6.1, where we go from the template $\mathcal{T}$ (left) to the model $\mathcal{M}$ (right). To go from one representation to the other, $\mathcal{T}$ is *grounded* into $\mathcal{M}$ given some plate cardinalities $\{\text{Card}(\mathcal{P}_p)\}_{p=0..P}$ (Koller and Friedman, 2009). $\text{Card}(\mathcal{P})$ represents the number of

elements in the plate $\mathcal{P}$, for instance the number of subjects in the study. Going from $\mathcal{T}$ to $\mathcal{M}$, a RV *template* θ_i is instantiated into multiple *ground* RVs $\{\theta_{i,n}\}_{n=0..N_i}$ with the same parametric form, where $N_i = \prod_{\mathcal{P}\in \text{Plates}(\theta_i)} \text{Card}(\mathcal{P})$. In Figure 6.1, the RV template θ_1 is grounded into the ground RVs $\theta_{1,0}, \theta_{1,1}, \theta_{1,2}$. There are as many ground RVs X_i as the product of the number of groups on the study $\text{Card}(\mathcal{P}_1)$ times the number of subjects per group $\text{Card}(\mathcal{P}_0)$.

We denote as $\pi(\theta_{i,n})$ the (potentially empty) set of parents of the RV $\theta_{i,n}$ in the ground graph corresponding to the model $\mathcal{M}$. $\pi(\theta_{i,n})$ are the RVs whose value condition the distribution of $\theta_{i,n}$. For instance, in Figure 6.1, a measured weight —the child RV— is a perturbation of the subject's weight —the parent RV. This is denoted as $\pi(X_0) = \{\theta_{1,0}\}$.

The *full* model $\mathcal{M}$ is associated with the density p. In p, the plate structure indicates that a RV template θ_i is associated to a conditional density p_i shared across all ground RV $\theta_{i,n}$:

$$\begin{aligned} \log p(\Theta, X) &= \log p(X|\Theta) + \log p(\Theta) \\ &= \sum_{n=0}^{N_X} \log p_X(x_n|\pi(x_n)) + \sum_{i=1}^{I}\sum_{n=0}^{N_i} \log p_i(\theta_{i,n}|\pi(\theta_{i,n})) \end{aligned} \tag{6.1}$$

where $\pi(\theta_{i,n})$ is the (potentially empty) set of parents of the RV $\theta_{i,n}$, which condition its distribution. We denote with a $\bullet_X$ index all variables related to the observed RVs X.

Exploiting the factorization visible in Equation (6.1), our goal is to obtain a variational distribution $Q(\Theta)$ usable to approximate the unknown posterior $P(\Theta|X)$ for the target model $\mathcal{M}$.

6.1.2 Plate amortization: leveraging the encoding/NF couple in structured VI

Full variational family Here we define an AVI scheme to derive automatically the variational distribution Q corresponding to the full model $\mathcal{M}$.

To implement Q we use trainable NFs, denoted as $\mathcal{F}$ —see Section 4.1 for a formal definition. To every ground RV $\theta_{i,n}$, we associate the learnable flow $\mathcal{F}_{i,n}$ to approximate its posterior distribution:

$$\begin{aligned} \log q(\Theta) &= \sum_{i=1}^{I} \sum_{n=0}^{N_i} \log q_{i,n}(\theta_{i,n}) \\ \theta_{i,n} &= \mathcal{F}_{i,n}(u_{i,n}) \end{aligned} \tag{6.2}$$

where $q_{i,n}$ is the push-forward density of the RV $u_{i,n}$ through the flow $\mathcal{F}_{i,n}$ —as visible in Figure 6.2. Our contributions differ in the distribution of the RV $u_{i,n}$:

- in ADAVI (Rouillard and Wassermann, 2022):

$$u_{i,n} \sim \mathcal{N}(\vec{0}, \vec{1}) \tag{6.3}$$

  This implies that ADAVI follows the blockwise MF approximation described in Section 5.2.1. Full dependencies are modeled *within* a RV $\theta_{i,n}$ (across its dimensions) but no dependencies are modeled *across* different RVs $\theta_{i,n}$.

- in PAVI (Rouillard, Bris, et al., 2023):

$$u_{i,n} \sim P_i(u_{i,n} | \pi(\theta_{i,n})) \tag{6.4}$$

  PAVI uses a more expressive structured VI dependency scheme, which replicates the prior dependencies. In particular, the flow $\mathcal{F}_{i,n}$ does not push forward a Gaussian, but a conditional prior distribution. This *cascading* scheme was first introduced by Ambrogioni, Silvestri, et al. (2021).

Plate amortization Here we introduce plate amortization: sharing the parameterization of density estimators across a model's plates. We leverage the encoding/NF couple inside the plate structure of the model. Plate amortization reduces the number of weights in a variational family as the cardinality of the inference problem augments. In Section 6.2.2, we show that plate amortization also results in faster inference.

In Section 3.3.2, we presented inference *amortization*: regressing the variational weights ϕ using an *encoder* f of the observed signal $\mathbf{X}$:

$$Q(\Theta; \phi = f(\mathbf{X}_i)) \simeq P(\Theta | X = \mathbf{X}_i) \tag{6.5}$$

We exploit the concept of amortization but apply it at a different granularity, leading to our notion of *plate amortization*.

Similar to amortizing across the different data samples $\mathbf{X}$, we amortize across the different ground RVs $\{\theta_{i,n}\}_{n=0..N_i}$ corresponding to the same RV template θ_i. Instead of casting every flow $\mathcal{F}_{i,n}$, defined in Equation (6.2), as a separate, fully-parameterized flow, we will share some parameters across the $\mathcal{F}_{i,n}$. To the template θ_i, we associate a conditional flow $\mathcal{F}_i(\,\cdot\,;\phi_i,\bullet)$ with weights ϕ_i shared across all the $\{\theta_{i,n}\}_{n=0..N_i}$. The flow $\mathcal{F}_{i,n}$ associated with a given ground RV $\theta_{i,n}$ will be an instance of this conditional flow, conditioned by an encoding $\mathbf{E}_{i,n}$:

$$\mathcal{F}_{i,n} = \mathcal{F}_i(\,\cdot\,;\phi_i,\mathbf{E}_{i,n}) \quad \text{yielding} \quad q_{i,n} = q_{i,n}(\theta_{i,n};\phi_i,\mathbf{E}_{i,n}) \tag{6.6}$$

The distributions $Q_{i,n}$ thus have 2 sets of weights, ϕ_i and $\mathbf{E}_{i,n}$, creating a parameterization trade-off. Concentrating all of $Q_{i,n}$'s parameterization into ϕ_i results in all the ground RVs $\theta_{i,n}$ having the same posterior distribution. On the contrary, concentrating all of $Q_{i,n}$'s parameterization into $\mathbf{E}_{i,n}$ allows the $\theta_{i,n}$ to have completely different posterior distributions. But in a large cardinality setting, this freedom results in a massive number of weights, proportional to the number of ground RVs times the encoding size. This double parameterization is therefore efficient when the majority of the weights of $Q_{i,n}$ is concentrated into ϕ_i.

Using NFs $\mathcal{F}_i$, the burden of approximating the correct parametric form for the posterior is placed onto ϕ_i, while the $\mathbf{E}_{i,n}$ encode lightweight summary statistics specific to each $\theta_{i,n}$. For instance, $\mathcal{F}_i$ could learn to model a Gaussian mixture distribution, while the $\mathbf{E}_{i,n}$ would encode the location and variance of each mode for each ground RV. Encodings $\mathbf{E}_{i,n}$ allow to individualize for $\theta_{i,n}$ only the strictly necessary information necessary to approximate $\theta_{i,n}$'s posterior.

The automatic derivation of the variational family $\mathcal{Q}$ from the model P is detailed in Algorithm 1. Algorithm 1 integrates different schemes to derive the encodings $\mathbf{E}_{i,n}$, that are described in Section 6.2.3 and Algorithm 2.

Summary This section defined Q, the variational distribution to approximate the full model $\mathcal{M}$'s posterior. Q features *plate amortization*, which helps maintain a tractable number of weights as the cardinality of $\mathcal{M}$ augments. The next section introduces a stochastic scheme to train Q.

Algorithm 1: ADAVI and PAVI architecture build

Input:

Model density:
$\log p(\Theta, X) = \sum_{n=0}^{N_X} \log p_X(x_n|\pi(x_n)) + \sum_{i=1}^{I}\sum_{n=0}^{N_i} \log p_i(\theta_{i,n}|\pi(\theta_{i,n}))$
Choice of variational family type: ADAVI, PAVI-F or PAVI-E

Output:

Variational family $\mathcal{Q}$

Algorithm:

Construct the conditional density approximator architecture:
for $i = 1..I$ **do**
  Construct conditional flow $\mathcal{F}_i$
  Define conditional posterior distributions $\theta_{i,n} = \mathcal{F}_i(u_{i,n}; \mathbf{E}_{i,n})$
  if *ADAVI variational family type* **then**
    Select $u_{i,n} \sim \mathcal{N}(\vec{0}, \vec{1})$
  else if *PAVI-F or PAVI-E variational family type* **then**
    Select $u_{i,n} \sim P_i(u_{i,n}|\pi(\theta_{i,n}))$
Combine the $q_{i,n}$ into the conditional variational density:
$\log q(\Theta) = \sum_{i=1}^{I}\sum_{n=0}^{N_i} \log q_{i,n}(\theta_{i,n})$
Construct the encodings $\mathbf{E}_{i,n}$:
if *ADAVI or PAVI-E variational family type* **then**
  Following Section 6.2.3 and Algorithm 2, construct a hierarchical encoder f
  Define the encodings $\mathbf{E} = f(\mathbf{X})$
  The variational family $\mathcal{Q}$ is the combination of the conditional density q and the encoder f
else if *PAVI-F variational family type* **then**
  Consruct full encoding arrays $\{\mathbf{E}_i = [\mathbf{E}_{i,n}]_{n=0..N_i}\}_{i=1..I}$ following Section 6.2.3 and Algorithm 2
  The variational family $\mathcal{Q}$ is the combination of the conditional density q and the full encoding arrays $\{\mathbf{E}_i\}_{i=1..I}$

6.2 Combining plate amortization with SVI to speed up inference

Section 6.1.2 combined NFs with structured VI, the first tool from VI to tackle large-scale HBMs. In this section, we leverage the second tool presented in Chapter 5: SVI. This addition is the main increment from ADAVI to PAVI. Interestingly, combining shared parameterization and stochastic training does not simply result in adding the advantages of both, but also speeds up inference. Hence, PAVI is not simply a stochastically-trained ADAVI, but rather the non-trivial synergies emerging from that design.

Contributions in this section relate to the PAVI publication (Rouillard, Bris, et al., 2023).

Note: This section features multiple **Implementation details** paragraphs. Those paragraphs can be skipped: they are not required to understand the rest of the thesis.

6.2.1 Generic structured stochastic training scheme

Our goal is to train the variational distribution $Q(\Theta)$, defined in Equation (6.2), to approximate the posterior $P(\Theta|X)$. Q corresponds to the *full* model $\mathcal{M}$. $\mathcal{M}$ typically features large plate cardinalities $\mathrm{Card}(\mathcal{P})$ —with multiple subjects and measures per subject. As an example, in Chapter 7 we study a population of 1,000 subjects, each associated with 2 measurement sessions across 60,000 vertices. $\mathcal{M}$ thus features too many ground RVs, making it computationally intractable We will therefore train Q stochastically, over smaller subsets of RVs at a time. In this section, we interpret stochastic training as the training over a *reduced* model $\mathcal{M}^{\mathrm{r}}$.

Instead of inferring directly over $\mathcal{M}$, we train over a smaller replica of $\mathcal{M}$. To this end, we instantiate the template $\mathcal{T}$ into a second HBM $\mathcal{M}^{\mathrm{r}}$, the *reduced* model, of tractable plate cardinalities $\mathrm{Card}^{\mathrm{r}}(\mathcal{P}) \ll \mathrm{Card}(\mathcal{P})$. $\mathcal{M}^{\mathrm{r}}$ has the same template as $\mathcal{M}$, meaning the same dependency structure and the same parametric form for its distributions. The only difference lies in $\mathcal{M}^{\mathrm{r}}$'s smaller cardinalities, resulting in fewer ground RVs, as visible in Figure 6.2.

At each optimization step t, we randomly choose inside $\mathcal{M}$ paths of reduced cardinality, as visible in Figure 6.2. Selecting paths is equivalent to selecting from X a subset $X^{\mathrm{r}}[t]$, and from Θ a subset $\Theta^{\mathrm{r}}[t]$. For a given θ_i, we denote as $\mathcal{B}_i[t]$ the batch

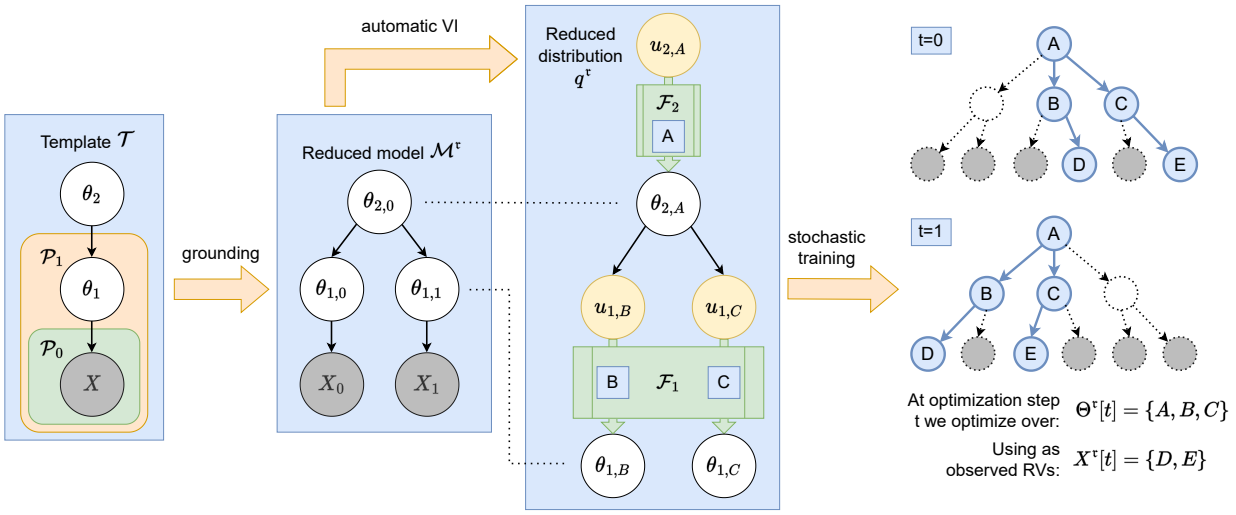

Fig. 6.2.: Structured stochastic training The full model $\mathcal{M}$ features large plate cardinalities. This makes inference over $\mathcal{M}$ computationally intractable. To circumvent this issue, we train over $\mathcal{M}$ stochastically. To this end, we instantiate $\mathcal{M}$'s template (first item) into a smaller replica: the reduced model $\mathcal{M}^{\mathfrak{r}}$ (second item). Following the AVI framework, we derive the reduced distribution $Q^{\mathfrak{r}}$ (third item) directly from $\mathcal{M}^{\mathfrak{r}}$. The reduced distribution $Q^{\mathfrak{r}}$ features 2 conditional normalizing flows $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively associated to the RV templates θ_1 and θ_2. During the stochastic training (fourth item), $Q^{\mathfrak{r}}$ is instantiated over different branchings of the full model $\mathcal{M}$ —highlighted in blue. The branchings have $\mathcal{M}^{\mathfrak{r}}$'s cardinalities and change at each stochastic training step t. The branching determines the encodings $\mathbf{E}$ conditioning the flows $\mathcal{F}$ —as symbolized by the letters A, B, C— and the observed data slice —as symbolized by the letters D, E.

of selected ground RVs, of size $N_i^{\mathfrak{r}}$. $\mathcal{B}_X[t]$ equivalently denotes the batch of selected observed RVs, of size $N_X^{\mathfrak{r}}$. The reduced model $\mathcal{M}^{\mathfrak{r}}$ is associated to the density over the observed RVs $X^{\mathfrak{r}}[t]$ and latent RVs $\Theta^{\mathfrak{r}}[t]$:

$$\log p^{\mathfrak{r}}(X^{\mathfrak{r}}[t], \Theta^{\mathfrak{r}}[t]) = \frac{N_X}{N_X^{\mathfrak{r}}} \sum_{n \in \mathcal{B}_X[t]} \log p_X(x_n | \pi(x_n)) + \sum_{i=1}^{I} \frac{N_i}{N_i^{\mathfrak{r}}} \sum_{n \in \mathcal{B}_i[t]} \log p_i(\theta_{i,n} | \pi(\theta_{i,n})) \tag{6.7}$$

where the factor $N/N^{\mathfrak{r}}$ emulates the observation of as many ground RVs as in $\mathcal{M}$ by repeating the RVs from $\mathcal{M}^{\mathfrak{r}}$ (Matthew D. Hoffman, David M. Blei, et al., 2013).

We apply the same reduction to the variational distribution Q:

$$\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t]) = \sum_{i=1}^{I} \frac{N_i}{N_i^{\mathfrak{r}}} \sum_{n \in \mathcal{B}_i[t]} \log q_{i,n}(\theta_{i,n} | \pi(\theta_{i,n})) \tag{6.8}$$

and to the ELBO loss used at step t:

$$\mathrm{ELBO}^{\mathfrak{r}}[t] = \mathbb{E}_{\Theta^{\mathfrak{r}} \sim Q^{\mathfrak{r}}} \left[\log p^{\mathfrak{r}}(X^{\mathfrak{r}}[t], \Theta^{\mathfrak{r}}[t]) - \log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t]) \right] \tag{6.9}$$

This scheme can be viewed as the instantiation of $\mathcal{M}^{\mathfrak{r}}$ over batches of $\mathcal{M}$'s ground RVs. This training is analogous to SVI (Matthew D. Hoffman, David M. Blei, et al., 2013) —presented in Section 5.2.2— generalized with multiple hierarchies, dependencies in the posterior, and mini-batches of RVs.

Implementation details: plate branchings In practice, we cannot randomly select arbitrary batches of ground RVs $\mathcal{B}_i[t]$ for the RV templates θ_i. Those batches have to be coherent with one another: they have to respect the conditional dependencies of the original model $\mathcal{M}$. As an example, if a ground RV is selected as part of $\mathcal{M}^{\mathfrak{r}}$, then its parent RVs needs to be selected as well. To ensure this, during the stochastic training we do not sample RVs directly but plates:

1. For every plate $\mathcal{P}_p$, we sample without replacement $\mathrm{Card}^{\mathfrak{r}}(\mathcal{P}_p)$ indices amongst the $\mathrm{Card}(\mathcal{P}_p)$ possible indices.

2. Then, for every RV template θ_i, we select the ground RVs $\theta_{i,n}$ corresponding to the sampled indices for the plates $\mathrm{Plates}(\theta_i)$.

3. The selected ground RVs $\theta_{i,n}$ constitute the set $\Theta^{\mathfrak{r}}[t]$ of parameters appearing in Equation (6.8). The same procedure yields the observed RV subset $X^{\mathfrak{r}}[t]$ and the data slice $\mathbf{X}^{\mathfrak{r}}[t]$.

For instance, in the toy example from Figure 6.2, X_2 will be chosen if and only if the index 1 is selected as part of sub-sampling $\mathcal{P}_1$ and the index 0 is selected as part of sub-sampling $\mathcal{P}_0$. Less formally, this is equivalent to going *middle*, then *left* in the full graph representing $\mathcal{M}$. This stochastic choice is illustrated in Figure 6.2 at $t = 1$ where X_2 corresponds to the node E.

6.2.2 Stochastic training and shared learning

Here we detail how our shared parameterization, detailed in Section 6.1.2, combined with our stochastic training scheme, results in faster inference.

In SVI (Matthew D. Hoffman, David M. Blei, et al., 2013), every $\theta_{i,n}$ corresponding to the same template θ_i is associated with individual weights. Those weights are trained only when the algorithm visits $\theta_{i,n}$, that is to say, at step t when $n \in \mathcal{B}_i[t]$. As plates become larger, this event becomes rare. If $\theta_{i,n}$ is furthermore associated with a highly-parameterized density estimator —such as a NF— many optimization steps are required for $q_{i,n}$ to converge. The combination of those two items leads to slow training and makes inference impractical in contexts such as Neuroimaging, which can feature millions of RVs. With plate amortization, we aim to unlock inference in those large regimes by reducing the training time.

Instead of treating the ground RVs $\theta_{i,n}$ independently, we share the learning across plates. Due to the problem's plate structure, we consider the inference over the $\theta_{i,n}$ as different instances of a common density estimation task. In PAVI, a large part of the parameterization of the estimators $q_{i,n}(\theta_{i,n}; \phi_i, \mathbf{E}_{i,n})$ is mutualized via the plate-wide-shared weights ϕ_i. This means that most of the weights of the flows $\mathcal{F}_{i,n}$, concentrated in ϕ_i, are trained at every optimization step across all the selected batches $\mathcal{B}_i[t]$. This results in drastically faster convergence compared to SVI, as illustrated in Section 6.3.2.

6.2.3 Encoding schemes

PAVI shares the parameterization and learning of density estimators across an HBM's plates. In practice the distributions $q_{i,n}(\theta_{i,n}; \phi_i, \mathbf{E}_{i,n})$ from Equation (6.6) with different n only differ through the value of the encodings $\mathbf{E}_{i,n}$. In the next two sections, we detail two schemes to derive those encodings.

PAVI free encoding scheme (PAVI-F)

In our core implementation, $\mathbf{E}_{i,n}$ are free weights. We define encoding arrays with the cardinality of the full model $\mathcal{M}$, one array $\mathbf{E}_i = [\mathbf{E}_{i,n}]_{n=0..N_i}$ per template θ_i. This means that an additional ground RV —for instance, adding a subject in a population study— requires an additional encoding vector. The associated increment in the total number of weights is much lighter than adding a fully parameterized normalizing flow, as would be the case in the non-plate-amortized regime.

The PAVI-F scheme cannot be sample amortized: when presented with an unseen $\mathbf{X}$, though ϕ_i can be kept as an efficient warm start, the optimal values for the encodings $\mathbf{E}_{i,n}$ have to be searched again.

During training, the encodings $\mathbf{E}_{i,n}$ corresponding to $n \in \mathcal{B}_i[t]$ are sliced from the arrays $\mathbf{E}_i$ and are optimized for along with ϕ_i. In the toy example from Figure 6.2, at $t = 0$, $\mathcal{B}_1[0] = \{1, 2\}$ and the trained encodings are $\{\mathbf{E}_{1,1}, \mathbf{E}_{1,2}\}$, and at $t = 1$ $\mathcal{B}_1[1] = \{0, 1\}$ and we train $\{\mathbf{E}_{1,0}, \mathbf{E}_{1,1}\}$.

Implementation details: plate levels and encoding tensors In practice, we have some amount of sharing across the encodings $\mathbf{E}_i$. Instead of defining separate encodings for every RV template, we define encodings for every *plate level*. A plate level is a combination of plates with at least one parameter RV template θ_i belonging to it:

$$\text{PlateLevels} = \{(\mathcal{P}_k..\mathcal{P}_l) = \text{Plates}(\theta_i)\}_{\theta_i \in \Theta} \tag{6.10}$$

For every plate level, we construct a large encoding array with the cardinalities of the full model $\mathcal{M}$:

$$\begin{aligned} \text{Encodings} &= \{(\mathcal{P}_k..\mathcal{P}_l) \mapsto \mathbb{R}^{\text{Card}(\mathcal{P}_k)\times..\times\text{Card}(\mathcal{P}_l)\times D}\}_{(\mathcal{P}_k..\mathcal{P}_l)\in\text{PlateLevels}} \\ \mathbf{E}_i &= \text{Encodings}(\text{Plates}(\theta_i)) \end{aligned} \tag{6.11}$$

Where D is an encoding size that we kept constant to de-clutter the notation but can vary between plate levels. The encodings for a given ground RV $\theta_{i,n}$ then correspond to an element from the encoding array $\mathbf{E}_i$.

PAVI deep set encoder encoding scheme (PAVI-E) and ADAVI

The parameterization of PAVI-F scales lightly but linearly with $\text{Card}(\mathcal{P})$. Though lighter than the non-plate-amortized case, this scaling could still become unafford-

able in large population studies. We thus propose an alternate scheme, PAVI-E, with a parameterization independent of plate cardinalities.

In this more experimental scheme, free encodings are replaced by an encoder f with weights η applied to the observed signal: $\mathbf{E} = f(\mathbf{X}; \eta)$. As encoder f we use a *deep-set* architecture (Zaheer et al., 2018; J. Lee et al., 2019a; Agrawal and Domke, 2021). Due to the plate structure, the observed X features multiple permutation invariances —across data points corresponding to i.i.d. RVs. Deep sets are attention architectures that can model generic permutation invariant functions. As such, they constitute a natural design choice to incorporate the problem's invariances.

The PAVI-E scheme allows for *sample amortization* across different data samples $\mathbf{X}_0, \mathbf{X}_1, ...$, as described in Section 3.3.2. Sample amortization actually does not imply any change in the PAVI-E architecture: an encoder f is used whether the variational family is sample-amortized or not.

During training, shared learning is further amplified as all the architecture's weights —ϕ_i and η— are trained at every step t. To collect the encodings to plug into $q^{\mathfrak{r}}$, we build up on a property of f: *set size generalization* (Zaheer et al., 2018). Instead of encoding the full-sized data $\mathbf{X}$, f is applied to the slice $\mathbf{X}^{\mathfrak{r}}[t]$. This amounts to aggregating summary statistics across a subset of the observed data instead of the full data (J. Lee et al., 2019a; Agrawal and Domke, 2021). The PAVI-E scheme has an even greater potential in the sample amortized context: we train a sample amortized family over the lightweight model $\mathcal{M}^{\mathfrak{r}}$ and use it "for free" to infer over the heavyweight model $\mathcal{M}$.

Disctinction between ADAVI and PAVI-E In terms of encoding scheme, PAVI-E is a generalization of the ADAVI architecture (Rouillard and Wassermann, 2022). In detail, the difference between ADAVI and the PAVI-E architecture is threefold:

- ADAVI is limited to *pyramidal* models. In the ADAVI paper (Rouillard and Wassermann, 2022), we define pyramidal models as plate-enriched models with a single "stack" of plates (no colliding plates) and a single observed RV X at the "bottom" of the pyramid. Figure 6.1 illustrates an example of a pyramidal model. In contrast, PAVI-E is generalized to arbitrary plate-enriched models (with possibly multiple observed RVs).
- As detailed in Section 6.1.2, ADAVI implements a blockwise MF dependency scheme. In contrast, PAVI-E follows a more expressive cascading dependency scheme (Ambrogioni, Lin, et al., 2021; Ambrogioni, Silvestri, et al., 2021).

- ADAVI is limited to the sample-amortized regime, whereas PAVI-E also implements a non-sample-amortized variant.

Implementation details: backward plate dependency structure Here we detail the design of the encoder $f(\,\cdot\,, \eta)$ applied to the observed data $\mathbf{X}$. As in Section 6.2.3, the role of the encoder is to produce one encoding per plate level. We start with a dependency structure for the plate levels (the symbol π denotes parents in a graph):

$$\begin{aligned} \forall (\mathcal{P}_a .. \mathcal{P}_b) \in \text{PlateLevels} \ , \\ \forall (\mathcal{P}_c .. \mathcal{P}_d) \in \text{PlateLevels} \ , \\ (\mathcal{P}_a .. \mathcal{P}_b) \in \pi((\mathcal{P}_c .. \mathcal{P}_d)) \Leftrightarrow \underset{\exists \theta_j / \, \text{Plates}(\theta_j) = (\mathcal{P}_c .. \mathcal{P}_d)}{\exists \theta_i / \, \text{Plates}(\theta_i) = (\mathcal{P}_a .. \mathcal{P}_b)} / \theta_j \in \pi(\theta_i) \end{aligned} \tag{6.12}$$

note that this dependency structure is in the *backward* direction: a plate level will be the parent of another plate level if the former contains a RV who has a child in the latter. We therefore obtain a plate level dependency structure that *reverts* the conditional dependency structure of the graph template $\mathcal{T}$. To avoid redundant paths in this dependency structure, we take the maximum branching of the obtained graph.

Given the plate level dependency structure, we will recursively construct the encodings, starting from the observed data:

$$\begin{aligned} \forall x \in X \text{ with } \text{Plates}(x) = (\mathcal{P}_a .. \mathcal{P}_b) : \\ \text{Encodings}(\text{Plates}(x)) = \rho(\mathbf{x}) \in \mathbb{R}^{\text{Card}(\mathcal{P}_a) \times .. \times \text{Card}(\mathcal{P}_b) \times D} \end{aligned} \tag{6.13}$$

where $\mathbf{x}$ is the observed data for the RV x, and ρ is a simple encoder that processes every observed ground RV's value independently through an identical MLP. Then, until we have exhausted all plate levels, we process existing encodings to produce new encodings:

$$\text{Encodings}((\mathcal{P}_c .. \mathcal{P}_d)) = g(\text{Encodings}(\pi(\mathcal{P}_c .. \mathcal{P}_d))) \in \mathbb{R}^{\text{Card}(\mathcal{P}_c) \times .. \times \text{Card}(\mathcal{P}_d) \times D} \tag{6.14}$$

where g is the composition of attention-based deep-set networks called *set transformers* (J. Lee et al., 2019b; Zaheer et al., 2018). For every plate $\mathcal{P}$ present in the parent plate level but absent in the child plate level, g will compute summary statistics *across* that plate, effectively contracting the corresponding batch dimensionality in the parent encoding (Rouillard and Wassermann, 2022).

In the case of multiple observed RVs, we run this "backward pass" independently for each observed data —with one encoder per observed RV. We then concatenate the resulting encodings corresponding to the same plate level.

Implementation details: encoding schemes

Encoding schemes are summarized in Algorithm 2.

Algorithm 2: Implementation details ADAVI and PAVI encoding schemes

---

Input:

Plate-enriched graph template $\mathcal{T}$ with:
- RV templates $\{\theta_i\}_{i=1..I}$
- Plates $\{\mathcal{P}_p\}_{p=0..P}$

Choice of variational family type: ADAVI, PAVI-F or PAVI-E

Output:

if *ADAVI or PAVI-E variational family type* **then**

Hierarchcial encoder $f(\bullet;\eta)$

else if *PAVI-F variational family type* **then**

Full encoding arrays $\{\mathbf{E}_i\}_{i=1..I}$

Algorithm:

Collect the plate levels: $\text{PlateLevels} = \{(\mathcal{P}_k..\mathcal{P}_l) = \text{Plates}(\theta_i)\}_{\theta_i \in \Theta}$

if *ADAVI or PAVI-E variational family type* **then**

Construct a backward plate level dependency graph:
$(\mathcal{P}_a..\mathcal{P}_b) \in \pi((\mathcal{P}_c..\mathcal{P}_d)) \Leftrightarrow \underset{\exists \theta_j / \text{Plates}(\theta_j)=(\mathcal{P}_c..\mathcal{P}_d)}{\exists \theta_i / \text{Plates}(\theta_i)=(\mathcal{P}_a..\mathcal{P}_b)} / \theta_j \in \pi(\theta_i)$

Prune the plate level dependency graph, taking its maximum branching

for $\mathcal{L} = (\mathcal{P}_c..\mathcal{P}_d) \in \text{PlateLevels}$ **do**

Recursively construct $\text{Encodings}(\mathcal{L}) = g_{\mathcal{L}}(\text{Encodings}(\pi(\mathcal{L})))$

$g_{\mathcal{L}}$ is a deep-set encoder contracting the plate(s) that belong to $\mathcal{L}$ and not to $\pi(\mathcal{L})$

By construction with $\mathcal{L} = (\mathcal{P}_c..\mathcal{P}_d)$,
$\text{Encodings}(\mathcal{L}) \in \mathbb{R}^{\text{Card}(\mathcal{P}_c)\times..\times\text{Card}(\mathcal{P}_d)\times D}$

The encoder $f(\bullet;\eta)$ is the hierarchical combination of the $g_{\mathcal{L}}$ encoders

else if *PAVI-F variational family type* **then**

for $\mathcal{L} = (\mathcal{P}_c..\mathcal{P}_d) \in \text{PlateLevels}$ **do**

Construct a full encoding tensor
$\text{Encodings}((\mathcal{P}_c..\mathcal{P}_d)) \in \mathbb{R}^{\text{Card}(\mathcal{P}_c)\times..\times\text{Card}(\mathcal{P}_d)\times D}$

Associate RV templates to plate-level encodings:

for $i = 1..I$ **do**

Set $\mathbf{E}_i = \text{Encodings}(\text{Plates}(\theta_i))$

Training and inference algorithms

Training Algorithm 3 summarizes PAVI's stochastic training. At each training step, a random branching is selected inside the full model. The branching determines the observed signal slice, the selected latent RVs and the associated encodings. The reduced ELBO is then evaluated, and its gradient is propagated to the variational family's weights.

Note: In the case of ADAVI, there is no stochasticity in the training. Training ADAVI amounts to approximating the ELBO via Monte Carlo and backpropagating its gradient to the conditional flows and encoder weights.

Inference Algorithm 4 summarizes inference with PAVI or ADAVI. Instead of collecting encodings over a subset of the model's graph, full encodings are collected, and fed to the conditional density estimators.

6.2.4 Stochastic training and bias

A key consideration is a potential bias introduced by the stochastic training scheme. When training stochastically over a variational family $\mathcal{Q}$, we want to converge to the same solution Q^* as if we trained over the entirety of $\mathcal{Q}$. In this section, we show that the PAVI-F scheme is unbiased. In contrast, the PAVI-E scheme is theoretically biased —though we seldom noticed any negative impact of that bias in practice.

Note: here the term *bias* refers to the stochastic training scheme. A different form of bias consists in the limited expressivity of the variational family $\mathcal{Q}$ which may not contain the true posterior $P(\Theta|X)$. We refer to this other bias as the variational family's *approximation gap*, as introduced in Section 3.3.1.

We first formalize the *plate sampling* strategy described in Section 6.2.1. To every plate $\mathcal{P}$ we associate the RV $I_{\mathcal{P}}$ corresponding to the $\mathrm{Card}^{\mathrm{r}}(\mathcal{P})$-sized set of indices sampled without replacement from the $\mathrm{Card}(\mathcal{P})$ possible index values. As an example, with a plate $\mathcal{P}$ with $\mathrm{Card}(\mathcal{P}) = 4$ and $\mathrm{Card}^{\mathrm{r}}(\mathcal{P}) = 2$, $\{0, 2\}$ or $\{2, 3\}$ can be 2 different samples from $I_{\mathcal{P}}$. At a given optimization step t, we sample independently from the RVs $\{I_{\mathcal{P}_p}\}_{p=0..P}$. This defines the batches $\mathcal{B}_i[t]$ in Equations (6.7) and (6.8).

Algorithm 3: PAVI stochastic training

Input:

Untrained variational distribution Q
Choice of encoding scheme: PAVI-F or PAVI-E
Reduced plate cardinalities $\{\mathrm{Card}^{\mathfrak{r}}(\mathcal{P}_p)\}_{p=0..P}$
Number of training steps T

Output:

Trained variational distribution Q

Algorithm:

for t=*1*..T **do**
  Sample plate indices with cardinalities $\{\mathrm{Card}^{\mathfrak{r}}(\mathcal{P}_p)\}_{p=0..P}$ following Section 6.2.1
  Based on the plate branchings select:
  - the ground observed RV batches $\mathcal{B}_X[t]$ which yield the observed signal slice $\mathbf{X}^{\mathfrak{r}}[t]$
  - the latent RV batches and encodings:
  for $i = 1..I$ **do**
    Collect the ground RV batch $\{\mathcal{B}_i[t]\}_{i=1..I}$ which yield $\theta_i^{\mathfrak{r}}[t]$
    Collect the encodings:
    if *PAVI-F encoding scheme* **then**
      Collect encodings $\mathbf{E}_{i,n}$ by slicing from the arrays $\mathbf{E}_i$ the elements $n \in \mathcal{B}_i[t]$
    else if *PAVI-E encoding scheme* **then**
      Compute encodings as $\mathbf{E} = f(\mathbf{X}^{\mathfrak{r}}[t]; \eta)$
  Assemble the reduced model density: $\log p^{\mathfrak{r}}(X^{\mathfrak{r}}[t], \Theta^{\mathfrak{r}}[t])$ as in Equation (6.7)
  Assemble the reduced variational density: $\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t])$ as in Equation (6.8)
  Approximate the reduced ELBO via Monte Carlo:
  $\mathrm{ELBO}^{\mathfrak{r}}[t] = \mathbb{E}_{\Theta^{\mathfrak{r}} \sim Q^{\mathfrak{r}}} \left[\log p^{\mathfrak{r}}(X^{\mathfrak{r}}[t], \Theta^{\mathfrak{r}}[t]) - \log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t])\right]$
  Back-propagate the reduced ELBO gradient
  Update the conditional flow weights $\{\phi_i\}_{i=1..I}$
  if *PAVI-F encoding scheme* **then**
    Update encodings $\{\mathbf{E}_{i,n}\}_{i=1..I, n \in \mathcal{B}_{i,t}}$
  else if *PAVI-E encoding scheme* **then**
    Update encoder weights η

Algorithm 4: ADAVI and PAVI inference

Input:

Trained variational distribution Q
Observed data $\mathbf{X}$
Choice of variational family type: ADAVI, PAVI-F or PAVI-E

Output:

Approximate posterior distribution $Q(\Theta) \simeq P(\Theta | X = \mathbf{X})$

Algorithm:

Collect the encodings for the full observed signal:
if *ADAVI or PAVI-E variational family type* **then**
  Compute encodings as $\mathbf{E} = f(\mathbf{X}; \eta)$
  In the case of PAVI-E, this leverages set size generalization as detailed in Section 6.2.3
else if *PAVI-F variational family type* **then**
  Collect full encoding arrays $\{\mathbf{E}_i\}_{i=1..I}$
Assemble the full variational density:
$\log q(\Theta) = \sum_{i=1}^{I} \sum_{n=0}^{N_i} \log q_{i,n}(\theta_{i,n}; \phi_i, \mathbf{E}_{i,n})$

To check the unbiasedness of our stochastic training, we need to show that:

$$\mathbb{E}_{I_{\mathcal{P}_0}} \dots \mathbb{E}_{I_{\mathcal{P}_P}} \left[\mathrm{ELBO}^{\mathfrak{r}}[t] \right] = \mathrm{ELBO} \tag{6.15}$$

Where:

$$\mathrm{ELBO} = \mathbb{E}_{\Theta \sim Q} \left[\log p(X, \Theta) - \log q(\Theta) \right] \tag{6.16}$$

And $\mathrm{ELBO}^{\mathfrak{r}}[t]$ is defined in Equation (6.9). In that expression, q and p have symmetrical roles. As the ELBO amounts to the difference between the logarithms of densities p and q, we can prove the equality in Equation (6.15) if we prove that the expectation of each reduced distribution is equal to the corresponding full distribution. To prove the equality in Equation (6.15), a sufficient condition is therefore to prove that:

$$\mathbb{E}_{I_{\mathcal{P}_0}} \dots \mathbb{E}_{I_{\mathcal{P}_P}} \left[\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t]) \right] = \mathbb{E}_{I_{\mathcal{P}}} \left[\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t]) \right] = \log q(\Theta) \tag{6.17}$$

where to de-clutter the notations we denote the expectation over the collection of RVs $\{I_{\mathcal{P}_p}\}_{p=0..P}$ as $\mathbb{E}_{I_{\mathcal{P}}}$.

Consider a given ground RV $\theta_{i,n}$ corresponding to the RV template θ_i and to the plates $\mathrm{Plates}(\theta_i)$. At a given stochastic step t, $\theta_{i,n}$ will be chosen if and only if its corresponding *branching* is chosen. Recall that when sampling equiprobably without replacement a set of k elements from a population of n elements, a given element

will be present in the set with probability k/n. We can apply this reasoning to the choice of *branching* corresponding to a given ground RV. For instance, in Figure 6.2, X_2 will be chosen if and only if the index 1 is selected as part of sub-sampling $\mathcal{P}_1$ and the index 0 is selected as part of sub-sampling $\mathcal{P}_0$. As $\mathrm{Card}^{\mathfrak{r}}(\mathcal{P}_1) = 2$ indices are chosen inside the plate $\mathcal{P}_1$ of full cardinality 3, and $\mathrm{Card}^{\mathfrak{r}}(\mathcal{P}_0) = 1$ indices are chosen inside the plate $\mathcal{P}_0$ of full cardinality 2, X_2 is therefore chosen with probability $2/3 \times 1/2$. More formally, for a ground RV $\theta_{i,n}$ we have:

$$\begin{aligned} \forall n = 0..N_i \ : \ \mathbb{P}(\theta_{i,n} \in \mathcal{B}_i[t]) &= \prod_{\mathcal{P} \in \mathrm{Plates}(\theta_i)} \frac{\mathrm{Card}^{\mathfrak{r}}(\mathcal{P})}{\mathrm{Card}(\mathcal{P})} \\ &= \frac{N_i^{\mathfrak{r}}}{N_i} \end{aligned} \tag{6.18}$$

Applying this reasoning to every RV template θ_i, we have that:

$$\begin{aligned} \mathbb{E}_{I_{\mathcal{P}}}\left[\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t])\right] &= \sum_{i=1}^{I} \frac{N_i}{N_i^{\mathfrak{r}}} \mathbb{E}_{I_{\mathcal{P}}}\left[\sum_{n \in \mathcal{B}_i[t]} \log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}))\right] \\ &= \sum_{i=1}^{I} \frac{N_i}{N_i^{\mathfrak{r}}} \mathbb{E}_{I_{\mathcal{P}}}\left[\sum_{n=0}^{N_i} \mathbb{1}_{n \in \mathcal{B}_i[t]} \log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}))\right] \\ &= \sum_{i=1}^{I} \frac{N_i}{N_i^{\mathfrak{r}}} \sum_{n=0}^{N_i} \mathbb{E}_{I_{\mathcal{P}}}\left[\mathbb{1}_{n \in \mathcal{B}_i[t]} \log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}); \phi_i, \mathbf{E}_{i,n})\right] \end{aligned} \tag{6.19}$$

where we exploited the fact that the expectation of the sum of RVs is the sum of the expectations, even in the case of dependent RVs. The term $\mathbb{1}_{n \in \mathcal{B}_i[t]} \times \log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}); \phi_i, \mathbf{E}_{i,n})$ is the product of 2 RVs —related to the stochastic choice of plate indices:

- the RV $\mathbb{1}_{n \in \mathcal{B}_i[t]}$ is an indicator that $\theta_{i,n}$'s *branching* has been chosen via the stochastic sampling of plate indices. By construction, this RV depends only on the indices of the plates $\mathcal{P} \in \mathrm{Plates}(\theta_i)$.
- the RV $\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}); \phi_i, \mathbf{E}_{i,n})$ depends on $\mathbf{E}_{i,n}$, whose construction depends on the encoding scheme:
  - In the PAVI-F scheme, $\mathbf{E}_{i,n}$ is a constant.
  - In the PAVI-E scheme, $\mathbf{E}_{i,n}$ results of the application of an encoder to the observed data of a subset of $\theta_{i,n}$'s descendants. By construction, this subset will only depend on the indices of plates containing θ_i's descendants, but not containing θ_i. The value of $\mathbf{E}_{i,n}$ therefore only depends on the indices of plates $\mathcal{P} \notin \mathrm{Plates}(\theta_i)$

As an example of this reasoning, consider the model $\mathcal{M}$ illustrated in Figure 6.2. We can evaluate both terms for the ground RV $\theta_{1,2}$ in the PAVI-E scheme:

- $\mathbb{1}_{2\in\mathcal{B}_1[t]}$ depends on whether the index 2 is chosen as part of sub-sampling the plate $\mathcal{P}_1$, and therefore only depends on the RV $I_{\mathcal{P}_1}$. In this case, the associated probability is $2/3$;

- to evaluate $\log q_{1,2}(\theta_{1,2}|\theta_{2,0};\phi_1,\mathbf{E}_{1,2})$, the value of $\mathbf{E}_{1,2}$ will result from the application of the encoder f over the value of either X_4 or X_5. This choice depends on whether the index 0 or 1 is chosen as part of sub-sampling the plate $\mathcal{P}_0$. Therefore, the value of the term $\log q_{1,2}$ only depends on the RV $I_{\mathcal{P}_0}$.

In summary, in both PAVI-F and PAVI-E, the terms $\mathbb{1}_{n\in\mathcal{B}_i[t]}$ and $\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\phi_i,\mathbf{E}_{i,n})$ depend on the sampled indices of *disjoint* sets of plates, and are therefore independent. This means that the expectation of their product can be rewritten as the product of their expectations:

$$\begin{aligned}\mathbb{E}_{I_\mathcal{P}}\left[\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t])\right] &= \sum_{i=1}^{I}\frac{N_i}{N_i^{\mathfrak{r}}}\sum_{n=0}^{N_i}\mathbb{E}_{I_\mathcal{P}}\left[\mathbb{1}_{n\in\mathcal{B}_i[t]}\right]\mathbb{E}_{I_\mathcal{P}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}))\right]\\ &= \sum_{i=1}^{I}\frac{N_i}{N_i^{\mathfrak{r}}}\sum_{n=0}^{N_i}\frac{N_i^{\mathfrak{r}}}{N_i}\mathbb{E}_{I_\mathcal{P}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n}))\right]\\ &= \sum_{i=1}^{I}\sum_{n=0}^{N_i}\mathbb{E}_{I_\mathcal{P}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\mathbf{E}_{i,n})\right]\end{aligned} \tag{6.20}$$

This equality can be further simplified in the PAVI-F case —proving its unbiasedness— but not in the PAVI-E case, as detailed in the sections below.

Unbiasedness of the PAVI-F scheme

In the PAVI-F scheme, detailed in Section 6.2.3, the encodings $\mathbf{E}_{i,n}$ are constants with respect to the branching choice, therefore we have:

$$\begin{aligned}\mathbb{E}_{I_\mathcal{P}}\left[\log q^{\mathfrak{r}}(\Theta^{\mathfrak{r}}[t])\right] &= \sum_{i=1}^{I}\sum_{n=0}^{N_i}\mathbb{E}_{I_\mathcal{P}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\mathbf{E}_{i,n})\right]\\ &= \sum_{i=1}^{I}\sum_{n=0}^{N_i}\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\mathbf{E}_{i,n})\\ &= \log q(\Theta)\end{aligned} \tag{6.21}$$

which proves Equation (6.17) and Equation (6.15). In the above example of $\theta_{1,2}$ in $\mathcal{M}$, in the PAVI-F scheme the expression $\mathbb{E}_{I_\mathcal{P}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\phi_i,\mathbf{E}_{i,n})\right]$ can be evaluated into $\log q_{1,2}(\theta_{1,2}|\theta_{2,0};\phi_1,\mathbf{E}_{1,2})$.

Equation (6.21) demonstrates that the PAVI-F scheme is unbiased: training over stochastically chosen sub-graphs for Q^r is in expectation equal to training over the full graph of Q.

Approximations in the PAVI-E scheme

In the PAVI-E scheme, detailed in Section 6.2.2, the encodings $\mathbf{E}_{i,n}$ are computed from the observed data $\mathbf{X}$. Specifically, considering the ground RV $\theta_{i,n}$, we have $\mathbf{E}_{i,n} = f(\mathbf{X}^\mathrm{r}_{i,n}[t])$ where $\mathbf{X}^\mathrm{r}_{i,n}[t]$ corresponds to the observed data of a subset of $\theta_{i,n}$'s descendants. Depending on the chosen branching *downstream* of $\theta_{i,n}$, the value of $\mathbf{E}_{i,n}$ can therefore vary. This means we cannot further simplify Equation (6.20): the terms $\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\mathbf{E}_{i,n})$ are not constants with respect to the RVs $I_\mathcal{P}$. In the above example of $\theta_{1,2}$ in $\mathcal{M}$, in the PAVI-E scheme the expression $\mathbb{E}_{I_\mathcal{P}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\phi_i,\mathbf{E}_{i,n})\right]$ can be evaluated into:

$$\frac{1}{2}(\log q_{1,2}(\theta_{1,2}|\theta_{2,0};\phi_1,f(\mathbf{X}_4)) + \log q_{1,2}(\theta_{1,2}|\theta_{2,0};\phi_1,f(\mathbf{X}_5)))$$

How could the PAVI-E scheme be made unbiased? Specifically, by making the value of $\mathbf{E}_{i,n}$ independent of the choice of downstream branching. A possibility would be to parameterize $\mathbf{E}_{i,n}$ as an average —an expectation— over all the possible sub-branchings downstream of $\theta_{i,n}$. Yet, in practical cases, the cardinalities of the reduced model are much inferior to the ones of the full model: $\mathrm{Card}^\mathrm{r}(\mathcal{P}) \ll \mathrm{Card}(\mathcal{P})$. This means that numerous $\mathrm{Card}^\mathrm{r}(\mathcal{P})$-sized subsets can be chosen inside the $\mathrm{Card}(\mathcal{P})$ possible descendants. In order to average over all those subset choices to compute $\mathbf{E}_{i,n}$, numerous encoding calculations would be required at each stochastic training step. For large-scale cases, we deemed this possibility impractical. Other possibilities could exist, revolving around the problem of aggregating collections of stochastic estimators into one general estimator —in an unbiased and efficient manner. To our knowledge, this is a complex and still open research question, whose advancement could much benefit our applications.

Practical approximation for the PAVI-E scheme In practice, we compute the encoding $\mathbf{E}_{i,n}$ based on the single downstream branching corresponding to the sampling of the RVs $I_\mathcal{P}$. Compared to the previous paragraph, this amounts to estimating the

expectation of $\mathbf{E}_{i,n}$ —over all downstream branchings— using a single one of those branchings. Note that, even if this encoding estimate was unbiased, $\log q_{i,n}$ would remain a highly non-linear function of $\mathbf{E}_{i,n}$. As a consequence, we need to rely on the approximation:

$$\mathbb{E}_{I_{\mathcal{P}}}\left[\log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\phi_i, f(\mathbf{X}^{\mathfrak{r}}_{i,n}[t]))\right] \simeq \log q_{i,n}(\theta_{i,n}|\pi(\theta_{i,n});\phi_i, f(\mathbf{X}_{i,n})) \quad (6.22)$$

which can theoretically introduce some bias in our gradients. The approximation Equation (6.22) can be interpreted as follows: "the expectation of the density of $\theta_{i,n}$ when collecting summary statistics over a stochastic subset of $\theta_{i,n}$'s descendants is approximately equal to the density of $\theta_{i,n}$ when collecting summary statistics over the entirety of $\theta_{i,n}$'s descendants". Another interpretation is that the distribution associated with the summary of the full data can be approximated by annealing the distributions associated with summaries of subsets of this data.

In practice, our approximation did not yield significantly worse performance for the PAVI-E scheme over the generative models we tested. At the same time, computing the encodings over a single branching allows the computation of all the $\mathbf{E}_{i,n}$ encodings in a single lightweight pass over the data $\mathbf{X}^{\mathfrak{r}}[t]$. This simple solution thus provided a substantial increase in training speed with seldom noticeable bias. Yet, we do not bar the existence of pathological generative HBMs where this approximation would become coarse. Experimenters should bear in mind this possibility when using the PAVI-E scheme. In practice, using the PAVI-F scheme as a sanity check over synthetic, toy-dimension implementations of the considered generative models is a good way to validate the PAVI-E scheme —before moving on to the real problem instantiating the same generative model with a larger dimensionality.

6.3 Experimental results

This section condenses essential results presented as part of our methodological contributions. Other results and discussion points are listed and explained in the next section.

6.3.1 Illustration of the approximation gap

This subsection is an adaptation from experiment 3.2 in the ADAVI paper.

Tab. 6.1.: **Effect of the approximation gap** Differences in ELBO directly translate differences in r-KL to the ground truth posterior —see Section 3.3.3. NF-based methods (CF and ADAVI) get closer to the ground truth posterior than a less expressive density approximator (a Gaussian). *Performance is averaged over 20 generative model samples, 20 random seeds per sample.*

Type	Method	ELBO (higher is better)
Fixed parametric form	Gaussian MF	-21.0 ($\pm$ 0.2)
NF-based	CF	-17.5 ($\pm$ 0.1)
	ADAVI (ours)	-17.6 ($\pm$ 0.3)

This toy experiment illustrates the approximation gap. It puts forward the interest in using expressive density approximators such as NFs. We pitch three example methods on a voluntarily non-canonical model:

$$\begin{aligned} D &= 2 \\ a &\sim \mathrm{Gamma}(\vec{1}_D, \vec{0.5}_D) \\ \forall n = 1..10 \quad b_n | a &\sim \mathrm{Laplace}(a, \vec{0.3}_D) \end{aligned} \tag{6.23}$$

We place ourselves in a setup where the posterior distribution of a given an observed value from b has no known parametric form. In particular, the posterior is not of the same parametric form as the prior. Such an example is called *non-conjugate* (Gelman, Carlin, et al., 2004).

We compare our method ADAVI to two baselines:

- **Gaussian MF** fits a Gaussian distribution with parametric mean and variance over the posterior;
- **Cascading Flows (CF)** (Ambrogioni, Silvestri, et al., 2021) is a non-plate-amortized structured VI architecture. CF pushes the prior P into the posterior Q using *Highway Flows*. CF follows a cascading dependency structure complemented by a backward auxiliary coupling. CF thus is an expressive structured baseline that does not pay particular attention to scalability.

Results are visible in Table 6.1. NF-based methods do not require the experimenter to specify a given parametric form for the posterior —such as a Gaussian. In doing so, they avoid variational family misspecification. Leveraging flows as part of automatic, structured VI, is an important step towards unbiased VI.

6.3.2 Plate amortization speeds up convergence during stochastic training

This subsection adapts experiment 5.1 from the PAVI paper. This experiment illustrates how plate amortization results in faster training.

We use the Gaussian random effects (GRE) model, described in the following equations:

$$\begin{array}{lll} & \theta_{2,0} & \sim \mathcal{N}(\vec{0}_D, \sigma_2^2) \\ \forall n_1 = 1..\,\mathrm{Card}(\mathcal{P}_1) & \theta_{1,n_1}|\theta_{2,0} & \sim \mathcal{N}(\theta_{2,0}, \sigma_1^2) \\ \forall n_0 = 1..\,\mathrm{Card}(\mathcal{P}_0) \quad \forall n_1 = 1..\,\mathrm{Card}(\mathcal{P}_1) & X_{n_1,n_0}|\theta_{1,n_1} & \sim \mathcal{N}(\theta_{1,n_1}, \sigma_x^2) \end{array} \tag{6.24}$$

Here we set $D = 8$, $\mathrm{Card}(\mathcal{P}_1) = 100$ and $\mathrm{Card}^{\mathrm{r}}(\mathcal{P}_1) = 2$. In this experiment, we set $\mathrm{Card}^{\mathrm{r}}(\mathcal{P}_1) \ll \mathrm{Card}(\mathcal{P}_1)$. This emulates a regime in which SVI is slow because only a small fraction —of size $\mathrm{Card}^{\mathrm{r}}(\mathcal{P}_1)$— of a large parameter space —of size $\mathrm{Card}(\mathcal{P}_1)$— gets optimized at a given stochastic training step. We compare our PAVI architecture to a baseline with the same architecture, trained stochastically with SVI (Matthew D. Hoffman, David M. Blei, et al., 2013), but without plate amortization. The only difference is that ground RVs $\theta_{i,n}$ are associated in the baseline to individual fully-parameterized flows $\mathcal{F}_{i,n}$ instead of sharing the same conditional flow $\mathcal{F}_i$, as described in Section 6.1.2.

Figure 6.3 displays the evolution of the ELBO across training steps for the baseline and with free encoding PAVI-F and deep-set encoders PAVI-E. Both plate-amortized methods reach asymptotic ELBO equal to the non-plate-amortized baseline's but with orders of magnitudes faster convergence and more numerical stability. This stems from the individual flows $\mathcal{F}_{i,n}$ in the baseline only being trained when the stochastic algorithm visits the corresponding $\theta_{i,n}$. In contrast, our shared flow $\mathcal{F}_i$ is updated at every optimization step in PAVI. Intuitively, the PAVI-E scheme should converge faster than PAVI-F by sharing the training not only of the conditional flows but also of the encoder across the different optimization steps. However, the computation required to derive the encodings from the observed data results in longer optimization steps and slower inference, as illustrated in Section 6.3.3.

6.3.3 Designing scalable variational families

This subsection reproduces experiment 5.3 from the PAVI paper.

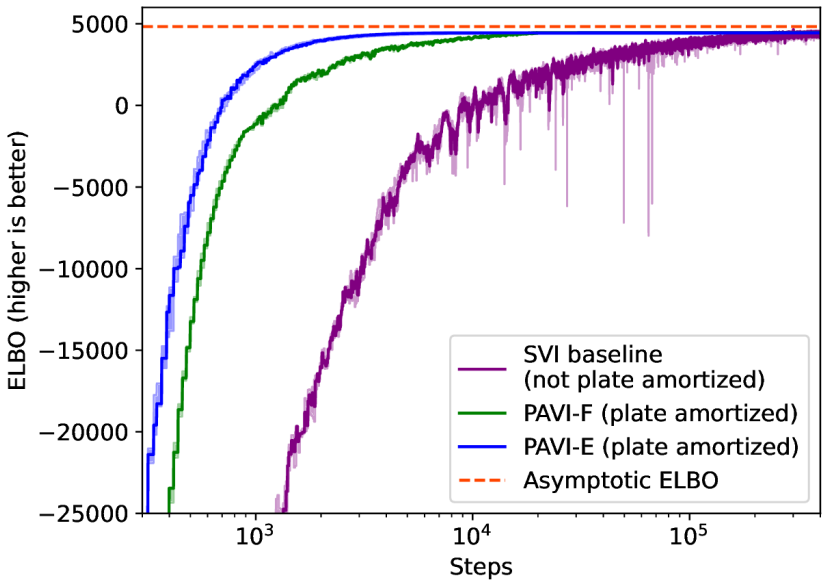

Fig. 6.3.: **Plate amortization increases convergence speed** We plot the ELBO (higher is better) as a function of the optimization steps (log-scale) for our methods PAVI-F (in green) and PAVI-E (in blue) versus a non-plate-amortized baseline (in purple). Due to plate amortization, our method converges ten to a hundred times faster to the same asymptotic ELBO as its non-plate-amortized counterpart. *Standard deviation across 20 samples, 5 random seeds per sample is displayed as a shaded area. A dashed line denotes the asymptotic closed-form performance, constructed using Gaussian distributions centered on the empirical group and population means.*

Here, we put in perspective the gains from plate amortization when scaling up an inference problem's cardinality. We consider the GRE model in Equation (6.24) with $D = 2$ and augment the plate cardinalities $(\mathrm{Card}(\mathcal{P}_1), \mathrm{Card}^{\mathfrak{r}}(\mathcal{P}_1)) : \ (2, 1) \rightarrow (20, 5) \rightarrow (200, 20)$. In doing so, we augment the number of parameters $\Theta : \ 6 \rightarrow 42 \rightarrow 402$.

Baselines We compare our PAVI and ADAVI architecture against two state-of-the-art AVI baselines.

- **Cascading Flows (CF)** (Ambrogioni, Silvestri, et al., 2021) is described in Section 6.3.1;
- **Unbiased Implicit VI (UIVI)** (Titsias and Ruiz, 2019) is an unstructured implicit VI architecture. UIVI infers over the full parameter space Θ, without any SVI-amenable factorization —contrary to CF, ADAVI, and PAVI. To do so, UIVI reparameterizes a base distribution with a stochastic transform. UIVI does not explicitly define a density q and relies on optimization steps intertwined with MCMC runs. UIVI thus consists of a non-structured VI baseline that does not pay particular attention to scalability to a large parameter space This means that UIVI could not be applied above a certain cardinality due to its impossibility to be stochastically trained.

For all architectures, we indicate with the suffix *(sa) sample amortization*, as defined in Section 3.3.2.

As the cardinality of the problem augments, Figure 6.4 shows how PAVI and ADAVI maintain a state-of-the-art inference quality while being more computationally attractive.

Parameterization In terms of parameterization, both ADAVI and PAVI-E provide a heavyweight but constant parameterization as the cardinality $\mathrm{Card}(\mathcal{P}_1)$ of the problem augments. This is due to both methods' usage of an encoder of the observed data, which makes their parameterization independent of the problem's cardinality. Comparatively, both CF and PAVI-F's parameterization scale linearly with $\mathrm{Card}(\mathcal{P}_1)$, but with a drastically lighter augmentation for PAVI-F. The difference can be explained by the additional weights each architecture requires for an additional ground RV. CF requires an additional fully parameterized normalizing flow, whereas PAVI-F only requires an additional lightweight encoding vector. In detail, PAVI-F's parameterization due to the plate-wide-shared ϕ_1 represents a constant $\approx 2k$ weights, while the part due to the encodings $\mathbf{E}_{1,n}$ grows linearly from 16 to 160 to $1.6k$ weights.

UIVI's parameterization scales quadratically with the size of the parameter space Θ. This is due to UIVI's usage of a neural network to regress the weights of a transform applied to a base distribution with the size of Θ. As the cardinality augments, UIVI's quadratic weight scaling would be limiting before CF and PAVI-F's linear scaling, which would be limiting before ADAVI and PAVI-E's constant scaling.

Memory Regarding computational budget, PAVI's stochastic training allows for controlled GPU memory during optimization. This removes the need for a larger memory as the cardinality of the problem augments, a hardware constraint that can become unaffordable at very large cardinalities. CF could be trained stochastically to remove this memory constraint but, without plate amortization, would suffer from slower inference, as illustrated in Section 6.3.2. In contrast, UIVI could not be trained stochastically, as it infers over the full parameter space Θ at once instead of factorizing it. As a result, UIVI would be ultimately limited by memory to infer over larger problems.

Speed Regarding convergence speed, PAVI benefits from plate amortization to have orders of magnitude faster convergence compared to structured VI baselines CF and ADAVI. This means that a stochastically-trained architecture (PAVI) trains faster than non-stochastic baselines (CF, ADAVI) This result is opposite to the result we would have obtained without plate amortization since SVI slows down inference. For UIVI, as the cardinality augments, training amounts to evaluating a neural network with increasingly larger layers. GPU training time is thus constant, but this property would not translate to larger problems as the GPU memory would become insufficient. Plate amortization is particularly significant for the PAVI-E(sa) scheme, in which a sample-amortized variational family is trained over a dataset of reduced cardinality yet performs "for free" inference over an HBM of large cardinality. Maintaining $\mathrm{Card}^{\mathrm{r}}(\mathcal{P}_1)$ constant while $\mathrm{Card}(\mathcal{P}_1)$ augments allows for a constant parameterization *and training time* as the cardinality of the problem augments. This property is not limited to any maximum cardinality, contrary to UIVI. This is a novel result with strong future potential. The effect of plate amortization is particularly noticeable at $\mathrm{Card}(\mathcal{P}_1) = 200$ between the PAVI-E(sa) and CF(sa) architectures, where PAVI performs sample-amortized inference with $10\times$ fewer weights and $100\times$ lower training time.

Scaling even higher the cardinality of the problem —$\mathrm{Card}(\mathcal{P}_1) = 2000$ for instance— renders ADAVI, CF and UIVI computationally intractable In contrast, PAVI maintains a light memory footprint and a short training time.

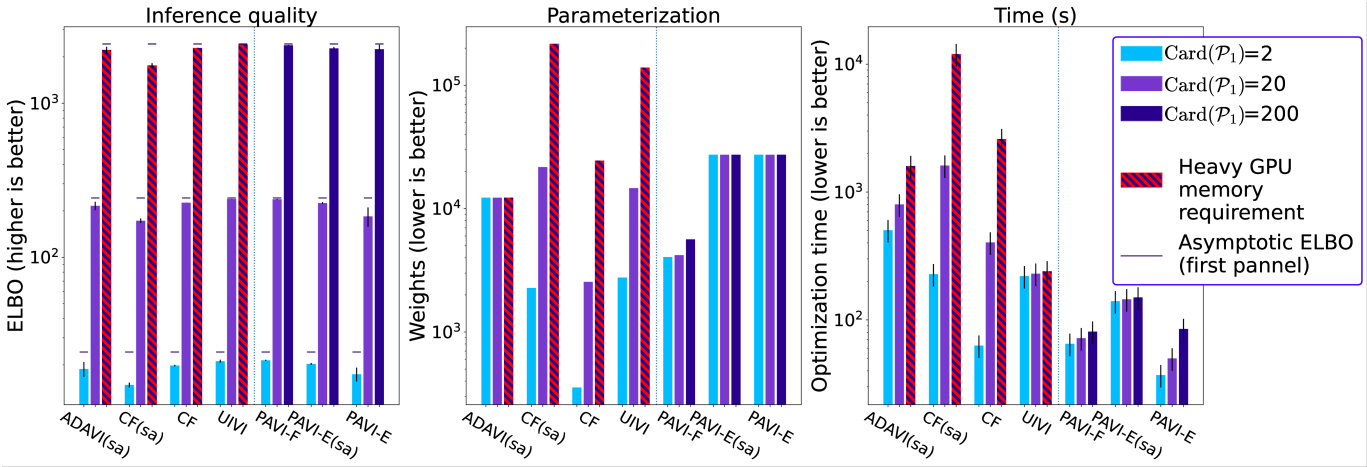

Fig. 6.4.: PAVI and ADAVI scale favorably as the cardinality of the target model augments Baselines are compared in each panel, with the suffix (sa) indicating *sample amortization* —as defined in Section 3.3.2. Our architecture PAVI is displayed on the right of each panel, and ADAVI on the left. We augment the cardinality $\mathrm{Card}(\mathcal{P}_1)$ of the GRE model, which is described in Equation (6.24). While doing so, we compare three different metrics. *In the first panel:* inference quality, as measured by the ELBO. An asymptotic closed-form ELBO is displayed using a dark blue dash. None of the presented state-of-the-art architecture's performance degrades as the cardinality of the problem augments. *In the second panel:* parameterization, comparing the number of trainable weights of each architecture. PAVI –similar to ADAVI– displays a constant number of weights as the cardinality of the problem increases —or almost constant for PAVI-F. *Third panel:* GPU training time. Benefiting from learning across plates, PAVI has a short and almost constant training time as the cardinality of the problem augments. At $\mathrm{Card}(\mathcal{P}_1) = 200$, CF, UIVI, and ADAVI required large GPU memory, a constraint absent from PAVI due to its stochastic training.

6.4 Other results and discussion points

This thesis reviewed essential theoretical and experimental results. Here we quickly review some additional points:

- ADAVI (Rouillard and Wassermann, 2022):
  - **Section 3.4: comparative experiment on a Gaussian mixture (GM) experiment** ADAVI is compared to a variety of methods, including the structured VI baseline cascading flows (Ambrogioni, Silvestri, et al., 2021), a SBI baseline (S)NPE-C (Greenberg et al., 2019), and a r-KL-trained NF baseline (Rezende and Mohamed, 2015). In this challenging setup —that involves the label switching issue (Jasra et al., 2005)— we show the comparative advantage of non-amortized over amortized baselines; of r-KL-trained baselines over f-KL-trained baselines; and of structured VI baselines over unstructured baselines. In particular, ADAVI yields the best performance amongst amortized baselines.
  - **Supplemental B.3: discussion on the relevance of likelihood-free methods (trained using the f-KL) in the presence of a likelihood** There is a general belief in the community that likelihood-free methods are not intended to be as competitive as likelihood-based methods in the presence of a likelihood (Cranmer et al., 2020). The ADAVI paper provides quantitative results to nourish this debate. Though likelihood-free methods generally scale poorly to high dimensions, they are dramatically faster to train and can perform on par with r-KL baselines on some examples (such as the GRE). Depending on the problem at hand, it is therefore not straightforward to systematically disregard likelihood-free methods.
- PAVI (Rouillard, Bris, et al., 2023):
  - **Section 5.2: impact of the encodings $\mathbf{E}_{i,n}$ on the approximation gap** —see Section 6.1.2 for a definition. We put forward the role of the $\mathbf{E}_{i,n}$ as ground RV's summary statistics. As the encoding size augments, so does the asymptotic performance of PAVI until reaching the dimensionality of the posterior's sufficient statistics, after which performance plateaus. Encoding size allows for a clear trade-off between memory and inference quality.
  - **Supplemental A.4: discussion on the *gaps* introduced by plate amortization**, which affects the variational family's expressivity. We show that in

theory plate amortization introduces a *plate amortization gap*, cumulative with the *sample amortization gap* and the *approximation gap* described in Sections 3.3.1 and 3.3.2. In practice, however, plate amortization can stabilize inference and yield higher ELBO compared to non-plate-amortized variants. We interpret this as a result of a simplified optimization problem —with fewer parameters to optimize for, and mini-batching effects across different ground RVs.

– **Supplemental B.3: impact on training speed of the reduced model cardinalities** $\mathrm{Card}^{\mathrm{r}}(\mathcal{P})$. Those define the "minibatch size" of paths taken inside the full model. We show that PAVI-F's and PAVI-E's asymptotic performance does not depend on the reduced model cardinalities. This confirms the unbiasedness of the stochastic training discussed in Section 6.2.3. We also show that PAVI-F converges faster as the reduced model cardinalities get closer to the full cardinalities. In contrast, PAVI-E's training speed is *constant* with respect to the reduced model cardinalities. This illustrates PAVI-E's potential for constant parameterization *and training time* as the cardinality of the problem augments.

– **Supplemental B.4: comparative experiments on a variety of HBMs** We compare PAVI-F and PAVI-E to cascading flows (Ambrogioni, Silvestri, et al., 2021), UIVI (Titsias and Ruiz, 2019) and ADAVI (Rouillard and Wassermann, 2022) as described in Sections 6.3.1 and 6.3.3. As HBMs, we compare over a GM model; a model featuring the aggregation of higher-order summary statistics; and over a smaller version of our parcellation model (P) model used in Chapter 7 —described in Equation (7.1). Across this large panel, we show the superior performance of PAVI-F, both in terms of asymptotic ELBO and convergence speed.

6.5 Summary of contributions

This chapter summarized our methodological contributions:

- We leverage NFs as expressive conditional density estimators inside structured VI. In particular, we propose the systematic use of NFs amortized across plates.
- We adjoin to those NFs schemes to replicate the causal structure of the model into encodings.
- The obtained architecture yields expressive yet parsimoniously parameterized variational families, adapted to the large scale.
- We propose algorithms to derive those architectures automatically from the model.
- We propose stochastic training schemes for those architectures, to control the compute in addition to the parameterization.
- Combined with plate amortization, this training scheme yields faster convergence, effectively unlocking large-scale inference in a reasonable time.

The following chapters apply those methodologies to different large-scale Neuroimaging hierarchical inference problems.

7

Application: individual parcellation of the human cortex

This chapter tackles parcellation over fMRI data, as presented in Section 1.2. The goal is to map the human cortex into parcels of distinct functional connectivity (Simon B. Eickhoff et al., 2018b). Functional connectivity corresponds to patterns of coactivation with the rest of the brain (Van Den Heuvel and Pol, 2010). The resulting networks of connected regions divide the brain into major units associated with dedicated cognitive functions (such as vision or motor control).

This application is complicated by a massive dimensionality (the signal contains several billion measures) and by a large variability across individuals: the same functions are not located in the same regions from one subject to another. We demonstrate the ability of VI to tackle such large-scale problems, aggregating the individual maps of a thousand individuals. We also show the ability of HBMs to act as transferrable representations, allowing to stabilize parcellations across different datasets.

Note on contributions: In this work, I have mostly focused on modeling and inference, obtaining the parcellations in Section 7.3.1. Cognition prediction in Section 7.3.2 has been jointly performed with another PhD student, Alexandre Le Bris. Regarding transfer learning in Section 7.3.3, I took on a middle-author role (theory, discussions, no experiments), with Alexandre assuming first authorship. Transfer learning is presented as an extension of the methodologies developed in this thesis.

7.1 Application context

7.1.1 Neuroimaging-based parcellations

Brain cartography has a long history (Simon B Eickhoff et al., 2018a). The earliest attempts to map the human brain were based on ex-vivo investigation of its micro- and macrostructure. Abrupt changes in markers such as the thickness of cortical layers were used to detect boundaries separating distinct areas (K. Brodmann and M. Brodmann, 1909). Interestingly, this anatomical segregation of the brain does not exactly correlate with its functional organization (Simon B. Eickhoff et al., 2018b). As a result, in the past decades, research has accumulated on **in-vivo parcellations based on Neuroimaging** (Simon B. Eickhoff et al., 2018b). This work focuses on imaging-based parcellations, specifically using fMRI.

A major hypothesis organizes brain function following two overarching principles (Tononi et al., 1994). One is the *segregation* of information into specialized regions. The other is the *integration* of information through large-scale networks. Both principles correspond to different functional granularities, that are reflected into maps with different numbers of parcels. Fine-granularity parcellations include hundreds, up to a thousand very specific parcels (Dadi et al., 2020; Schaefer et al., 2018; Fan et al., 2016). In contrast, this work investigates coarse-granularity parcellations, containing dozens of **integrative networks** (Thomas Yeo et al., 2011; Kong, J. Li, et al., 2019; Power et al., 2011). A prominent example is the parcellation from Thomas Yeo et al. (2011), who subdivided the brain into 7 networks, as illustrated in Figure 7.1. The goal of those parcellations is to highlight distributed systems of (sometimes disconnected) brain areas.

Following the taxonomy from Simon B. Eickhoff et al. (2018b), parcellations can be broadly separated into two groups. On the one hand are parcellations based on *local* architecture or function, which are often combined with border detection algorithms. By design, those parcelations yield spatially contiguous parcels. On the other hand are parcellations based on *global* fingerprints, associated with clustering algorithms grouping brain regions potentially distant in space (Churchland and Sejnowski, 1988). This work belongs to the latter category. Each point in the cortex is associated with a vector describing its connectivity with the entirety of the brain (Van Den Heuvel and Pol, 2010). Points with a similar fingerprint are then grouped into networks. To compute this **connectivity fingerprint**, *resting-state* fMRI is used (Van Den Heuvel and Pol, 2010; J. Bijsterbosch et al., 2017; Poldrack et al., 2011). This is in opposition to *task* fMRI data, in which subjects perform given certain

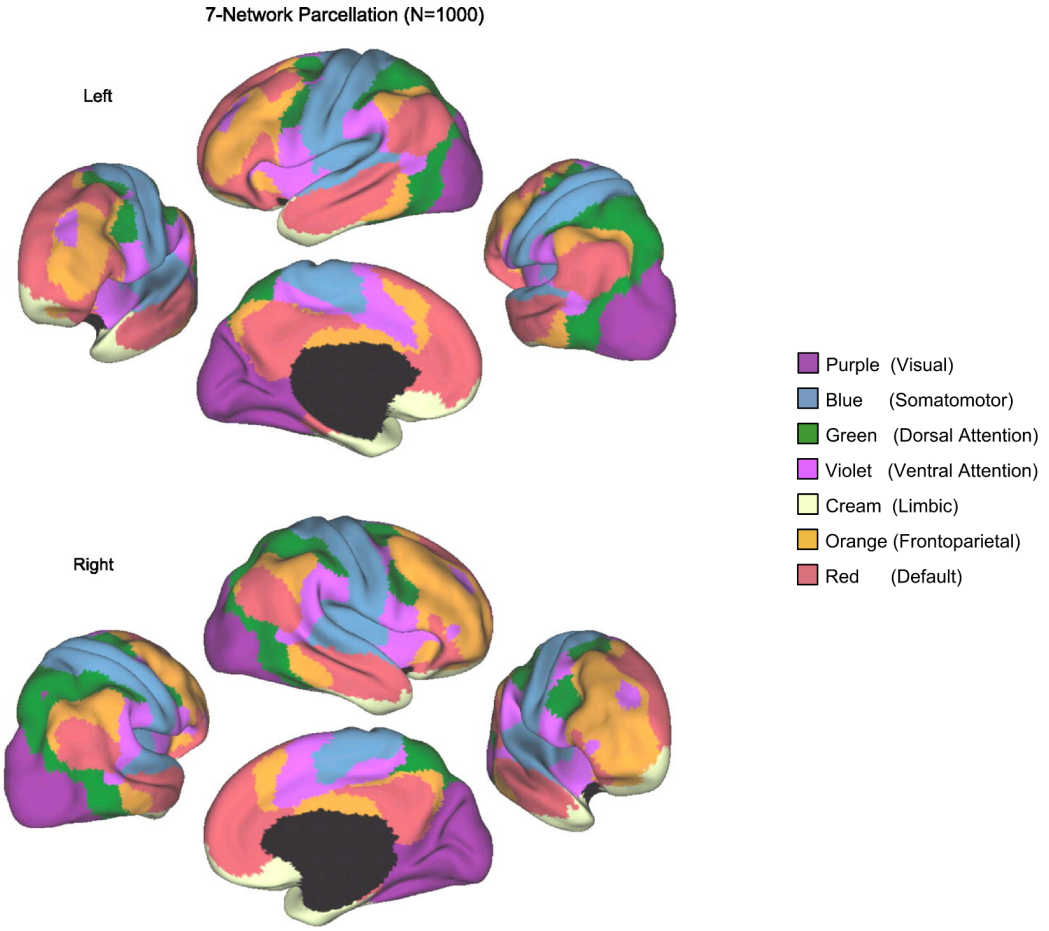

Fig. 7.1.: **Yeo 7 functional networks** (Thomas Yeo et al., 2011) The human cortex is subdivided into 7 *networks*: regions that co-activate when the brain is at rest. Those networks can be broadly associated with cognitive functions, such as vision or motor control. *Figure adapted from Thomas Yeo et al. (2011)*

cognitive activities (reading, remembering previous items, etc...). Using data from rest, the goal is to investigate the "default" organization of the brain, which can be modulated as part of performing cognitive tasks.

Thomas Yeo et al. (2011) averaged data across a thousand individuals (Van Essen et al., 2012), yielding two parcellations into 7 and 17 networks considered as a reference in Neuroimaging (Simon B. Eickhoff et al., 2018b). At a later stage, Kong, J. Li, et al. (2019) performed a similar analysis to obtain individual parcellations. Interestingly, they showed that the subject's *topography* —the localization of the networks on their cortex— was predictive of the subject's cognition and behavior (as measured via test scores) (J. D. Bijsterbosch et al., 2018). One goal of this work is to reproduce such an analysis: **to obtain individual parcellations and test their predictive power over cognition**.

From a methodological standpoint, both parcellations (Thomas Yeo et al., 2011; Kong, J. Li, et al., 2019) are based on a HBM associated with an *EM algorithm*. This creates a strong link between these works and the VI methods studied in this thesis. In particular, EM algorithms are very close to the pen-and-paper derivations described in Section 3.3.4. EM requires to manually derive an HBM-specific optimization routine that converges to optimal parameter estimates. This creates a methodological barrier to entry for experimenters to produce similar analyses. What's more, should the HBM modeling the observed signal change, a new routine would need to be designed. Finally, not every HBM would yield a simple-to-derive closed-form routine, which reduces modeling possibilities. In contrast, we **demonstrate the ability of AVI (a general method) to tackle HBMs similar to the ones from Kong, J. Li, et al. (2019)**. This way, we hope to reduce the barriers to entry for experimenters who are not methodological experts.

7.1.2 Transfer Learning in fMRI

Transfer learning is a general concept in machine learning, which amounts to learning signal representations on a source task/dataset, and leveraging those representations on a target task/dataset (Weiss et al., 2016). In the MRI context, the source domain typically features abundant, high-quality labeled data: e.g. a large annotated dataset of general-purpose images such as `ImageNet` (Deng et al., 2009). In contrast, the target domain contains scarcer, sometimes noisier medical images, such as anatomical MRI T1/T2 scans used to detect dementia symptoms (Ardalan and Subbian, 2022; Valverde et al., 2021; Stangor and Walinga, 2014). In this context, the transferred representations are low-level visual features common

to all images: geometrical shapes and patterns. To conserve those representations, the first layers of a neural network are "frozen" —that is to say, not re-trained. On top of those, a trainable neural network is plugged and trained using data from the target domain (medical images), on the target task (e.g. dementia detection, or segmentation). Leveraging the general-purpose representations from the high-quality source domain, the neural network can achieve good performance on the target task even in the absence of a large target dataset.

In this work, we apply transfer learning to fMRI data, on parcellation and behavioral/cognitive score prediction tasks (Gao et al., 2019; Hejia Zhang et al., 2018). The long-term objective is to design methods capable of learning on a large, qualitative dataset of healthy subjects (e.g. the HCP (Van Essen et al., 2012)), and to transfer learned representations over a smaller, noisier clinical dataset (Mandonnet, 2011). In this context, the task would typically be the same across the source and target domains, but we expect an important *distribution shift* when moving from research to clinical data (Weiss et al., 2016). As an intermediate objective, we study here the **transfer across different subject groups in the HCP dataset.** (Van Essen et al., 2012; Hejia Zhang et al., 2018).

From a methodological standpoint, our objective is also to **explore the intersection between transfer learning and hierarchical Bayesian modeling** (Weiss et al., 2016; Suder et al., 2023). On the one hand, Bayesian inference over a source domain recovers latent parameter posterior distributions —see Section 3.1. Those distributions can be used as priors to perform inference over a target domain. The Bayesian formalism is thus naturally amenable to transfer learning. On the other hand, we also leverage parametric neural networks in our variational family —see Section 6.1.2. Those source-domain-trained networks can be re-used to infer over the target domain, which constitutes another instance of transfer learning. In this work, we explore combinations of both instances of transfer learning —Bayesian-centric and machine learning-centric— bridging a syntactic gap across communities (Suder et al., 2023).

7.2 Individual parcellation of the human cortex through hierarchical modeling and AVI

This section describes our methodology to tackle individual-level parcellation:

1. Preprocessing to compute vertex-level connectivity fingerprints;

2. Hierarchical modeling, inspired from the work of Kong, J. Li, et al. (2019), along with the inference method;

The main insight is that we tried to keep each step as minimal and principled as possible. In particular, we show the applicability of a general method (AVI) to infer effectively and automatically over large-scale, hierarchical models.

7.2.1 Preprocessing: extracting vertex-level connectivity fingerprints

We use data from the HCP dataset (Van Essen et al., 2012). We randomly select a cohort of $S = 1,000$ subjects from this dataset, each subject is associated with $T = 2$ resting-state fMRI sessions. Each session consists of 15 minutes-long volume time series. We minimally pre-process the signal using the `nilearn` library (Abraham et al., 2014) (high-variance confounds removal, detrending, band-pass filtering and spatial smoothing).

For every subject, we extract the surface BOLD signal of $N = 59,412$ vertices across the whole cortex. We compare this signal with the extracted signal of $D = 64$ dictionary of functional modes (DiFuMo) components: a dictionary of brain spatial maps allowing for an effective fMRI dimensionality reduction (Dadi et al., 2020). Specifically, we compute the one-to-one Pearson's correlation coefficient of every vertex with every DiFuMo component. The resulting connectome, with S subjects, T sessions, N vertices and a connectivity signal with D dimensions, is of shape $(S \times T \times N \times D)$. We project this data —correlation coefficients lying in $]-1; 1[$— in an unbounded space using an inverse sigmoid function.

The total signal $\mathbf{X}^{\text{population}} \in \mathbb{R}^{S \times T \times N \times D}$ is of massive dimensionality: 7.6 billion measures. How to perform inference in this large-scale context?

7.2.2 Modeling and inference: high-dimensional mixture with hierarchical vertex labeling

This chapter features two variants of HBM —illustrated in Figure 7.2— inspired by the work of Kong, J. Li, et al. (2019). Both models amount to a GM clustering over brain vertices, using a mixture of either subject-level or population-level components. We also apply in both models a spatial regularization that encourages vertices to have the same label across subjects.

Parcellation-hierarchical model (P) We hypothesize that every vertex in the cortex belongs to either one of $L = 7$ functional networks. Each network l corresponds to the connectivity fingerprint μ_l, represented as the correlation of the BOLD signal with the signal from the $D = 64$ DiFuMo components.

At the population level, for a vertex n we denote as $\textit{logits}_n \in \mathrm{R}^L$ the probability logits to belong to each network l. Each subject s is associated with the logits $\textit{logits}_{s,n}$, which are a perturbation of the population logits. This creates a regularization across subjects: a vertex n is encouraged to have the same label across all subjects. The variable γ_l controls the inter-subject spatial variability across all subjects, for the network l.

For every subject s, session t, and vertex n, we denote as $X_{s,t,n}$ the observed connectivity. This connectivity is modeled via a mixture model: $X_{s,t,n}$ is a perturbation of the connectivity fingerprint μ_l of the network corresponding to its label $\text{label}_{s,n}$. κ_l controls the variability between $X_{s,t,n}$ and the mixture component $\mu_{\text{label}_{s,n}}$.

The resulting model (P) model is summarized as:

$$
\begin{aligned}
S, T, N, D, L &= 1000, 2, 59412, 64, 7 \\
\forall l{=}1..L: \quad \mu_l &\sim \mathcal{N}(\vec{0}_D, \vec{6}_D) \\
\forall l{=}1..L: \quad \log \kappa_l &\sim \mathcal{N}(\vec{0}_L, \vec{1}_L) \\
\forall n{=}1..N: \quad \textit{logits}_n &\sim \mathcal{N}(\vec{0}_L, \vec{6}_L) \\
\forall l{=}1..L: \quad \log \gamma_l &\sim \mathcal{N}(\vec{0}_L, \vec{1}_L) \\
\forall s{=}1..S,\ \forall n{=}1..N: \quad \textit{logits}_{s,n} | \textit{logits}_n, [\gamma_l]_{l=1..L} &\sim \mathcal{N}(\textit{logits}_n, [\gamma_1 ... \gamma_L]) \\
\forall s{=}1..S,\ \forall n{=}1..N: \quad \text{labels}_{s,n} | \textit{logits}_{s,n} &\sim \text{Categorical}(\textit{logits}_{s,n}) \\
\forall s{=}1..S,\ \forall t{=}1..T,\ \forall n{=}1..N: \quad X_{s,t,n} | [\mu_l]_{l=1..L}, [\kappa_l]_{l=1..L}, \text{label}_{s,n} &\sim \mathcal{N}(\mu_{\text{label}_{s,n}}, \kappa_{\text{label}_{s,n}})
\end{aligned}
\tag{7.1}
$$

The model contains 4 plates: the *network* plate of cardinality L (that we did not exploit in our implementation), the *subject* plate of cardinality S, the *session* plate of cardinality T and the *vertex* plate of cardinality N.

Parcellation and connectivity-hierarchical model (P&C) This variant slightly modifies the (P) model by adding another differentiation across subjects. In addition to having specific parcellations, subjects are associated with specific connectivity fingerprints $\mu_{s,l}$. $X_{s,t,n}$ is no longer a mixture of the population components μ_l, but of the subject-specific components $\mu_{s,l}$.

The resulting model (P&C) model is summarized as:

$$
\begin{aligned}
S, T, N, D, L &= 1000, 2, 59412, 64, 7 \\
\forall_{l=1..L}: \quad \mu_l &\sim \mathcal{N}(\vec{0}_D, \vec{6}_D) \\
\forall_{l=1..L}: \quad \log \epsilon_l &\sim \mathcal{N}(\vec{0}_L, \vec{1}_L) \\
\forall_{\substack{l=1..L \\ s=1..S}}: \quad \mu_{s,l} | \mu_l, \epsilon_l &\sim \mathcal{N}(\mu_l, \epsilon_l) \\
\forall_{l=1..L}: \quad \log \kappa_l &\sim \mathcal{N}(\vec{0}_L, \vec{1}_L) \\
\forall_{n=1..N}: \quad \textit{logits}_n &\sim \mathcal{N}(\vec{0}_L, \vec{6}_L) \\
\forall_{l=1..L}: \quad \log \gamma_l &\sim \mathcal{N}(\vec{0}_L, \vec{1}_L) \\
\forall_{\substack{s=1..S \\ n=1..N}}: \quad \textit{logits}_{s,n} | \textit{logits}_n, [\gamma_l]_{l=1..L} &\sim \mathcal{N}(\textit{logits}_n, [\gamma_1 ... \gamma_L]) \\
\forall_{\substack{s=1..S \\ n=1..N}}: \quad \text{labels}_{s,n} | \textit{logits}_{s,n} &\sim \text{Categorical}(\textit{logits}_{s,n}) \\
\forall_{\substack{s=1..S \\ t=1..T \\ n=1..N}}: \quad X_{s,t,n} | [\mu_{s,l}]_{l=1..L}, [\kappa_l]_{l=1..L}, \text{label}_{s,n} &\sim \mathcal{N}(\mu_{s,\text{label}_{s,n}}, \kappa_{\text{label}_{s,n}})
\end{aligned}
\tag{7.2}
$$

Across both models, our goal is to recover the posterior distribution of the networks' fingerprints μ and the labels label given the observed connectome described in Section 7.2.1. To this end, we use a variant of the PAVI-F scheme —described in Section 6.2.3. Every RV is associated with a plate-amortized NF, except the second-order parameters $\epsilon_l, \kappa_l, \gamma l$, associated with MAP regressors. To allow for the optimization over the discrete $\text{label}_{s,n}$ RV, we used the Gumbell-Softmax trick as described in Section 3.3.4 (Grathwohl et al., 2018).

7.3 Applications in large-scale fMRI

7.3.1 Fullcortex parcellation over 1,000 subjects from the HCP (P model)

In a first experiment, we demonstrate the ability of AVI to tackle a large-scale Neuroimaging inference problem. We apply the hierarchical inference method described in Section 7.2 over the connectivity from $S = 1,000$ subjects from the HCP (Van Essen et al., 2012). This represents a signal containing 7.6 billion measures. For each of the $N = 59,412$ vertices of every subject, we infer a probabilistic label to belong to one of the $L = 7$ functional networks. This amounts to inferring over 400 million latent parameters.

Results are visible in Figure 7.3. Using PAVI, *in under 5 hours of GPU time*, we recover a smooth population parcellation that replicates some of the major networks

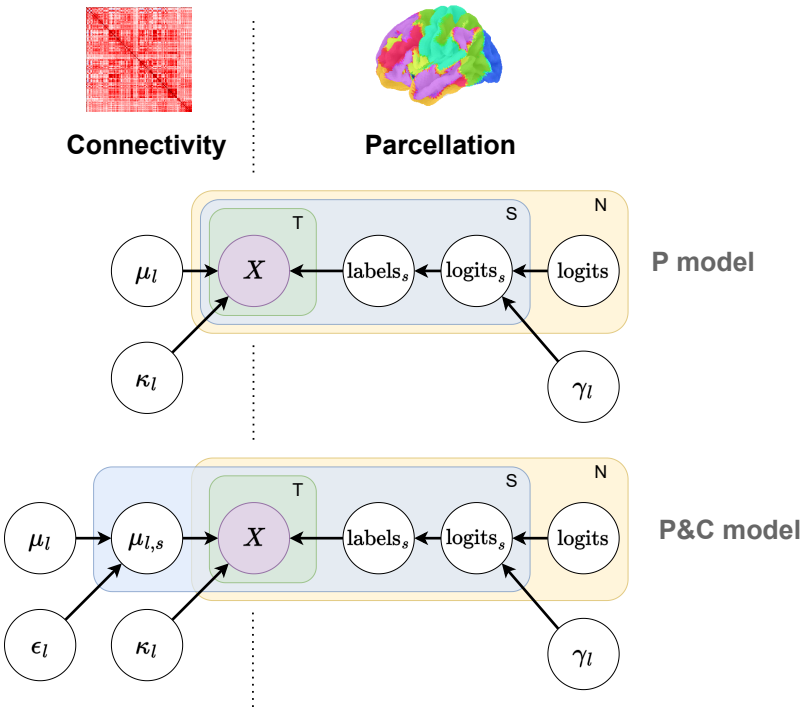

Fig. 7.2.: **Parcellation (P) and (P&C) HBMs** Models are composed of two parts. On the left, we represent the connectivity fingerprints μ associated with the mixture components. On the right, we represent the parcellations denoting where each network is expressed on the cortex. In the (P) model in Equation (7.1), only the parcellation part is hierarchical —via the subject-specific logits. In the (P&C) model in Equation (7.2), subjects are also associated with individual connectivity fingerprints.

from Figure 7.1. Subject parcellations are noisier perturbations of this population map.

There is a large uncertainty in the labeling of a given vertex —based solely on $T = 2$ resting-state connectivity fingerprints. To represent this uncertainty, we display a probabilistic parcellation: incertain labeling is associated with bland colors, while certain labeling is associated with vivid color. This probabilistic parcellation underlines the richness of the Bayesian framework, which features an interpretable notion of uncertainty.

7.3.2 Out-of-sample cognition and behavior prediction based on the probabilistic parcellation (P model)

This experiment tests out the predictive power of individual parcellations on behavior and cognition. Can the general location of function inside the brain inform us about the cognitive ability of a subject?

Method We reproduce the methodology from Kong, J. Li, et al. (2019). After inference, we use as feature the subject-specific parcellation logits and perform a cross-validation across subjects to predict out-of-sample cognitive and behavioral scores. We start from the $logits_{s,n}$ associated with each subject. We use PCA (33% explained variance) followed by a linear regression to predict each of the 13 cognitive scores. We report the test performance on the test fold averaged across the 13 cognitive measures. This process is reproduced 100 times. The reported scores are the triple average across folds, measures, and repetition, while the standard deviation is computed across the 100 repetitions only.

Results Table 7.1 shows our prediction performance, on par with the state-of-the-art (Kong, J. Li, et al., 2019; Calhoun and Adali, 2012; E. M. Gordon et al., 2017; Danhong Wang et al., 2015). This demonstrates that our parcellations —recovered using minimal preprocessing, and a principled AVI methodology—- are relevant to cognition.

Stronger baseline To put our results in perspective, Kong, Q. Yang, et al. (2021) recently extended the work from Kong, J. Li, et al. (2019). Authors notably illustrated the positive impact of individual parcellations over the extraction of functional connectomes (as defined in Chapter 9). On a behavioral scores prediction task,

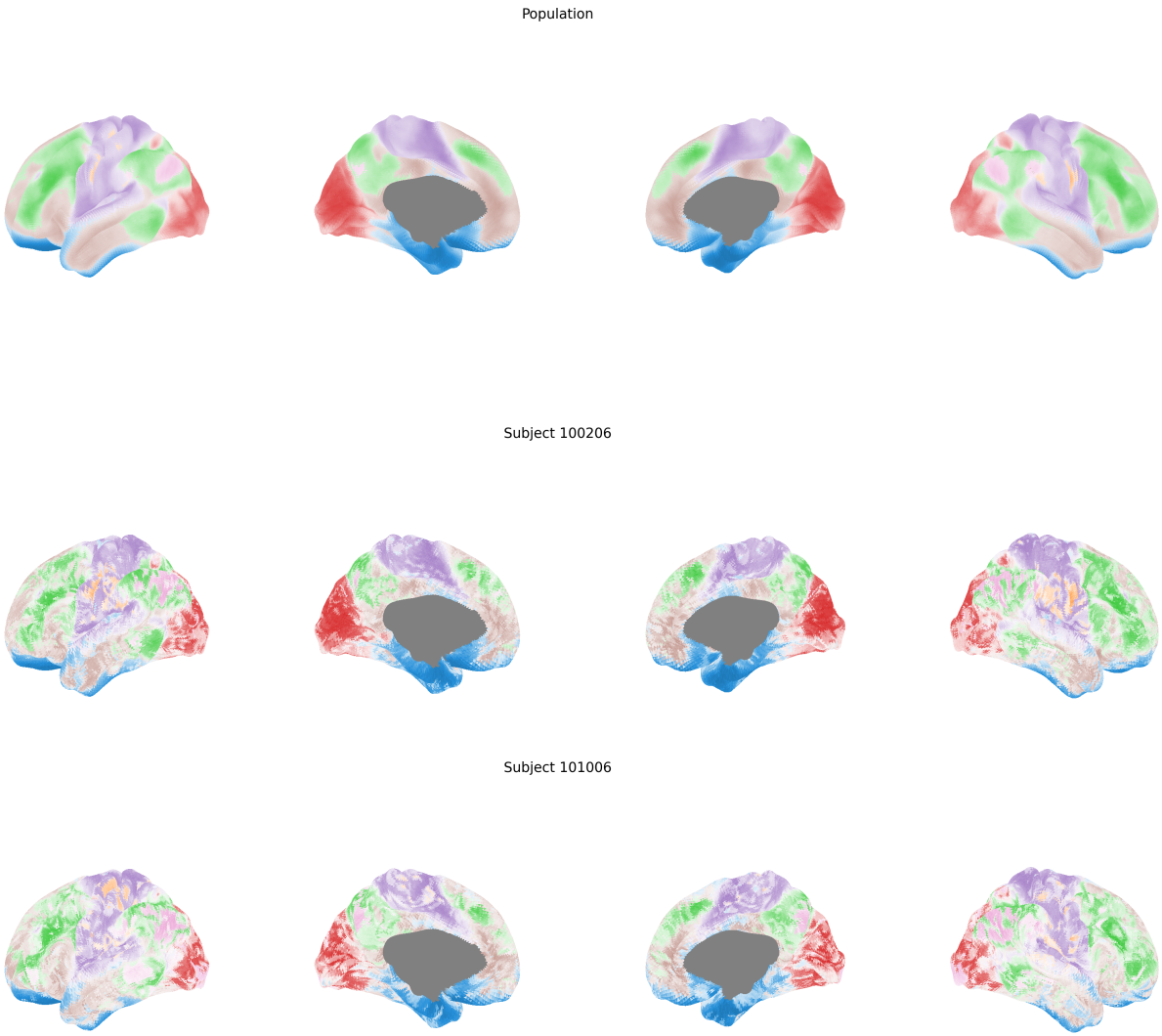

Fig. 7.3.: Probabilistic full cortex parcellation For a cohort of $1,000$ subjects, 2 of which are represented here (in the bottom 2 lines) we cluster $60,000$ cortex vertices according to their connectivity with the rest of the brain. We show the obtained probabilistic *parcellations*. Each color in the parcellation corresponds to one of 7 functional *network* (Thomas Yeo et al., 2011). Networks represent groups of neurons that co-activate in the brain and can be associated with certain cognitive functions, such as vision or motor control. Through our method, we recover each subject's parcellation (at the bottom), which are i.i.d. perturbations of the population's parcellation (at the top). Our method also models uncertainty: coloring represents the dominant label for each vertex and the level of white increases with the uncertainty in the labeling.

Tab. 7.1.: **Individual probabilistic parcellation predicts a subject's cognition and behavior** We can use the subject parcellations as features for a cognitive score prediction task. The table shows the mean predictive accuracy across 13 cognitive measures, including memory, pronunciation, processing speed or spatial orientation. The baseline methods scores are reproduced from Kong, J. Li, et al. (2019)'s implementation. Our method produces individual maps that are predictive of the subject's cognitive ability. *Reported accuracy is the correlation of the predicted scores with the true score. Performance is averaged over 1,000 subjects in our method, versus only 881 in the implementation from Kong, J. Li, et al. (2019). This can in part explain our higher performance, as per the learning curve featured in Section 7.3.3*

Method	Accuracy for 13 cognitive measures (Higher is better)
MS-HBM from Kong, J. Li, et al. (2019)	0.1321 ($\pm$0.0053)
YeoBackProject from Calhoun and Adali (2012)	0.1057 ($\pm$0.0060)
Gordon2017 from E. M. Gordon et al. (2017)	0.0545 ($\pm$0.0062)
Wang2015 from Danhong Wang et al. (2015)	0.1202 ($\pm$0.0054)
Ours (PAVI)	0.1645 ($\pm$0.0047)

authors reported a correlation of 18% over a cohort of 1,200 subjects from the HCP (as opposed to 1,000 in our case), and a set of 36 measures (as opposed to 13 in our case) (Van Essen et al., 2012). Kong, Q. Yang, et al. (2021) thus constitutes a stronger baseline than the results presented in Table 7.1 (but a weaker baseline than our coupling-based results from Figure 9.12).

7.3.3 Extension: Bayesian transfer learning yields more informative parcellation in the small-sample regime (P&C model)

In this extension, we consider transfer learning across different subject groups in the HCP as a way to stabilize cognitive scores prediction. We consider successive inference over two subject sets:

- a large training set, containing 750 subjects. Over this set, inference is performed "from scratch".

- a smaller test set, containing up to 200 subjects. Optionally, we transfer representations from the training set to help with inference.

Method To evaluate the effect of transfer learning, we consider three experimental setups:

1. **No transfer learning** We transfer no representation from the training to the test subject sets. We use a random initialization for the variational family weights —the PAVI-Fs flows and encodings as described in Sections 6.1.2 and 6.2.3. We then perform inference using the same weak priors that were used for the training set;

2. **Machine learning warm start** we use the trained population-level variational family weights as an initialization. We then perform inference using the same weak priors that were used for the training set;

3. **Warm start + Bayesian posterior predictive transfer** we use the trained population-level variational family weights as an initialization. In addition, we use the training set posteriors (of the non-subject-specific RVs) as priors to infer over the subject test set. This corresponds to the RVs $\mu_l, \epsilon_l, \kappa_l, \textit{logits}_n, \gamma_l$ described in Equation (7.2).

Results and discussion From a theoretical point of view, only the Bayesian transfer (in step 3) is supposed to improve inference. The machine learning warm start is a better initialization for the r-KL optimization problem —as described in Section 3.3.3. But the warm start does *not* modify the optimized ELBO loss, nor the theoretical minimum of this loss —that depends solely on the HBM, the observed signal, and the variational family. In contrast, by modifying the priors in the HBM, the Bayesian transfer modifies the loss minimum by injecting some precision from the training set. Results are visible in Figure 7.4. The machine learning warm start (step 2) appears as an effective strategy, while the effect of the Bayesian transfer (step 3) is only marginal. Those results are the opposite of what theory would have predicted! This suggests that **in the small sample regime, parcellation is a complex optimization problem, whose loss minimum is (statistically) not reached**. We interpret the warm start as a practical simplification of optimization, which helps with downstream task performance. In contrast, the added value of the Bayesian transfer is more theoretical, and its effect appears marginal.

Those results motivate the development of robust inference methods, as further discussed in Section 10.4.

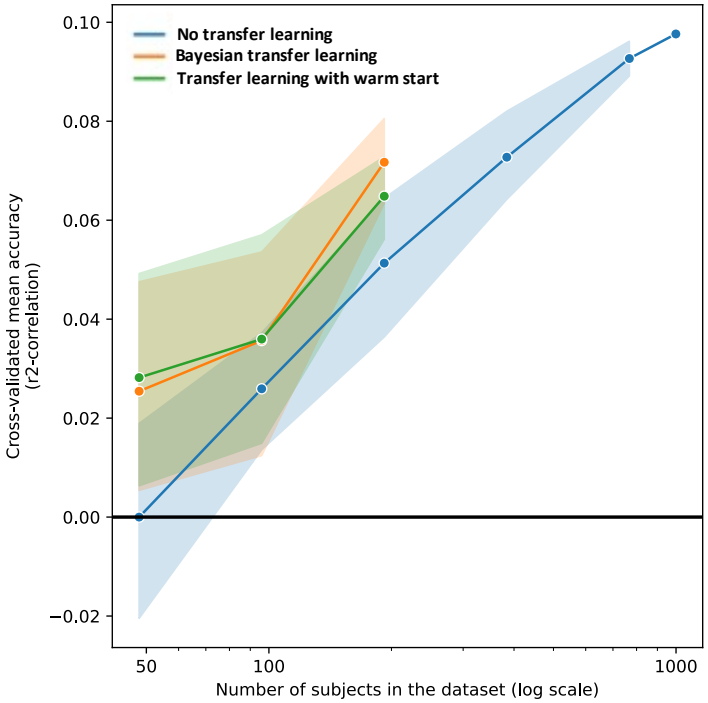

Fig. 7.4.: Effect of transfer learning on cognitive scores prediction The setups 1, 2 and 3 described in Section 7.3.3 are respectively represented in blue, red and yellow. The blue line shows a logarithmic learning curve: training over larger subject groups, we improve out-of-sample cognitive scores prediction. To plot the blue line, only information from the validation set subjects is considered. In contrast, on the red and yellow lines, information is transferred from a training set of 750 subjects. This results in higher prediction performance at equal validation set size. The machine learning warm start (from blue to red) significantly improves cognitive scores prediction. In contrast, the Bayesian transfer (red to yellow) only marginally improves performance. *Performance averaged across 10 population bootstraps (with a full population size represented on the x-axis.) Shaded areas represent 95% confidence intervals. Cognitive scores performance is computed following the same steps as in Section 7.3.2.*

7.3.4 Preliminary: robustifying individual parcellations (P&C model)

In this section, we showcase more recent results from our ongoing work on parcellations. This rich application setting allows us to test out methods to robustify inference at scale. In particular, compared to Section 7.3.1, we test out:

- Augmenting the number of parcels to $L = 17$, the number of sessions to $T = 4$, and the dimensionality of the connectivity fingerprints to $D = 128$;
- Using simplified amortized variational distributions, including Gaussian approximations;
- Using a KL annealing schedule, transitioning from a ML to an ELBO loss (Krishnan et al., 2017);
- Regularizing inference through the encodings $\mathbf{E}$'s structure (see Sections 6.1.2 and 6.2.3). In the original PAVI-F design, we associate every (subject,vertex) pair to an individual encoding, resulting in a full-rank encoding tensor $\mathbf{E}^{S\times N}$ —with $\propto (S \times N)$ parameters. In contrast, here we combine subject-specific encodings $\mathbf{E}^{S}$ with vertex-specific encodings $\mathbf{E}^{N}$, resulting in a low-rank encoding tensor $\mathbf{E}^{S\times N} = \mathbf{E}^{S} \otimes \mathbf{E}^{N}$ —with $\propto (S + N)$ parameters. This encoding scheme is lighter in memory and favors smoother parcellations across subjects.

Figure 7.5 illustrates two example individual parcellations. Both subdivide the cortex in a similar way to the 17 networks illustrated in Figure 7.6 (Thomas Yeo et al., 2011).

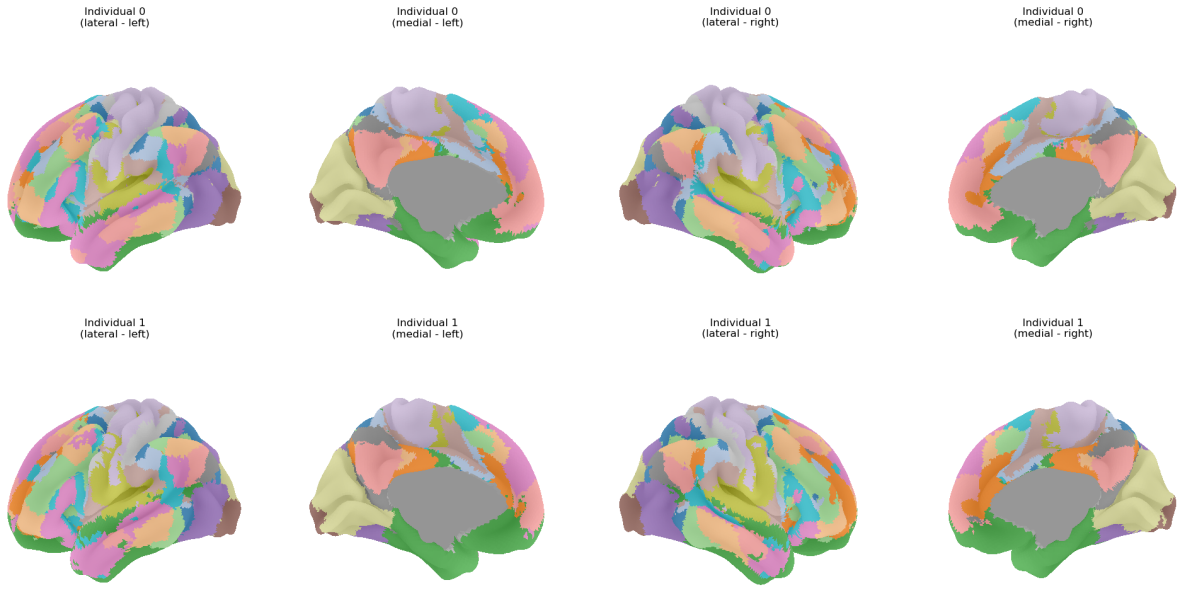

Fig. 7.5.: **Stabilized individual hard parcellations** Thanks to various improvements described in Section 7.3.4, we obtain more spatially stable individual parcellations. Fractionations are similar to the ones reported by Thomas Yeo et al. (2011), as visible in Figure 7.6.

7.4 Summary of contributions

This chapter summarized our contributions in fMRI-based parcellation:

- We illustrate the applicability of AVI to large-scale Neuroimaging problems. Compared to pen-and-paper EM algorithms, AVI:
  1. requires little technical mastery,
  2. does not limit the choice of HBM,
  3. could potentially reduce technical barriers to entry for experimenters.
- We design original HBMs, taking as a basis the work from Kong, J. Li, et al. (2019). We propose models that incorporate a spatial regularization across subjects, jointly obtaining individual and population-level parcellations.
- We map the brain of a thousand individuals via probabilistic full-cortex parcellations. Those parcellations naturally incorporate a notion of uncertainty, leveraging the richness of the Bayesian probabilistic framework.

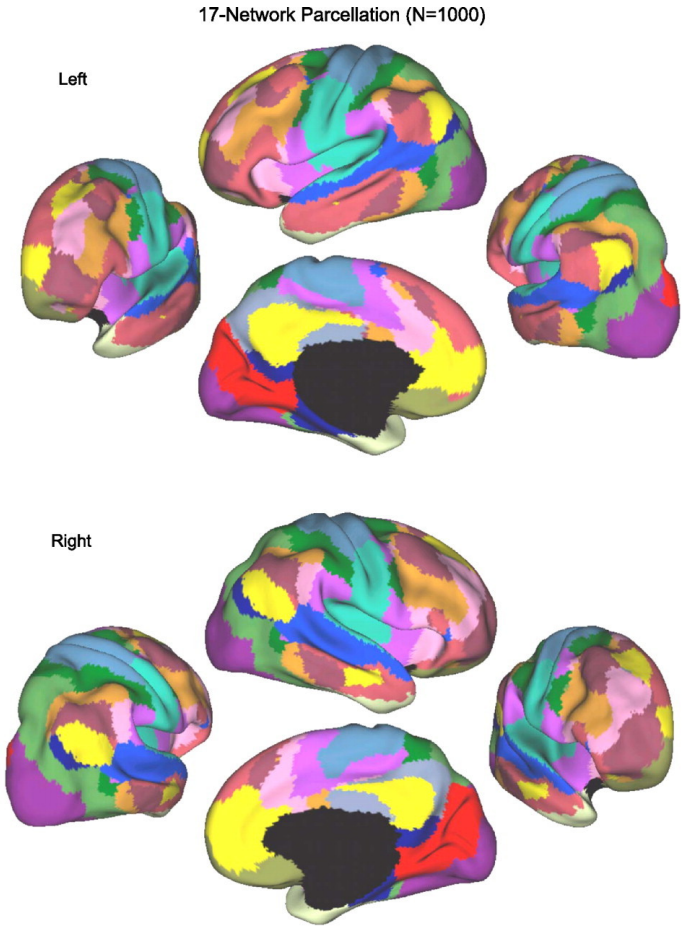

Fig. 7.6.: Yeo 17 functional networks (Thomas Yeo et al., 2011) Compared to the parcellation in 7 networks in Figure 7.1, the cortex is sub-divided into finer parcels. We use this parcellation as a reference for Figure 7.5 *Figure adapted from Thomas Yeo et al. (2011).*

- We demonstrate the predictive power of those parcellations over cognition and behavior, obtaining scores relevant to the state-of-the-art (Kong, Q. Yang, et al., 2021)
- We pave the way for transfer learning as a means to stabilize individual parcellations in potentially degraded datasets.

8 Application: reducing uncertainty in tissue microstructure estimation via hierarchical modeling

This work tackles tissue microstructure estimation, an application of dMRI presented in Section 1.3. The goal is to infer the statistical cellular composition of brain voxels from a dMRI signal. As introduced in Section 1.3, this application suffers from large uncertainty. In this chapter, we combine SBI with hierarchical modeling to reduce this uncertainty. We inject a distribution surrogate inside a HBM and infer over this composite model using the methods from Chapter 6. By combining the information across multiple voxels in the brain, we can infer the parameters in each voxel with greater precision.

First, we present in detail tissue microstructure estimation. We then present the hierarchical modeling and inference method we use to tackle this task. Finally, we present some results, notably in epileptic lesion segmentation.

8.1 Tissue microstructure estimation

This section presents in greater detail tissue microstructure estimation, which we introduced in Section 1.3. Microstructure estimation relies on dMRI, which measures the diffusion of water molecules inside tissues (Alexander et al., 2019; Jelescu, Palombo, et al., 2020). This diffusion can be impeded by two elements:

- cellular membranes, which separate the different cells in brain tissue from the extracellular space. In this application, we model those membranes as impermeable (Jelescu, Skowronski, et al., 2022);
- other molecules, proteins and large metabolites, which crowd the space inside and outside cells. This is summarized via the space's *diffusivity*, denoted using the letter D with unit $\mu m^2/s$.

To simplify this complex phenomenon, and link it to interpretable parameters, various microstructure models have been posited by experts (Hui Zhang et al., 2012; Novikov et al., 2019; Palombo et al., 2020). This work focuses on one such model, the **standard model (SM)** (Novikov et al., 2019), which has been designed for white matter and simplifies tissues into two compartments:

1. **linear fibers**, representing a fraction f of the total signal in a voxel. Inside those fibers, it is assumed that water can only diffuse *parallel* to their direction, with a diffusivity D_a. This amounts to modeling fibers as sticks. A fiber segment (composed of multiple fibers) is oriented in a given direction. A voxel features multiple segments, whose directions can be more or less aligned, as measured by the orientation dispersion index (ODI);

2. the **extracellular space**, representing a fraction $1 - f$ of the total signal in a voxel. Outside the cells, water can diffuse both *parallel* and *perpendicular* to the fibers, with two distinct diffusivities $D_{e\parallel}$ and $D_{e\perp}$.

The SM is illustrated in Figure 8.1.

In summary, we infer 5 parameters in each voxel $n = 1..N$, collectively denoted as $\theta_n^{\text{vox}} \in \mathbb{R}^5$:

- the neurite fraction f;
- the intra-neurite diffusivity D_a;
- the ODI;
- the parallel diffusivity within the extra-neurite space $D_{e\parallel}$;
- the perpendicular diffusivity within the extra-neurite space $D_{e\perp}$.

To infer those parameters, we observe the dMRI signal $X_n^{\text{vox}} \in \mathbb{R}^D$ coming from each voxel. This signal is composed of D measures, one measure per magnetic field direction and intensity —where the intensity is called the b-value (Le Bihan and Breton, 1985).

How to infer the tissue microstructure parameters $\theta^{\text{vox}} \in \mathbb{R}^{N\times 5}$ from the observed signal $X^{\text{vox}} = \mathbf{X}^{\text{subject}} \in \mathbb{R}^{N\times D}$?

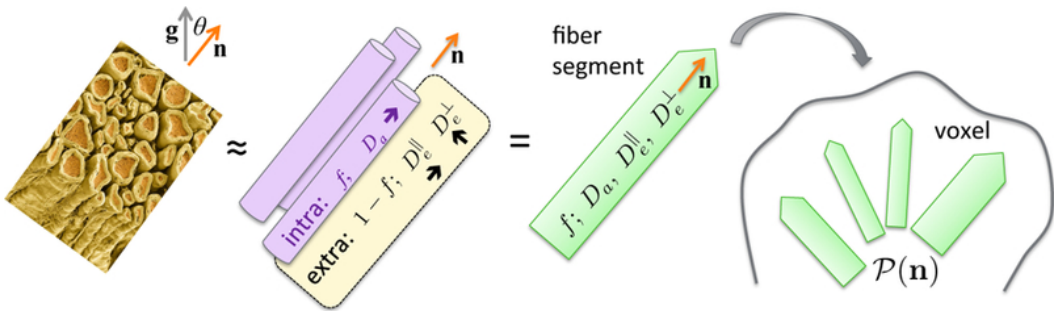

Fig. 8.1.: The standard model (SM) in dMRI microstructure estimation Tissues are approximated into two compartments —fibers and the extracellular space— in which water can diffuse with diffusivity D. The ODI is linked to the orientation distribution function $\mathcal{P}$. *Figure reproduced from Novikov et al. (2019)*

8.2 Hierarchical μ-GUIDE: combining SBI, hierarchical modeling and PAVI

Our method is composed of three successive steps:

1. A **SBI** phase, where we learn a surrogate to the voxel-independent microstructure parameters posterior (Section 8.2.1);

2. A **hierarchical modeling** phase, where we inject this distribution surrogate inside a hierarchical model to reduce the microstructure parameter uncertainty using a meaningful prior (Section 8.2.2);

3. An **inference** phase, where we leverage PAVI to infer effectively over the composite model (Section 8.2.3).

8.2.1 μ-GUIDE: learning an independent-case posterior surrogate

To learn the statistical link between the microstructure parameters and the observed dMRI signal, we rely on a **simulator** (Ianuş et al., 2017). This simulator takes as input voxel parameters θ_i^{vox}, simulates simplified geometrical fibers, and computes a noisy synthetic signal X_i^{vox} using analytical formulas (which is possible because of the hypothesized simplified geometry). We used Rician noise, with a signal-to-noise ratio (SNR) of 50. Repeating this simulation procedure, using a uniform prior defined on physically plausible ranges for the θ_i^{vox}, we obtain a synthetic dataset $\{\theta_i^{\text{vox}}, X_i^{\text{vox}}\}_{i=1..N}$.

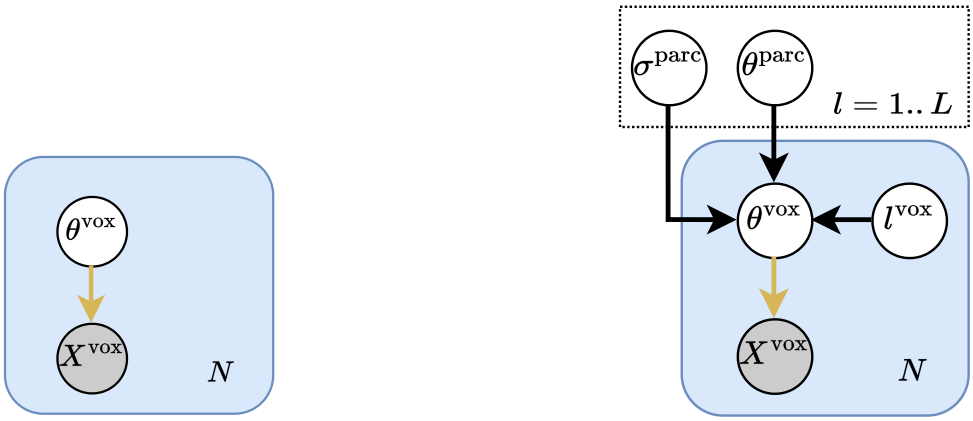

Fig. 8.2.: Microstructure estimation graphs On the left, we illustrate the voxel-independent graph corresponding to the posterior surrogate learning in Section 8.2.1. On the right, we illustrate the latent mixture graph corresponding to the hierarchical modeling in Section 8.2.2. In both graphs, the yellow edge from θ^{vox} to X^{vox} symbolizes a learned surrogate distribution.

Using this synthetic dataset, we learn a surrogate to the posterior distribution $P_{\text{ind}}(\theta_i^{\text{vox}}|X_i^{\text{vox}})$. P_{ind} is indexed "ind" to mark that this posterior corresponds to the case when voxels are considered as **independent** (as opposed to dependent through a hierarchical prior as in Section 8.2.2). This independent case is illustrated by the graph on the left in Figure 8.2. To learn $P_{\text{ind}}(\theta_i^{\text{vox}}|X_i^{\text{vox}})$, we use NPE —as described in Section 4.2.2 (Papamakarios and Murray, 2016). We combine an encoder with a MAF to learn a sample amortized posterior surrogate (Papamakarios, Pavlakou, et al., 2017).

Using this surrogate on the subject data, we obtain an independent-case posterior: $P_{\text{ind}}(\theta^{\text{vox}}|X^{\text{vox}} = \mathbf{X}^{\text{subject}})$. This methodology has been dubbed μ-GUIDE by Jallais and Palombo (2023) (this work is a continuation of μ-GUIDE). The μ-GUIDE posterior combines two elements:

1. for each voxel n, an encoding $\mathbf{E}_n^{\text{ind}} = f(\mathbf{X}_n)$ resulting from the application an encoder f to the voxel's dMRI signal $\mathbf{X}_n$;

2. a trained NF $\mathcal{F}^{\text{ind}}(\bullet; \phi^{\text{ind}}, \mathbf{E}_\bullet^{\text{ind}})$ that approximates the conditional posterior distribution of a voxel's parameters given that voxel's signal's encoding.

We use this independent-case posterior as a baseline, but also as an initialization for PAVI, as detailed in Section 8.2.3. Due to voxels being considered independently, this posterior features a large variance. Our goal is to reduce that variance.

8.2.2 Switching to a hierarchical mixture prior

To reduce the uncertainty in the voxel-wise estimates, we switch from an independent prior to a hierarchical prior (Powell et al., 2021; Orton et al., 2014). In doing so, we remove the assumption of independence across voxels, by injecting the original hypothesis/bias:

> "It is unlikely that voxels feature wildly different tissue microstructures across the brain, that would span over the entirety of the Θ-space. Most likely, only a **few microstructure configurations** exist in the brain, corresponding to canonical tissue types such as gray matter, white matter, or the ventricles. Each voxel can be associated with one of those sub-types, and its **microstructure is likely similar to the other voxels of that sub-type**."

To implement that hypothesis, we use a **hierarchical mixture prior**. We refer to the different tissue types as "parcels", because their spatial localization will partition the brain into different macroscopic tissues.

We group the brain voxels into L parcels of similar microstructure. Each parcel $l = 1..L$ is associated with the average parameters θ_l^{parc}. The microstructure θ_n^{vox} of each voxel $n = 1..N$ is assumed to be a perturbation of the parcel parameters it belongs to, with variability σ^{parc}. We denote l_n^{vox} the label of the voxel n, that is to say, the parcel it belongs to. We denote $\textit{logits}_n^{\text{vox}}$ the probability logits corresponding to l_n^{vox}. We can summarize the new HBM $P_{\text{hier}}(X^{\text{vox}}, \theta^{\text{vox}}, l^{\text{vox}}, \textit{logits}_n^{\text{vox}}, \theta^{\text{parc}})$ with the following equations:

$$
\begin{aligned}
\forall l = 1..L \quad & \theta_l^{\text{parc}} \sim \mathcal{U}(\theta_{\text{min}}, \theta_{\text{max}}) \\
\forall n = 1..N \quad & \textit{logits}_n^{\text{vox}} \sim \mathcal{N}(\vec{0}_L, \vec{0.5}_L) \\
\forall n = 1..N \quad & l_n^{\text{vox}} | \textit{logits}_n^{\text{vox}} \sim \text{Categorical}(\textit{logits}_n^{\text{vox}}) \\
\forall n = 1..N \quad & \theta_n^{\text{vox}} | \{\theta_l^{\text{parc}}\}_{l=1..L}, l_n^{\text{vox}} \sim \mathcal{N}(\theta_{l_n^{\text{vox}}}^{\text{parc}}, \sigma_{l_n^{\text{vox}}}^{\text{parc}}) \\
\forall n = 1..N \quad & X_n^{\text{vox}} | \theta_n^{\text{vox}} \sim P(X^{\text{vox}} | \theta^{\text{vox}})
\end{aligned}
\tag{8.1}
$$

where $\mathcal{U}(\theta_{\text{min}}, \theta_{\text{max}})$ is a uniform prior over physically-possible microstructure parameters. $\mathcal{N}(\vec{0}_L, \vec{0.5}_L)$ translates into an equiprobable possibility of a voxel to belong to each parcel. σ^{parc} is a hyper-parameter to more or less strongly uniformize voxel distributions inside each parcel. $P(X^{\text{vox}}|\theta^{\text{vox}})$ corresponds to the likelihood *implicitly* learned in Section 8.2.1. Critically, $P(X^{\text{vox}}|\theta^{\text{vox}})$ **does not depend on the choice of prior for θ^{vox}** —as put forward in Section 4.2.2. We clarify this point in Section 8.2.3.

The resulting composite HBM —injecting a learned surrogate inside an explicit model— in Equation (8.1) is illustrated in Figure 8.2 (right) (Glöckler et al., 2022).

Technical notes on modeling:

- A non-parametric setting could be used to derive the number of parcels L from the data. For now, we default to a conservative dozen parcels, with only a few significantly expressed through the tissue.
- A non-parametric setting could also be used for σ^{parc}, to derive the level of variability inside parcels from the data. Implementing this however resulted in a mode collapse for σ, and variance underestimation for θ^{parc}. As a result, we chose to keep σ^{parc} as a user-defined constant, setting the value from synthetic experiments with known ground truth.

Inferring over the HBM in Equation (8.1), we jointly perform two tasks:

1. **learn a parcellation** of the brain tissue, by inferring the labels l^{vox};
2. **reduce the uncertainty** in each voxel's microstructure estimation θ^{vox} by sharing the information across voxels in the same parcel.

as illustrated in Figure 8.3. Having switched the prior, we infer the new *hierarchical* posterior: $P_{\text{hier}}(\theta^{\text{vox}}, l^{\text{vox}}, \mathit{logits}^{\text{vox}}, \theta^{\text{parc}} | X^{\text{vox}})$. Combining hierarchical modeling with μ-GUIDE, we dub our method **hierarchical μ-GUIDE** (Rouillard, Wassermann, et al., 2024).

Note: the same methodology could be applied to a **different hierarchical prior**, to implement a different hypothesis/bias. As an example, consider a Markov random field that would smooth the posterior distributions of neighboring voxels.

8.2.3 Inference using PAVI

Variational family architecture As described in Algorithm 5, we use a modified MF version of the PAVI-F design described in Sections 6.1.2 and 6.2.3:

- the posterior for the θ_n^{vox} and θ_l^{parc} with $l = 1..L$ are approximated via a combination of a plate-amortized NFs and encodings;
- we infer the parcellation $\mathit{logits}_n^{\text{vox}}$ using parametric Gaussians, enumerating over the discrete labels l_n^{vox} —as described in Section 3.3.4.

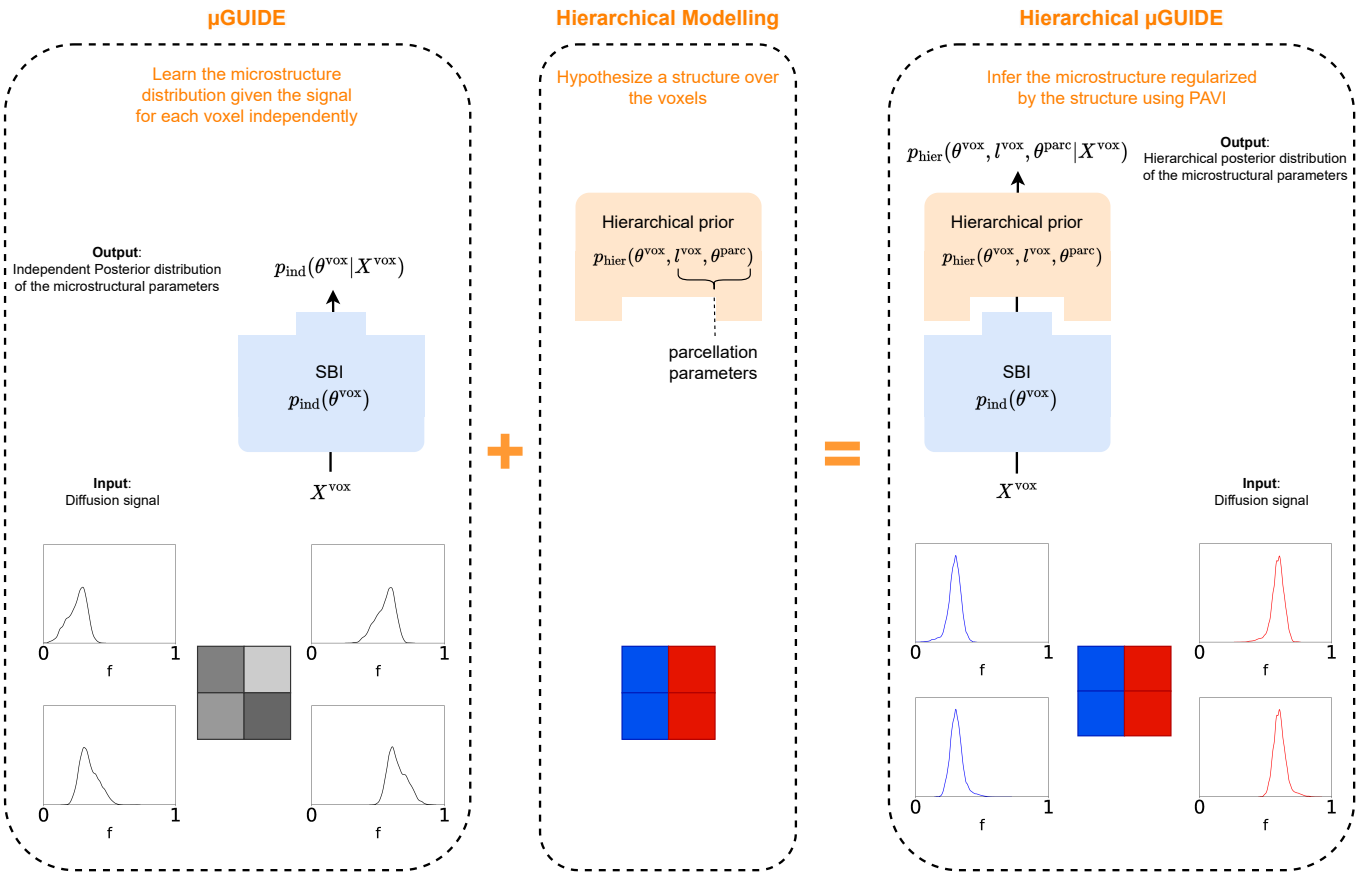

Fig. 8.3.: Hierarchical μ-GUIDE working principle μ-GUIDE (left) corresponds to the application of NPE to the microstructure inference problem (Jallais and Palombo, 2023; Papamakarios and Murray, 2016). μ-GUIDE yields voxel-independent posteriors, with a large variance (bottom). We combine μ-GUIDE with a hierarchical prior (middle), grouping voxels into parcels of similar microstructure. Using this meaningful prior, the resulting hierarchical μ-GUIDE (right) reduces microstructure parameter uncertainty.

Algorithm 5: Hierarchical-μ-GUIDE variational family initialization

Input:

Number of voxels N
Number of parcels L
Trained independent-case conditional flow $\mathcal{F}^{\text{ind}}(\bullet; \phi^{\text{ind}}, \mathbf{E}^{\text{ind}}_{\bullet})$
foreach $n = 1..N$
  Trained independent-case signal encoding $\mathbf{E}^{\text{ind}}_n = f(\mathbf{X}_n)$

Output:

Initial hierarchical-case voxel conditional flow $\mathcal{F}^{\text{vox}}(\bullet; \phi^{\text{vox}}, \mathbf{E}^{\text{vox}}_{\bullet})$
foreach $n = 1..N$
  Initial hierarchical-case voxel encoding $\mathbf{E}^{\text{vox}}_n$
  Initial voxel logits $\mathit{logits}^{\text{vox}}_n$
Initial parcel conditional flow $\mathcal{F}^{\text{parc}}(\bullet; \phi^{\text{parc}}, \mathbf{E}^{\text{parc}}_{\bullet})$
foreach $l = 1..L$
  Initial parcel encoding $\mathbf{E}^{\text{parc}}_l$

Algorithm:

Initialize $\mathcal{F}^{\text{vox}}$ with the same architecture as $\mathcal{F}^{\text{ind}}$ and set $\phi^{\text{vox}} = \phi^{\text{ind}}$
for $n = 1..N$ **do**
  Initialize $\mathbf{E}^{\text{vox}}_n = \mathbf{E}^{\text{ind}}_n$
  Define the independent-case voxel parameter posterior $Q^{ind}_n(\theta^{\text{vox}}_n)$ as the push-forward distribution of the flow $\mathcal{F}^{\text{ind}}(\bullet; \phi^{\text{ind}}, \mathbf{E}^{\text{ind}}_n)$
  Compute the voxel parameter expected value: $\tilde{\theta}^{\text{vox}}_n = \frac{1}{K} \sum_{\theta^{\text{vox}}_{n,k} \sim Q^{ind}_n} \theta^{\text{vox}}_{n,k}$
Run a GM EM algorithm over the $\{\tilde{\theta}^{\text{vox}}_n\}_{n=1..N}$ to obtain the initial labels $\{l^{\text{vox,EM}}_n\}_{n=1..N}$
for $n = 1..N$ **do**
  Initialize $\mathit{logits}^{\text{vox}}_n$ as the one-hot of $l^{\text{vox,EM}}_n$
Initialize $\mathcal{F}^{\text{parc}}$ with the same architecture as $\mathcal{F}^{\text{ind}}$ and set $\phi^{\text{parc}} = \phi^{\text{ind}}$
for $l = 1..L$ **do**
  Initialize the parcel's encoding to the mean of the voxels' encodings that belong to it: $\mathbf{E}^{\text{parc}}_l = \frac{1}{\#_{\text{parcel l}}} \sum_{\text{voxel n with } l^{\text{vox,EM}}_n = l} \mathbf{E}^{\text{vox}}_n$

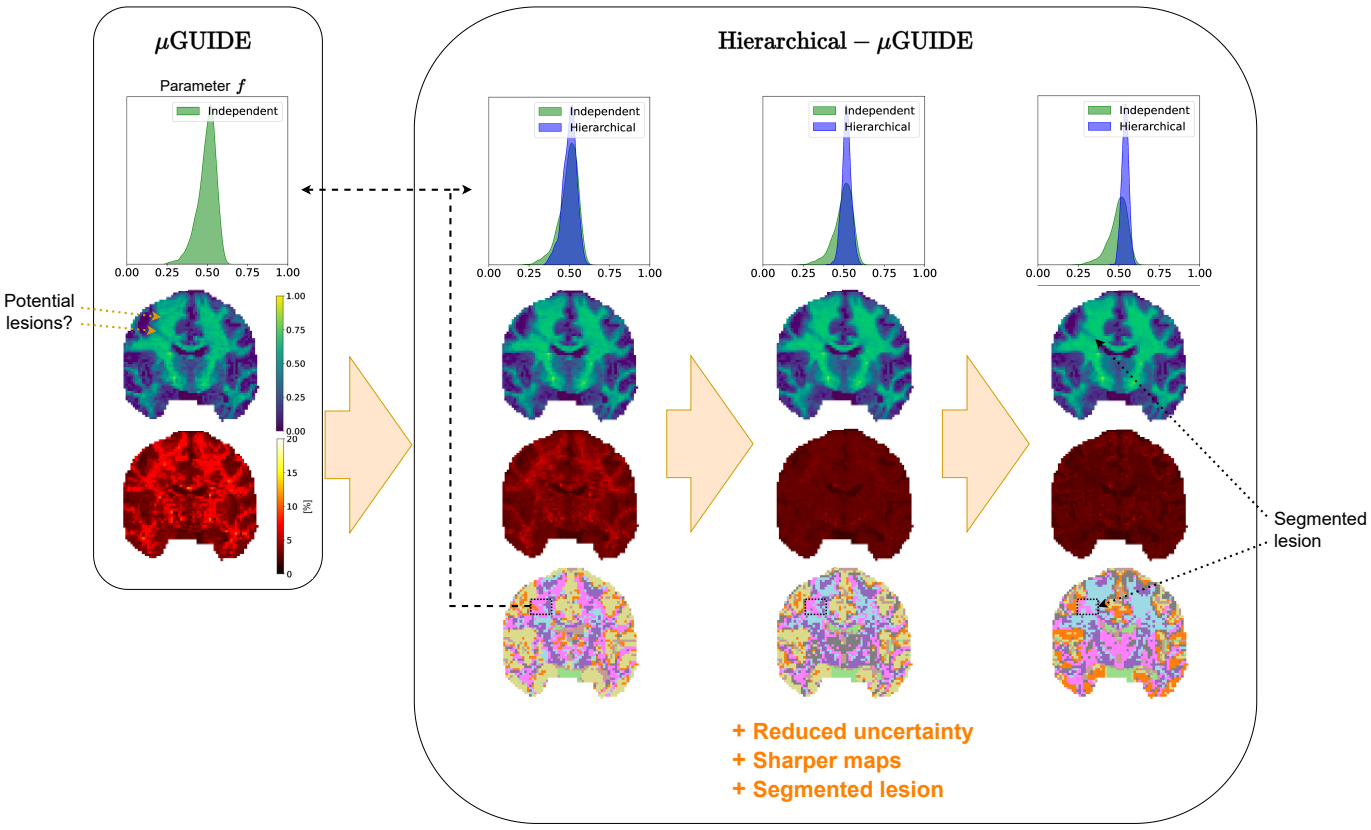

Fig. 8.4.: Illustration of the Hierarchical-μ-GUIDE optimization Evolution of the mean, uncertainty (relative standard deviation), and parcellation during training on a slice of a participant with epilepsy. Hierarchical-μ-GUIDE starts from the independent posterior distributions estimated using μ-GUIDE and progressively regularises those into distributions with reduced uncertainty. Contrary to spatial smoothing, neighboring voxels are not averaged together, which maintains the sharpness of the parameter maps and highlights lesions while preserving tissue heterogeneity.

Strong independent-case initialization In practice, given the complexity of this inference problem, a random initialization of the variational family's weights fails. Instead, we use a **principled scheme that initializes hierarchical posteriors to their independent-case counterparts**. In Section 8.2.1, we put forward that μ-GUIDE outputs both encodings $\mathbf{E}_n^{\text{ind}}$ and a trained NF $\mathcal{F}^{\text{ind}}(\bullet; \phi^{\text{ind}}, \mathbf{E}_\bullet^{\text{ind}})$. We use those as part of our initialization, as described in Algorithm 5. This initialization scheme puts forward the interplay between PAVI and the encoding/conditional density estimator couple described in Section 4.1.3. At time $t = 0$, the hierarchical posteriors are equal to the independent-case posterior, before being progressively regularized by the hierarchical prior. This is illustrated in Figure 8.4.

Optimization In Equation (8.1) we construct a composite model that includes $P(X^{\text{vox}}|\theta^{\text{vox}})$. Yet in Section 8.2.1, we don't learn this likelihood directly but the posterior $P_{\text{ind}}(\theta^{\text{vox}}|X^{\text{vox}})$. To go from one to the other, we rely on Baye's theorem —defined in Section 3.1. The likelihood does not depend on the prior, so the same likelihood term can be used across the independent and hierarchical models:

$$\begin{aligned} P_{\text{ind}}(X^{\text{vox}}, \theta^{\text{vox}}) &= P(X^{\text{vox}}|\theta^{\text{vox}}) \times P_{\text{ind}}(\theta^{\text{vox}}) \\ &= P_{\text{ind}}(\theta^{\text{vox}}|X^{\text{vox}}) \times P_{\text{ind}}(X^{\text{vox}}) \\ \implies \log p(X^{\text{vox}}|\theta^{\text{vox}}) &= \log p_{\text{ind}}(\theta^{\text{vox}}|X^{\text{vox}}) + \log p_{\text{ind}}(X^{\text{vox}}) - \log p_{\text{ind}}(\theta^{\text{vox}}) \end{aligned} \tag{8.2}$$

where:

- $\log p_{\text{ind}}(X^{\text{vox}})$ does not depend on θ^{vox};
- $\log p_{\text{ind}}(\theta^{\text{vox}})$ corresponds to a uniform prior, so is a constant with respect to θ^{vox}.

As a result, injecting this rewriting in the ELBO yields:

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_Q \left[\log p_{\text{hier}}(X^{\text{vox}}, \theta^{\text{vox}}, \Theta^{\text{mixture}}) - \log q(\theta^{\text{vox}}, \Theta^{\text{mixture}})\right] \\ &= \mathbb{E}_Q \left[\log p(X^{\text{vox}}|\theta^{\text{vox}}) + \log p_{\text{hier}}(\theta^{\text{vox}}, \Theta^{\text{mixture}}) - \log q(\theta^{\text{vox}}, \Theta^{\text{mixture}})\right] \\ &\propto \mathbb{E}_Q \left[\log p_{\text{ind}}(\theta^{\text{vox}}|X^{\text{vox}}) + \log p_{\text{hier}}(\theta^{\text{vox}}, \Theta^{\text{mixture}}) - \log q(\theta^{\text{vox}}, \Theta^{\text{mixture}})\right] \end{aligned} \tag{8.3}$$

were $\Theta^{\text{mixture}} = (l^{\text{vox}}, \mathit{logits}^{\text{vox}}, \theta^{\text{parc}})$ abstracts the mixture-related parameters. During optimization, the independent-case, uniform-prior posterior can thus be used in lieu of the likelihood. Using this modified loss instead of the original ELBO, we can train the variational family using the r-KL.

Note: Even if the prior used in Section 8.2.1 is not uniform, its density —which is not a constant anymore— can be injected in Equation (8.3). This amounts to importance-resampling the model prior, using the log ratio of the hierarchical and independent-case priors.

8.3 Results: joint tissue parcellation and reduced-uncertainty inference

8.3.1 Synthetic experiment validation

First, we validate our method on a synthetic example where we know the ground truth parameters. We start from realistic white matter configurations θ^{vox} (Coelho et al., 2022). We simulate corresponding synthetic signals X^{vox} using the same procedure as in Section 8.2.1. Results are visible in Figure 8.5, where injecting a meaningful structure helps recover the ground truth with greater precision.

8.3.2 Application to a healthy subject

Next, we apply our method to the signal coming from a healthy subject. Results are visible in Figure 8.6. The parcellation partitions the tissue into meaningful clusters: grey matter, white matter, ventricles, and thick fiber bundles.

8.3.3 Application to a subject with epilepsy

Finally, we apply our method to the signal coming from a subject with epilepsy. Results are visible in Figure 8.7. The epileptic lesion (in the left hemisphere) is segmented via the parcellation.

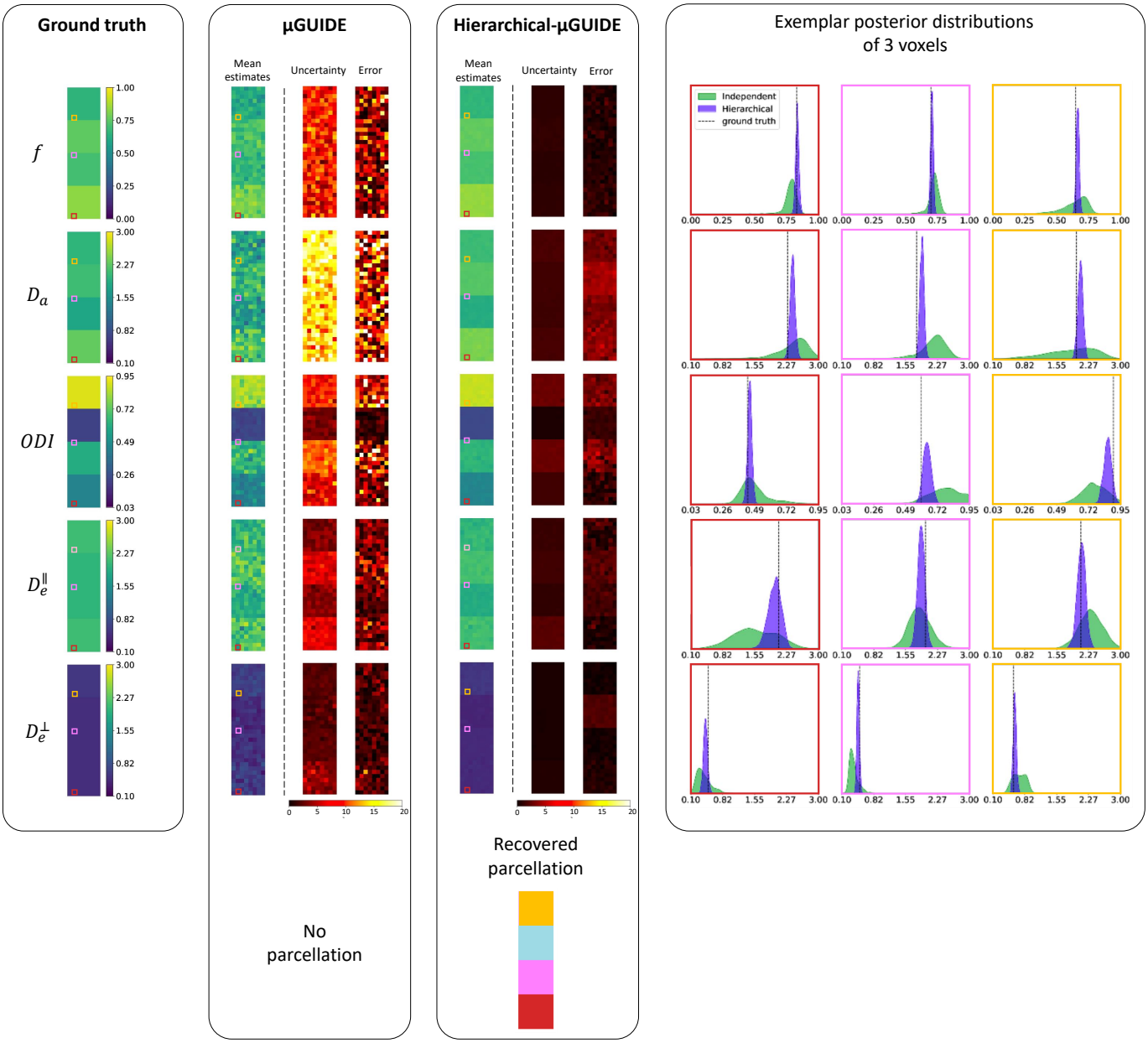

Fig. 8.5.: Experiment over a synthetic tissue composed of four distinct realistic subtypes This experiment simulates a synthetic tissue composed of four square parcels. As visible on the left, each parcel is associated with a distinct realistic white matter tissue configuration (Coelho et al., 2022). We aim at recovering the ground truth parameters (on the left). Both methods μ-GUIDE and hierarchical-μ-GUIDE output a full posterior distribution for the voxel parameters. In the middle, we report the posterior means, uncertainty —the posterior standard deviation— and error compared to the ground truth. On the right, we report 3 exemplar voxel's posterior distributions. Hierarchical μ-GUIDE reduces the uncertainty and error of the estimates and provides a parcellation of the tissue.

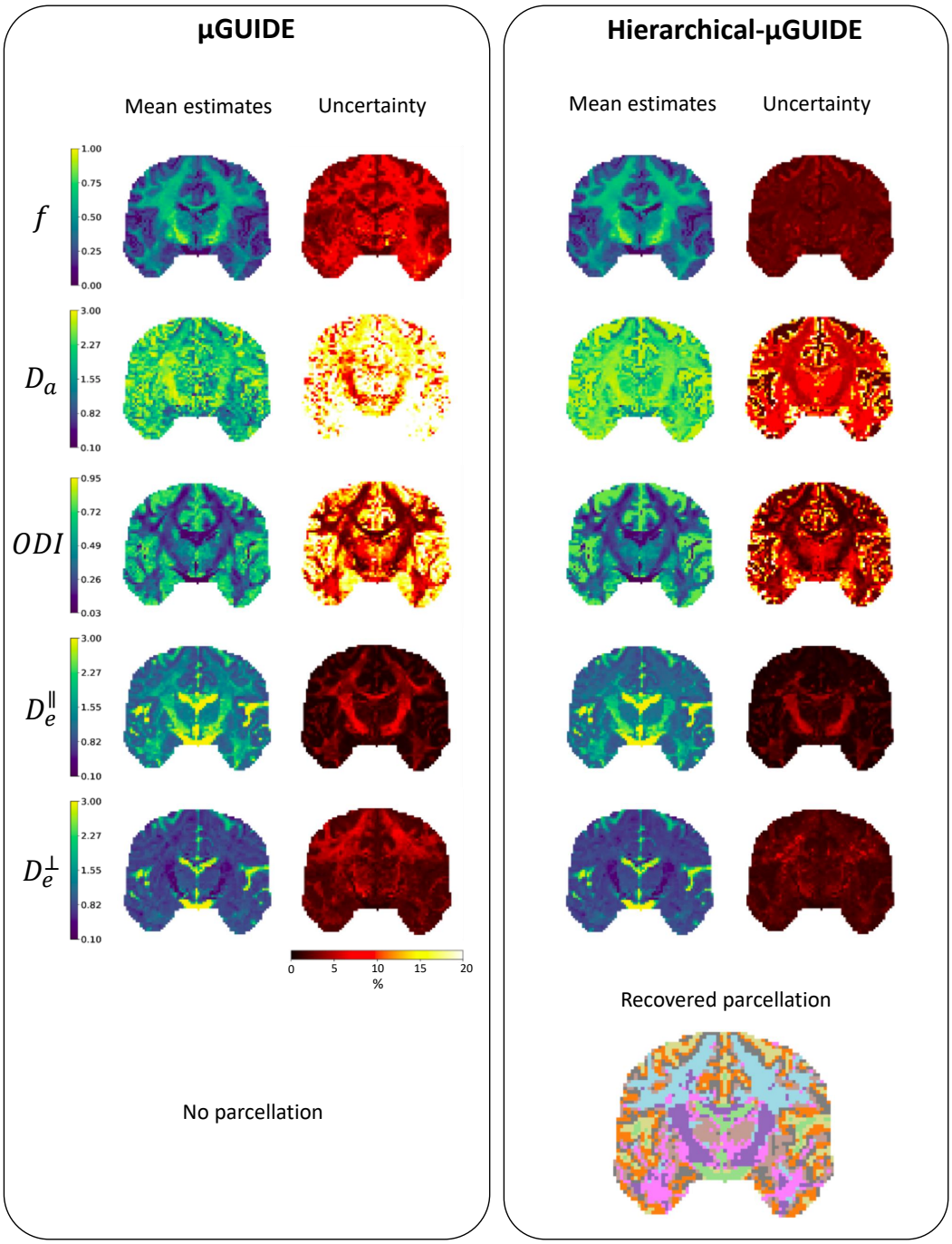

Fig. 8.6.: Application on a healthy subject Parametric maps of a healthy participant using μ-GUIDE and hierarchical μ-GUIDE (L=8 parcels). We report the mean and uncertainty of the estimates, although full posterior distributions are estimated in each voxel. Uncertainty is reduced using hierarchical μ-GUIDE and a meaningful parcellation is recovered.

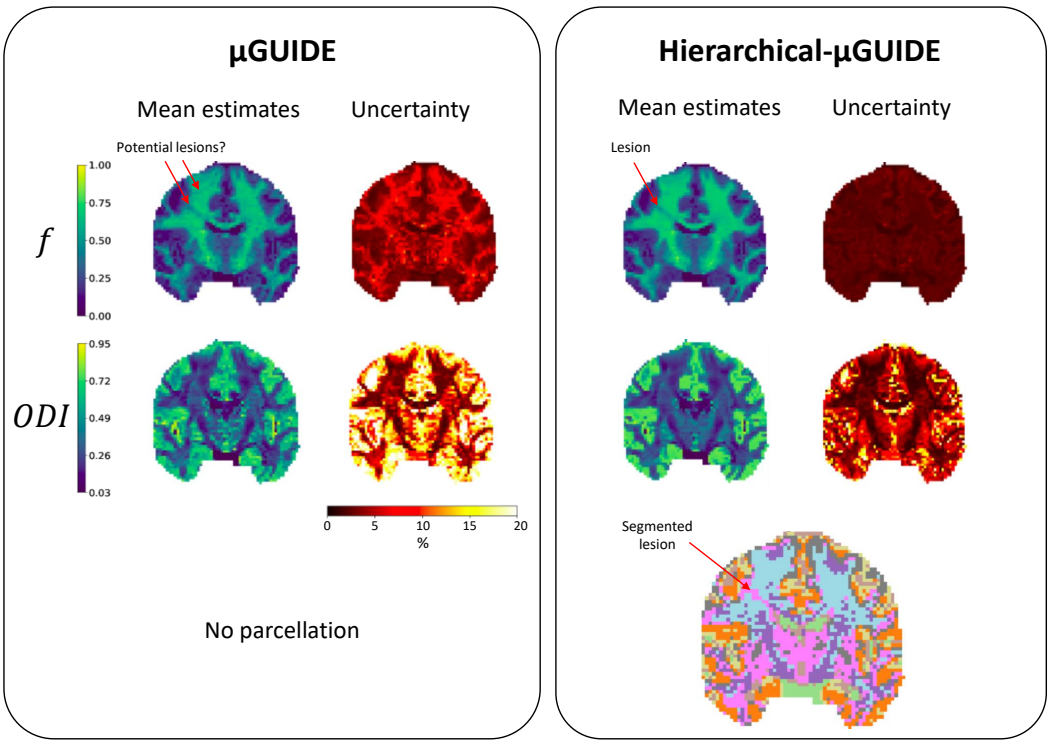

Fig. 8.7.: Application to a subject with epilepsy Parametric maps of a participant with epilepsy using μ-GUIDE and hierarchical μ-GUIDE (L=8 parcels). The lesion is clearly segmented in the obtained parcellation. hierarchical μ-GUIDE preserves tissue heterogeneity.

8.4 Summary of contributions

This chapter summarized our contributions in dMRI-based tissue microstructure estimation (Rouillard, Wassermann, et al., 2024):

- Uncertainty is considered a major challenge in microstructure estimation (Alexander et al., 2019; Jelescu, Palombo, et al., 2020). We reduce this uncertainty by injecting meaningful biases in the form of hierarchical priors. Compared to the work from Powell et al. (2021), we do not rely on an initial segmentation of brain tissues to average distribution across large compartments such as the white and gray matter. In contrast, we infer a tissue segmentation *from* the data, by jointly parcellating voxels and estimating their microstructure.
- We propose the first "global" biophysical-model-based microstructure parcellation scheme (using the global vs local taxonomy from Simon B. Eickhoff et al. (2018b)). To this end, we design an original *latent* GM HBM, clustering distributions rather than observed data points.
- We design novel architectures injecting a learned distribution surrogate inside a HBM. While Glöckler et al. (2022) concomitantly combined VI with NLE, we leverage NPE in hierarchical modeling. Specifically, we propose to plate-amortize f-KL-trained distributions inside composite HBMs, inferred upon using the r-KL.
- We derive original initialization schemes, illustrating the interplay between PAVI and NPE. Specifically, we propose to learn hierarchical posteriors as a perturbation of their independent counterpart.
- We illustrate the potential of our method in tissue lesion segmentation.
- We design a general method that could be applied to:
  - different microstructure models (Hui Zhang et al., 2012; Novikov et al., 2019; Palombo et al., 2020)
  - different priors implementing different hypothesis
  - multi-subject, multi-scans hierarchical setups.

9 Application: reliable large-scale directional coupling estimation in fMRI

This work tackles functional connectivity, as introduced in Section 1.2. The goal is to investigate the connections linking regions in the brain when performing different cognitive tasks. As introduced in Section 1.2, this task is complexified by several factors: the confounding by the BOLD response; the confounding by unobserved regions; and the limitations of correlation as a measure of connectivity. We tackle all three problems via a method that marginalizes the HRFs, scales up to the full brain, and infers the directional coupling between regions. From a methodological standpoint, we avoid mode collapse in this complex inference task by injecting a f-KL plate-amortized estimator inside a scalable r-KL inference.

First, we provide some context on functional connectivity. We then present the hierarchical modeling and inference method we use to tackle this task. Finally, we present some results, including full-brain coupling experiments.

9.1 Functional connectivity, and directional coupling estimation

One of the prime goals of fMRI is the investigation of brain **function**. That is to say, the association of given brain regions (e.g. the occipital lobe, at the "back" of the brain) with given cognitive functions (in this case, vision) —see Figure 7.1 for an illustration. Historically, this association between anatomy and function has been investigated using task fMRI contrasts (Poldrack et al., 2011; K. J. Friston, 2009). Individual region activations are measured during a certain activity (e.g. looking at human faces) and compared to a baseline condition (at rest). A significant difference across the conditions indicates the association of a given brain region with a cognitive task. A different viewpoint is to consider different brain regions not individually, but through their *connections* as part of larger **networks** (Van Den Heuvel and Pol,

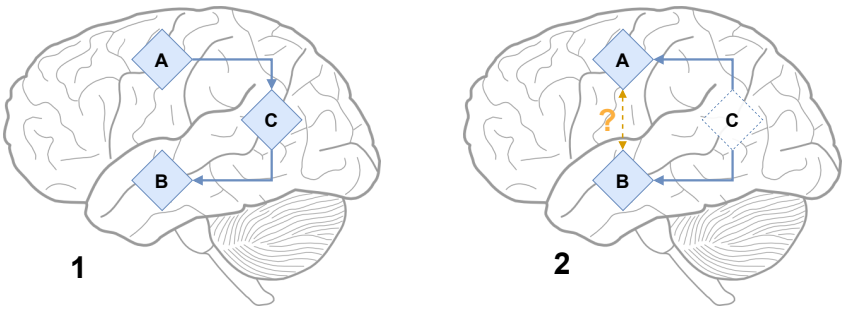

Fig. 9.1.: Various coupling scenarios *1-* Region A activates region B, which activates region C. All three regions have correlated time series. *2-* Region C (unobserved) activates both A and B. Because C is not observed, this could be confounded as a coupling between A and B.

2010). Those networks would be dynamically mobilized during cognitive tasks, leading to a hierarchy of brain function (Simon B. Eickhoff et al., 2018b). In this "connectivity" viewpoint, the focus is moved from the regions themselves (the nodes in a graph) to their connections (the edges).

But how to define those connections?

The most straightforward tool to investigate connectivity is *correlation* (Van Den Heuvel and Pol, 2010; Varoquaux and Craddock, 2013). If different brain regions —even located in different parts of the cortex— have correlated time series, we can assume they are connected at some level, and partake in the same brain functions. Correlation —which we use in Chapter 7— has thus been the historical tool in the Neuroscience community to construct connectomes. Yet, correlation is lacking in several ways. First, correlation is a symmetrical metric: the correlation between regions A and B is the same as between B and A. In contrast, it would be informative to know which region, A or B, activates the other sequentially. Second, correlation is not informative about the finer role of regions as part of correlated networks. Consider the situation 1 (left) in Figure 9.1. In the presence of a $A \rightarrow C \rightarrow B$ network, all three regions will have correlated signals, as would be the case in a $C \rightarrow A \rightarrow B$ network. Thus **correlation lacks the *granularity* that would help unveil precise region functions**.

As a result, Neuroscientists have looked into refining connections into so-called "causal" or "effective" links (Roebroeck et al., 2005; K. J. Friston et al., 2003; Valdes-Sosa et al., 2011). For instance, Granger causal analysis (GCA) investigates directional coupling using information theory (Roebroeck et al., 2005). Another example is dynamic causal modeling (DCM) which relies on Bayesian modeling and inference (K. J. Friston et al., 2003). Though those metrics yield the finer granularity

that correlation lacks, they are significantly harder to compute and suffer from at least two issues.

First, due to their computational complexity, causal methods are usually applied to a dozen or so pre-selected regions. This **biases the investigation** towards the existing knowledge of experimenters —unforeseen regions would be completely missed by such an analysis. What's more, investigating the connections only between pre-selected regions is sensitive to the confounding from unobserved regions —as illustrated in Figure 9.1 2 (right). Activation from an unobserved region could be mistaken as a coupling between the observed regions. A brute-force solution to this problem is to observe the *entirety* of the brain at once. This limits the risk of confounds and allows for a more data-driven approach to discovering region function. Because of the significant computation they entitle, full-brain analysis has only been recently investigated (Frässle, Lomakina, et al., 2017; Arab et al., 2023). As an example, regression dynamical causal modeling (r-DCM) recently revisited DCM to improve its scalability (Frässle, Lomakina, et al., 2017). **In this work, we develop a method capable of full brain directional coupling estimation.**

The second limitation of causal methods is the confounding by the hemodynamic response function (HRF) (Rangaprakash, Barry, et al., 2023; Rangaprakash, G.-R. Wu, et al., 2018). fMRI does not measure directly regional activity, but a delayed blood oxygenation response —as illustrated in Figure 1.1. The HRF can vary across brain regions (Devonshire et al., 2012; Handwerker et al., 2004; Taylor et al., 2018). Different delays in the hemodynamic response can thus be mistaken as the precedence of the underlying neuronal signal, which may for instance significantly confound GCA (Deshpande et al., 2010). **In this work, we specifically look into the marginalization of the uncertainty in the HRF, and its effect on directional coupling estimation.**

The next section describes our method in greater detail.

9.2 MDSI using hybrid variational Bayes (MDSI-h-VB): reliable large-scale estimation of directional coupling

9.2.1 The multivariate dynamical system (MDS) model

Our goal is to infer the coupling between different brain regions. To this end, we describe in this section the multivariate dynamical system (MDS) model (Ryali et al., 2011). We hypothesize that the observed BOLD signal is generated by the convolution of some latent activation by the HRF. We assume this coupling between regions to be linear and at the latent activation level. In the next paragraphs, we explain our modeling from the observed BOLD signals down to the latent coupled activations.

Notations We denote scalars using lowercase symbols: $x \in \mathbb{R}$. We denote cardinalities, such as the number of brain regions, using sans serif uppercase letters: $\mathsf{M} \in \mathbb{N}$. We denote vectors using lowercase bold symbols: $\boldsymbol{x} \in \mathbb{R}^{\mathsf{T}}$. We denote matrices and tensors using uppercase bold symbols: $\mathbf{X} \in \mathbb{R}^{\mathsf{M} \times \mathsf{T}}$. Indexing denotes the selection of an element inside a tensor: $\boldsymbol{x}_m \in \mathbb{R}^{\mathsf{T}}$ denotes the mth element for the matrix $\mathbf{X} \in \mathbb{R}^{\mathsf{M} \times \mathsf{T}}$. We denote the realization of a random variable using a typewriter font: $\mathbf{Y} = \mathtt{Y}$

BOLD response We denote as M the number of regions. $\boldsymbol{y}_m \in \mathbb{R}^{\mathsf{T}}$ denotes the BOLD time series for the region m, where T denotes the temporal duration of the signal. We model $\boldsymbol{y}_m$ as the convolution of some latent activation by the HRF. $\boldsymbol{x}_m \in \mathbb{R}^{\mathsf{T}}$ denotes the latent activation. The HRF is assumed to be region-specific: the region m is associated with the HRF $\boldsymbol{h}_m \in \mathbb{R}^{\mathsf{K}}$ of temporal duration K. We obtain $\boldsymbol{y}_m$ as:

$$\begin{aligned} y_m[t] &= (\boldsymbol{h}_m * \boldsymbol{x}_m)[t] + \eta \\ \eta &\sim \mathcal{N}(0, r_m) \\ \boldsymbol{r} &= [r_1 \ \ldots \ r_{\mathsf{M}}] \in \mathbb{R}^{+\mathsf{M}} \end{aligned} \tag{9.1}$$

where $[t]$ denotes the time indexing, and $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{r})$ denotes some BOLD-level white Gaussian noise that we assume to be independent and of different amplitude across regions.

We model the region-specific HRF $\boldsymbol{h}_m$ as a linear combination of the canonical HRF vector and its time derivative (Glover, 1999), both of temporal duration K:

$$\begin{aligned} \boldsymbol{h}_m &= \cos(\alpha_m) \times \mathbf{hrf} + \sin(\alpha_m) \times \dot{\mathbf{hrf}} \\ \alpha_m &\in]-\pi/4, \pi/4[\\ \mathbf{hrf},\ \dot{\mathbf{hrf}} &\in \mathbb{R}^{\mathsf{K}} \end{aligned} \tag{9.2}$$

where following Steffener et al. (2010), the HRF coefficients are parameterized on the unit circle. This means that the HRF for a region m is entirely described by the angle α_m. When $\alpha_m = 0$, the HRF is equal to the canonical HRF, while positive or negative values induce differences in the response's time-to-peak, the presence of an initial dip, and the response's amplitude. The range $]-\pi/4, \pi/4[$ simulates the time-to-peak variability observed in the human HRF (Taylor et al., 2018). We model the hemodynamic response as independent across regions. Considering all the regions at once, we respectively denote $\mathbf{Y}, \mathbf{X} \in \mathbb{R}^{\mathsf{M}\times\mathsf{T}}$ and $\mathbf{H} \in \mathbb{R}^{\mathsf{M}\times\mathsf{K}}$ the concatenated BOLD signals, activations, and HRFs. Vectorizing the convolution operation across regions, we can write $\boldsymbol{y}[t] = (\mathbf{H} * \mathbf{X})[t] + \boldsymbol{\eta}$.

Latent activation dynamics In the MDS, the latent activations $\mathbf{X}$ are subject to the coupling between different regions. We assume this coupling to be linear and parameterized by a coupling matrix $\mathbf{A}$. The evolution of the latent signal follows the linear Gaussian state-space model:

$$\begin{aligned} \boldsymbol{x}[t+1] &= \mathbf{A}[t] \times \boldsymbol{x}[t] + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{q}) \\ \boldsymbol{q} &\in \mathbb{R}^{+\mathsf{M}} \end{aligned} \tag{9.3}$$

where $\boldsymbol{\epsilon}$ denotes some latent white Gaussian noise that we assume to be independent and of different amplitude across regions. We hypothesize that the coupling matrix $\mathbf{A}[t]$ varies through time, depending on the experimental condition —amongst C possible conditions:

$$\begin{aligned} \mathbf{A} &\in \mathbb{R}^{\mathsf{C}\times\mathsf{M}\times\mathsf{M}} \\ \boldsymbol{c} &\in \{1, \ \ldots \ , \mathsf{C}\}^{\mathsf{T}} \\ \mathbf{A}[t] &= \mathbf{A}_{c[t]} \in \mathbb{R}^{\mathsf{M}\times\mathsf{M}} \end{aligned} \tag{9.4}$$

The matrix $\mathbf{A}$, representing the coupling for all the conditions, is our analysis's main quantity of interest. A positive coefficient $a_{c,m,n}$ denotes a positive coupling from the region m to the region n under condition c. If the region m is active at time $[t]$ (with a positive value for the activation $x_m[t]$), then the region n will be more active

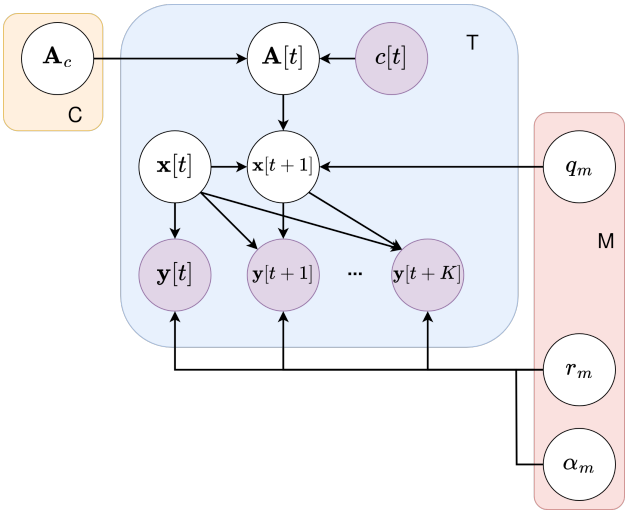

Fig. 9.2.: The MDS model. The latent signal $\boldsymbol{x}$ follows a linear Gaussian state-space model governed by the coupling matrix $\mathbf{A}$. The coupling matrix depends on the experimental condition, denoted by the vector $\boldsymbol{c}$. The observed BOLD signal $\boldsymbol{y}$ results from the convolution of the latent signal with the region-specific HRF, described by the angle α_m. q_m and r_m denote region-specific noise levels.

at time $[t+1]$. Note that this coupling is directed: a positive coupling from m to n does not imply the converse, as for a correlation analysis (Van Den Heuvel and Pol, 2010; Varoquaux and Craddock, 2013).

Summary We can summarize the MDS model using the set of equations:

$$
\begin{aligned}
\mathbf{Y} &= [\boldsymbol{y}[1] \ \dots \ \boldsymbol{y}[\mathsf{T}]] \\
\boldsymbol{y}[t] &= (\mathbf{H} * \mathbf{X})[t] + \boldsymbol{\eta} \\
\boldsymbol{\eta} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{r}) \\
\mathbf{H} &= \begin{bmatrix} \cos(\alpha_1) & \sin(\alpha_1) \\ \vdots & \vdots \\ \cos(\alpha_{\mathsf{M}}) & \sin(\alpha_{\mathsf{M}}) \end{bmatrix} \times \begin{bmatrix} \mathbf{hrf} \\ \cdot \\ \mathbf{hrf} \end{bmatrix} \\
\mathbf{X} &= [\boldsymbol{x}[1] \ \dots \ \boldsymbol{x}[\mathsf{T}]] \\
\boldsymbol{x}[t+1] &= \mathbf{A}[t] \times \boldsymbol{x}[t] + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{q}) \\
\mathbf{A}[t] &= \mathbf{A}_{c[t]}
\end{aligned}
\tag{9.5}
$$

where our main goal is to infer the coupling matrix $\mathbf{A}$ from the observed BOLD signal $\mathbf{Y}$. A graphical representation of the MDS model is visible in Figure 9.2.

9.2.2 Problem statement: directional coupling inference from the BOLD signal

Given the MDS model described in Equation (9.5), the observed BOLD signal $\mathbf{Y}$ and experimental conditions $\boldsymbol{c}$, we aim to infer the parameters susceptible to generating $\mathbf{Y}$. The MDS model described in Equation (9.5) is associated with the joint distribution P, which factorizes as:

$$\begin{aligned} P(\mathbf{Y}, \boldsymbol{c}, \mathbf{X}, \mathbf{A}, \boldsymbol{q}, \boldsymbol{r}, \mathbf{H}) = & P(\mathbf{Y}|\mathbf{X}, \boldsymbol{r}, \mathbf{H}) \\ & \times P(\mathbf{X}|\boldsymbol{c}, \mathbf{A}, \boldsymbol{q}) \\ & \times P(\boldsymbol{c})P(\mathbf{A})P(\boldsymbol{r})P(\boldsymbol{q})P(\mathbf{H}) \end{aligned} \tag{9.6}$$

where $P(\boldsymbol{c})$ is a uniform categorical prior, $P(\mathbf{A})$ is a sparsity-inducing Laplace prior, $P(\mathbf{H})$ corresponds to a uniform prior over the angle α between the bounds $]-\pi/4, \pi/4[$, and $P(\boldsymbol{q})$ and $P(\boldsymbol{r})$ are log-normal priors. $P(\mathbf{X}|\boldsymbol{c}, \mathbf{A}, \boldsymbol{q})$ and $P(\mathbf{Y}|\mathbf{X}, \boldsymbol{r}, \mathbf{H})$ correspond to the Normal distributions described in Equation (9.5).

Following the Bayesian inference formalism, we search for the **posterior distribution of the coupling matrix**: $P(\mathbf{A}|\mathbf{Y}, \boldsymbol{c})$. $P(\mathbf{A}|\mathbf{Y}, \boldsymbol{c})$ denotes a distribution because there are several sources of uncertainty in the problem, and therefore $\mathbf{A}$ cannot be inferred unequivocally. In particular, the noise levels at the BOLD level $\boldsymbol{r}$ and latent level $\boldsymbol{q}$ are unknown. In addition, the HRF $\mathbf{H}$ for the different regions is also unknown. When estimating the latent signal $\mathbf{X}$ and the coupling matrix $\mathbf{A}$, we want to ensure that the uncertainty in all the other parameters is properly *marginalized*. That is to say, we do not want to underestimate the uncertainty when inferring the parameters of interest $\mathbf{A}$.

In detail, our method focuses on the proper marginalization of the HRF $\mathbf{H}$. Each combination of different HRFs for the brain regions yields —via de-convolution— a different set of latent signals $\mathbf{X}$. In turn, each different set of latent signals yields a different estimate for the coupling matrix $\mathbf{A}$. Theoretically, the Bayesian framework allows weighting all those scenarios by their likelihood of generating the observed BOLD signal $\mathbf{Y}$. This results in a single posterior distribution $P(\mathbf{A}|\mathbf{Y}, \boldsymbol{c})$ that integrates all the sources of uncertainty in the problem.

However, in practice, inference methods may fail to recover the true posterior $P(\mathbf{A}|\mathbf{Y}, \boldsymbol{c})$, resulting in uncertainty underestimation and biased estimation. This is due to the **mode collapse** problem described in Section 3.4. In particular, without special attention, inference **methods focus on specific HRFs** and the associated underlying signals. As a consequence, those methods "miss" entire parts of the

A solution space. Critically, while inference methods may fail to recover the true uncertainty in $P(\mathbf{A}|\mathbf{Y}, \boldsymbol{c})$, they still output the distribution corresponding to the mode they are stuck into. This can be a misleading result: recovering a probabilistic output, experimenters may assume that *all* the uncertainty in the problem has been captured. Yet, in practice, off-the-shelf methods may only recover *part* of the problem's uncertainty. In the context of the MDS generative model —described in Equation (9.5)— mode collapse can result in over-inflated statistical confidence when inferring the connections between regions, and even in spurious connections discovery. In this work, we propose an inference method to **marginalize the uncertainty in the HRF H and the noise levels q and r properly when inferring the latent signal X and the coupling matrix A**.

9.2.3 Hybrid Variational Bayes: leveraging a plate-amortized f-KL-trained estimator to prevent mode collapse in a scalable r-KL inference

In this section, we describe our hybrid variational Bayes method (h-VB) to tackle the multi-modality in inference. The term *hybrid* refers to separating the parameters $(\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}, \mathbf{X}, \mathbf{A})$ into two groups treated using different inference methods. Specifically, as described below, we use a r-KL gradient-based VI loss for the coupling and latent signal parameters as they correspond to a well-behaved unimodal conditional optimization problem. On the other hand, we use a forward amortized VI (FAVI) loss (Ambrogioni, Güçlü, Berezutskaya, et al., 2019) for the noise and HRF parameters since their posterior distribution is often highly multi-modal. This results in a hybrid approach that combines the efficiency and scalability of r-KL VI for large-scale inference of large coupling matrices with the robustness of FAVI on a smaller set of key (hyper-)parameters.

Composite variational family We optimize the variational distribution Q to approximate the unknown posterior $P(\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}, \mathbf{X}, \mathbf{A}|\mathbf{Y}, \boldsymbol{c})$. We factorize Q into two distributions:

$$Q(\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}, \mathbf{X}, \mathbf{A}; \phi) = Q_{\text{HP}}(\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}; \phi_{\text{HP}}) \times Q_{\text{P}}(\mathbf{X}, \mathbf{A}|\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}; \phi_{\text{P}}) \tag{9.7}$$

where Q_{HP} denotes our *hyper-parameter* estimator, and Q_{P} our *parameter* estimator. Per our "hybrid" method, both factors are trained using different losses, as explained in the next two sections.

Hyper-parameter (HP) estimation Our main goal when training Q_{HP} is to avoid mode collapse, the phenomenon described in Section 3.4. We consider the different regions as independent inference problems and factorize Q_{HP} as:

$$Q_{\text{HP}}(\boldsymbol{q}, \boldsymbol{r}, \mathbf{H} | \mathbf{Y} = \mathrm{Y}; \phi_{\text{HP}}) = \prod_{m=1..\mathrm{M}} Q_{\text{region}}(q_m, r_m, \alpha_m; f(\mathrm{y}_m; \phi_{\text{HP}})) \tag{9.8}$$

where Q_{region} approximates a location's noise levels and HRF given a realization of the region's observable signal y. To approximate Q_{region}, we use the NPE approach described in Section 4.2.2. A MAF approximates the distribution of (α, q, r), conditioned by an encoding of the observed region's observable signal $f(\mathrm{y}_m)$. As encoder f, we use a time convolutional neural network.

We train Q_{region} to minimize the amortized f-KL loss, that is to say, to maximize the probability of (q, r, α) given $\boldsymbol{y}$:

$$\begin{aligned} \phi^*_{\text{HP}} &= \min_{\phi_{\text{HP}}} \mathcal{L}^{\text{f-KL}}_{\text{HP}} \\ &= \min_{\phi_{\text{HP}}} \mathbb{E}_{q,r,\alpha,\boldsymbol{y}\sim P} \left[-\log q_{\text{region}}(q, r, \alpha; f(\boldsymbol{y}; \phi_{\text{HP}})) \right] \end{aligned} \tag{9.9}$$

where the expectation $\mathbb{E}_{q,r,\alpha,\boldsymbol{y}\sim p}$ denotes the training over a large synthetic dataset sampled from the MDS model described in Section 9.2.1. The training of Q_{region} is *amortized*, which means that once trained, Q_{region} can estimate the hyper-parameters of *any* brain region by feeding the region's BOLD signal y to the encoder f. We can then reuse Q_{region} across symmetrical inference problems inside the MDS generative model, leveraging *plate amortization* as defined in Section 6.1.2.

In the next paragraph, we feed to this amortized estimator the observed signal $\mathrm{Y}^{\text{observed}}$, resulting in a posterior of $\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}$ for that particular signal. By training over a sizeable synthetic dataset and using the f-KL loss, we avoid mode collapse for Q_{HP} and, consequently, for Q_{P} as detailed below.

Parameter (P) estimation Our main goal when training Q_{P} is inference speed and scalability. This is due to the large dimensionality of $\mathbf{X}$ and $\mathbf{A}$, which scale badly with

the number of regions and time points in our experiments. We use the r-KL loss to ensure this scalability. We maximize the ELBO under the variational distribution:

$$\begin{aligned}\phi_{\mathrm{P}}^{*} &= \min_{\phi_{\mathrm{P}}} \mathcal{L}_{\mathrm{P}}^{\text{r-KL}} \\ &= \min_{\phi_{\mathrm{P}}} \mathbb{E}_{\substack{\boldsymbol{q},\boldsymbol{r},\mathbf{H}\sim Q_{\mathrm{HP}} \\ \mathbf{X},\mathbf{A}\sim Q_{\mathrm{P}}}} \quad \log p(\mathbf{Y},\boldsymbol{c},\mathbf{X},\mathbf{A},\boldsymbol{q},\boldsymbol{r},\mathbf{H}) \\ &\qquad\qquad - \log q_{\mathrm{P}}(\mathbf{X},\mathbf{A}|\boldsymbol{q},\boldsymbol{r},\mathbf{H};\phi_{\mathrm{P}}) \\ &\qquad\qquad - \log q_{\mathrm{HP}}(\boldsymbol{q},\boldsymbol{r},\mathbf{H}|\mathbf{Y}=\mathrm{Y}^{\text{observed}})\end{aligned} \tag{9.10}$$

where the estimator Q_{HP} —described in the previous paragraph— evaluated on the true observed signal $\mathrm{Y}^{\text{observed}}$ is used as the variational posterior for the hyper-parameters $\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}$. $Q_{\mathrm{HP}}(\boldsymbol{q}, \boldsymbol{r}, \mathbf{H})$ is not re-trained during this second phase to prevent mode collapse.

Theoretically, if the HRF $\mathbf{H}$ and the noise levels $\boldsymbol{q}$ and $\boldsymbol{r}$ were known, $\mathbf{X}$ can be inferred via Wiener de-convolution. In turn, given the latent signals, $\mathbf{A}$ can be inferred in closed form via Bayesian linear regression. Informed by those considerations, we choose a Gaussian variational family to approximate the exact $\mathbf{X}$ and $\mathbf{A}$ posterior distributions. To scale our method to hundreds of regions, we do not model the covariance between the different coefficients of $\mathbf{A}$, hence the covariance matrix for the posterior of $\mathbf{A}$ is modeled as diagonal. To obtain the mean and variance of the Gaussian approximations, we considered either regressing those from the value of the hyper-parameters $\mathbf{H}, \boldsymbol{q}, \boldsymbol{r}$, or keeping those as free parameters. In the $\alpha \in]-\pi/4, \pi/4[$ regime, those two parameterizations yielded identical results on numerous synthetic experiments. As a result, in the interest of simplicity, we used free parameters in our default implementation. The only exception lies in the synthetic experiment Section 9.3.1, where we increased the range of α to $]-\pi/2, \pi/2[$ to more clearly illustrate border cases where mode collapse yields biased inference.

Summary Our hybrid variational Bayes method can be summarized as follows:

1. We separate our latent parameters into two groups: the hyper-parameters which are susceptible to mode collapse ($\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}$), and the parameters which are high-dimensional ($\mathbf{X}, \mathbf{A}$);

2. First, we train an amortized estimator for the hyper-parameters $\boldsymbol{q}, \boldsymbol{r}, \mathbf{H}$. We use the f-KL loss, which prevents mode collapse. The amortized estimator can further be re-used across the different brain regions, using *plate amortization* (Rouillard, Bris, et al., 2023);

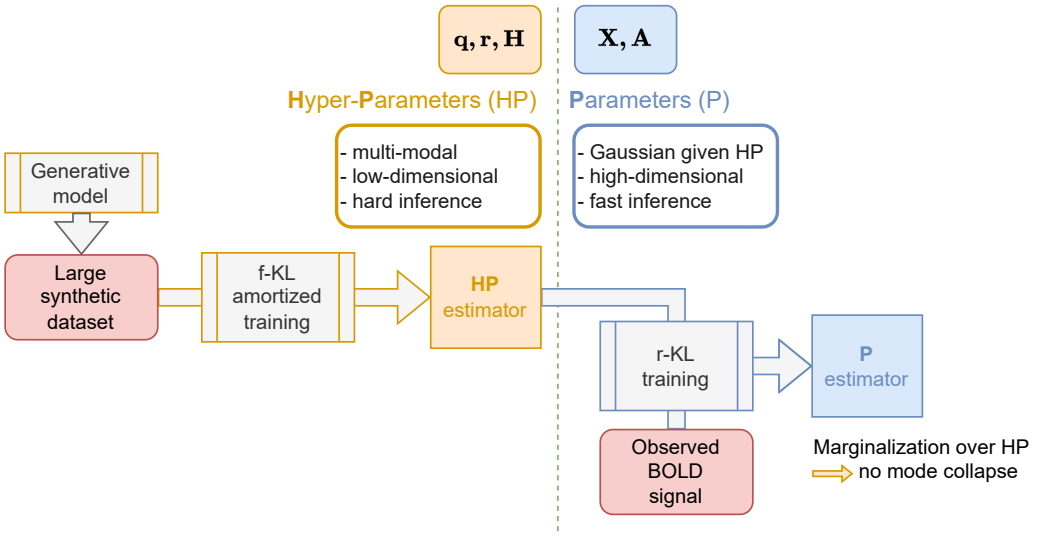

Fig. 9.3.: **Hybrid inference scheme** We separate our latent parameters into the hyper-parameters (HP) on the left and the parameters (P) on the right. We train an amortized HP-estimator using f-KL to prevent mode collapse. We plug the trained HP estimator into a scalable r-KL inference to infer the parameters. Marginalizing over the multi-modal HP posterior, we prevent mode collapse for the parameters, including the coupling matrix **A**. The term hybrid comes from combining different losses for different latent parameter groups.

3. Second, we train another estimator for the parameters $\mathbf{X}, \mathbf{A}$. We use the r-KL loss for fast convergence while using the pre-trained hyper-parameter estimator, which prevents mode collapse.

a graphical depiction of our method is visible in Figure 9.3. Combining the advantages of both losses, MDSI-h-VB allows us to infer the coupling matrix **A** from a high-dimensional observable signal **Y** in minutes while ensuring proper hyper-parameters marginalization, as exemplified in our experiments.

9.3 Results: from synthetic results to full-brain directional coupling estimation

9.3.1 Synthetic example: avoiding mode collapse

The goal of this synthetic experiment is to illustrate our methodological claims. MDSI-h-VB avoids mode collapse —the phenomenon described in Section 3.4— via the separate f-KL training of the HP estimator, as described in Section 9.2.3. In practice, this helps us recover the true uncertainty in the inference of the coupling

matrix **A** —described in Section 9.2.1. We show that, on the contrary, an off-the-shelf inference method underestimates the uncertainty in the coupling matrix **A**.

Data In this experiment, we use a synthetic sample from the MDSI generative model —described in Section 9.2.1. This means that the ground truth HRF **H**, variances levels $\boldsymbol{q}, \boldsymbol{r}$ and coupling **A** are known. We compare two inference methods that are fed with the synthetic BOLD signal $\mathtt{Y}$.

Baseline As a baseline for comparison, we use a variational Bayes method. Contrary to MDSI-h-VB, the entirety of the parameters —including $\mathbf{H}, \boldsymbol{q}, \boldsymbol{r}$— are inferred using the r-KL loss. As a result, the baseline focuses on certain HRFs only and misses part of the solution space for **A**. The baseline uses a Gaussian approximation for **A** and **X** (similar to MDSI-h-VB). The baseline approximates the posterior for $\boldsymbol{r}$ and $\boldsymbol{q}$ using log-normal distributions. The baseline approximates the posterior for α using a Normal distribution soft clipped to the range $]\pi/2; \pi/2[$ (using a rescaled sigmoid function).

Hyper-Parameter inference: HRF and variance levels Figure 9.4 displays the $(\alpha, \boldsymbol{q}, \boldsymbol{r})$ posterior distributions of MDSI-h-VB and the baseline. The baseline's posterior collapses to a small fraction of the posterior's support, thereby missing the ground truth parameters. On the contrary, MDSI-h-VB correctly recovers the entirety of the solution space. Note that, without strong priors on the underlying signal **X**, inferring the HRF **H** from the BOLD signal **Y** is ill-posed (Taylor et al., 2018). As a result, the support of MDSI-h-VB's α posterior is very large.

Parameter inference: coupling matrix Figure 9.5 displays the **A** posterior distributions of MDSI-h-VB and the baseline. Since the baseline ignored most of the HRF **H** solution space, it features peaked posteriors on spurious coupling values. This means that the baseline outputs biased results with strong statistical confidence. On the contrary, MDSI-h-VB correctly considers all the different HRF scenarios that could have generated the BOLD signal **Y**. As an example, consider the only non-null coupling in this synthetic example: a strong negative coupling from region 1 to region 2. Placing the threshold of the existence of a coupling at a 0.1 value, the baseline outputs a 1% chance of a positive coupling and a 0% chance of a negative coupling (the ground truth). On the contrary, MDSI-h-VB outputs a 46% chance for a positive coupling and a 30% chance for a negative coupling (the ground truth). MDSI-h-VB helps the experimenter determine that, though a coupling is likely to

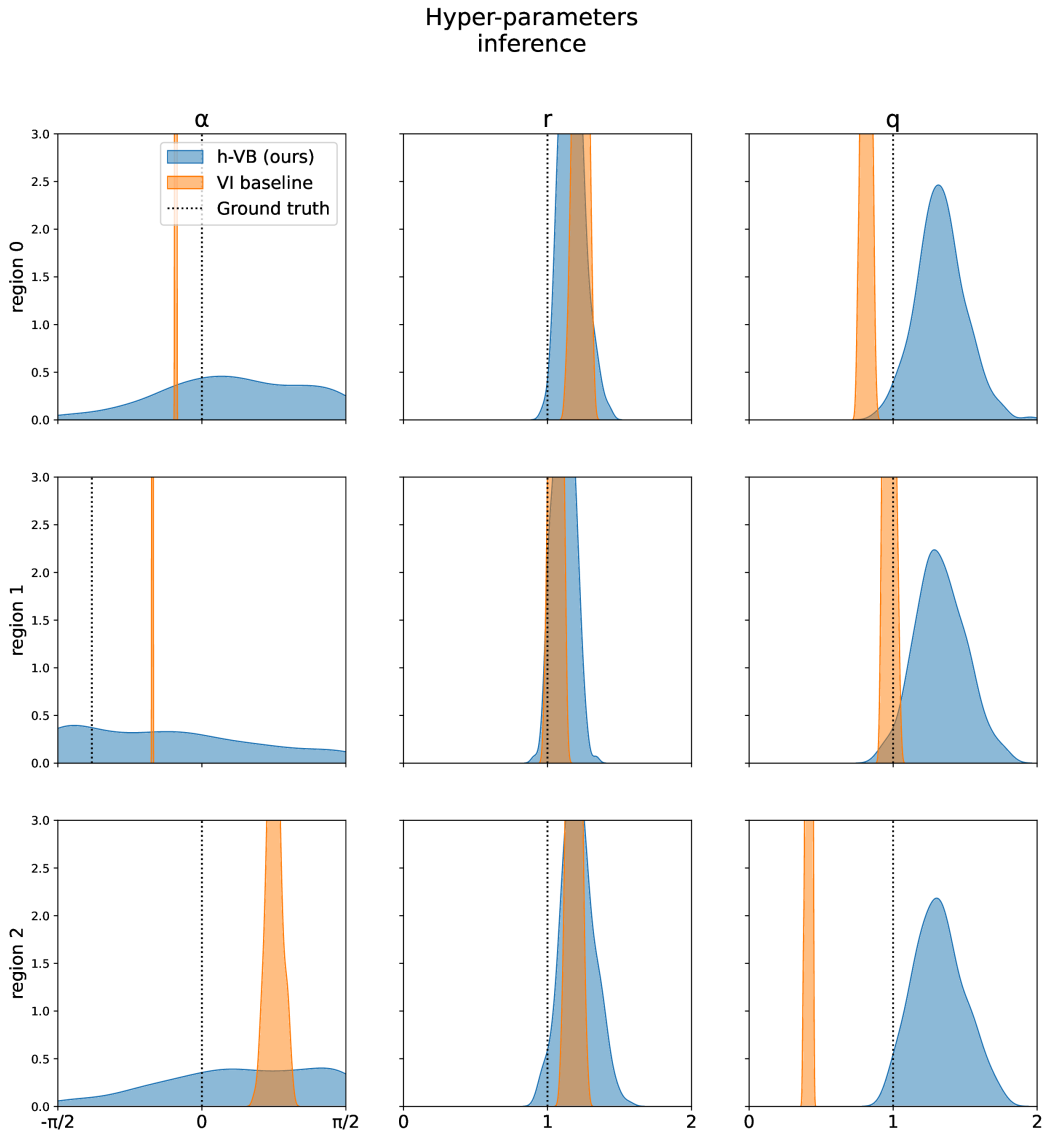

Fig. 9.4.: Mode collapse synthetic example: HRF and noise levels inference We display the posterior distributions of the hyper-parameters —as described in Section 9.2.1. Each line corresponds to a different region and each column to a different parameter: α (which conditions the HRF $\mathbf{H}$) and the variance levels $\boldsymbol{q}$ and $\boldsymbol{r}$. An off-the-shelf inference method (in orange) features mode collapse, outputing peaked distributions on a subset of the solution space —as described in Section 3.4. On the contrary, MDSI-h-VB (in blue) recovers the full support of the posterior distribution. As a sanity check, we see that the ground truth parameters (dashed lines) fall within the MDSI-h-VB's posterior but are missed by the baseline.

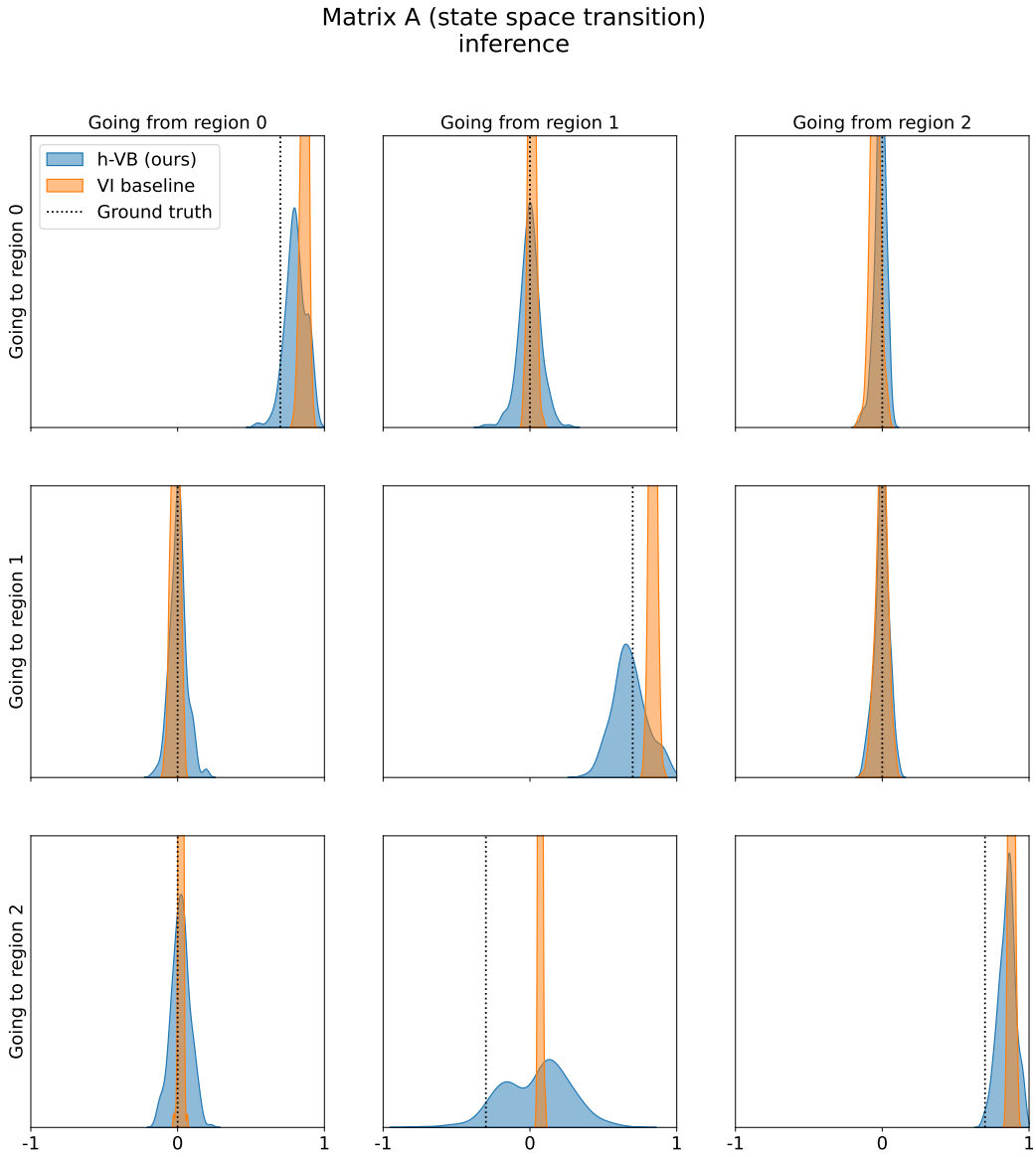

Fig. 9.5.: Mode collapse synthetic example: coupling inference We display the posterior distributions of the coupling matrix **A** —as described in Section 9.2.1. The matrix **A** is a 3×3 matrix, containing the coupling from every region (columns) to every region (lines). As shown in Figure 9.4, the baseline (in orange) features mode collapse and misses some of the solution space for the HRFs **H**. As a result, the baseline outputs a very narrow posterior for **A**, focusing on specific posterior modes. This is particularly visible for inferring the coupling coefficient from region 1 to region 2 (bottom center plot). MDSI-h-VB (in blue) recovers a bi-modal distribution, whereas the baseline collapses in only the positive mode, thereby missing the correct negative coupling (dashed line). In a downstream analysis, the baseline would spuriously output the existence of a weak positive coupling $1 \rightarrow 2$ with strong confidence. In contrast, MDSI-h-VB would also output the possibility of a negative $1 \rightarrow 2$ coupling and show that inferring the sign of the $1 \rightarrow 2$ coupling remains inconclusive.

exist between the 2 regions, inferring its sign is inconclusive.

In this experiment, we showed that off-the-shelf inference methods, though featuring a probabilistic output, can lead to over-estimated statistical confidence and spurious results. MDSI-h-VB, on the contrary, recovers the true uncertainty in the problem and can lead to more nuanced and richer conclusions.

9.3.2 Mode collapse in practice: effect on ground truth coupling coverage

This experiment validates statistically the effect of mode collapse illustrated in Section 9.3.1.

Data We generate a synthetic dataset using the MDS model —described in Section 9.2.1. We generate 20 random networks with 5 regions, each network associated with a different sparse coupling matrix. Non-diagonal elements of **A** have a 70% chance to be null, 20% to be 0.2, and a 10% chance to be -0.2. For each network, we simulate 10 "subjects", corresponding to independent runs of the MDS model with the same coupling matrix.

Baseline We use the same r-KL baseline as described in Section 9.3.1. In addition, we compare to r-DCM, a recent scalable extension of DCM (Frässle, Lomakina, et al., 2017; Frässle and Stephan, 2022). r-DCM uses a similar linear-coupling modeling as in the MDS model described in Equation (9.5). To invert its model, r-DCM uses Fourier analysis and Bayesian linear regression. One major difference with MDSI-h-VB is that r-DCM does *not* take into account HRF variability, and assumes that every region is associated with the default HRF. Mis-specification of the HRF is identified by Frässle, Lomakina, et al. (2017) as one of their method's main limitations.

Metric We leverage the probabilistic output of the compared methods. Once the posterior is fitted, we compute the log density over the off-diagonal coupling coefficients. This metric translates if the ground truth is statistically contained in the posterior distribution.

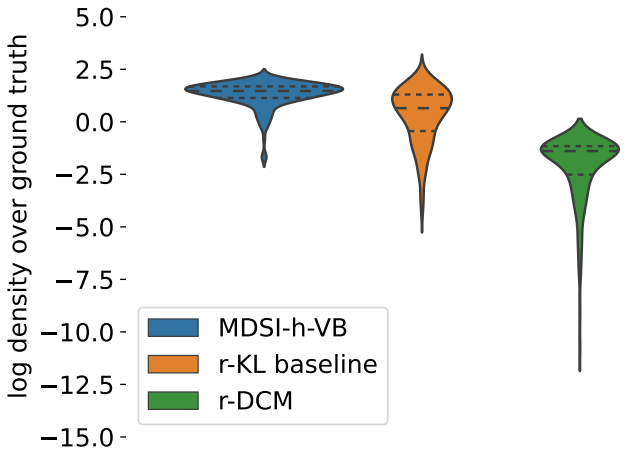

Fig. 9.6.: Effect of mode collapse on ground truth coupling posterior coverage We report the posterior log density over the off-diagonal ground truth coupling coefficient values. MDSI-h-VB (left) has a superior coverage of the ground truth. Baselines (middle and right) yield biased estimation, which results statistically in lower ground truth coverage. r-DCM (right) does not consider HRF variability, and hypothesizes the default HRF for every region. Yet, the HRF *does* vary across regions in this synthetic dataset. This results in biased r-DCM coupling estimation. Misspecification of the HRF is identified by Frässle, Lomakina, et al. (2017) as one of their method's main limitations. However, naively integrating the HRF variability also results in biased results, this time because of mode collapse, as illustrated with the r-KL baseline (middle). Due to mode collapse, the r-KL baseline misses parts of the true posterior's support and (statistically) the ground truth coupling value. *Performance is averaged over 10 independent runs over the same coupling matrix, across 20 networks. We report the* log *density over the ground truth* **A** *off-diagonal coefficients (ignoring self-coupling).*

MDSI-h-VB recovers the ground truth coupling more reliably Results are visible in Figure 9.6. By taking into account HRF variability, yet avoiding mode collapse, MDSI-h-VB covers the full support of the posterior for the coupling matrix **A**. This posterior thus contains the ground truth coupling value. In contrast, the baselines feature more peaked posteriors that tend to "miss" the ground truth —as illustrated in Figure 9.5. The baseline's posterior density over the ground truth is thus lower than for MDSI-h-VB.

This experiment shows that off-the-shelf inference can statistically miss the ground truth. Our hybrid method prevents this degenerate behavior.

9.3.3 Application on a neurophysiological synthetic dataset: connection detection

The goal of this experiment is to validate our method on samples coming from a different generative model than the MDS. The ground truth coupling is binary: either there is a positive coupling between regions, or there is no coupling (the strength of the coupling does not vary). As a result, we test our method in terms of the accuracy of connection detection.

Data We use synthetic data sampled using a neurophysiological process (Sanchez-Romero et al., 2018). Underlying neural dynamics are simulated using the linear differential equation $\partial z/\partial t = \sigma \mathbf{A} z + Cu$, where **A** denotes the ground-truth connectivity. To simulate resting-state data, the u input was modeled using a Poisson process for each of the regions. The neuronal signals z were then passed through the Balloon-Windkessel model (K. J. Friston, 2009) to obtain simulated BOLD data. The networks 1-9 feature small-scale synthetic graphs, which vary widely in their density and number of cycles. The SmallDegree and FullDegree networks consist of two larger graphs extracted from the macaque connectome.

Baseline We compare ourselves to a state-of-the-art directional coupling estimation method: regression dynamical causal modeling (r-DCM) introduced in Section 9.3.2 (Frässle, Lomakina, et al., 2017; Frässle and Stephan, 2022). r-DCM has been designed with scalability in mind, to be applied in the context of full-brain analysis.

Tab. 9.1.: **Physiological synthetic model: connection detection area under the curve (AUC)** We use the inferred off-diagonal coupling matrix **A** mean coefficient as data. We compute the t-score of the data across subjects and feed the score to a binary classifier. We report the AUC of the classifier. Note that this dataset does *not* feature any variability in the HRF. This implies that the performance of MDSI-h-VB is competitive with the one of r-DCM even in the default HRF case, where MDSI-h-VB's marginalization of the HRF is an over-parametrization.

Network	Number of nodes	MDSI-h-VB AUC (ours)	r-DCM AUC
1	5	0.82	**0.92**
2	5	0.79	**0.92**
3	5	**0.95**	0.88
4	10	**0.94**	0.83
5	5	**0.91**	0.70
6	8	**0.93**	0.88
7	6	**0.82**	0.72
8	8	**0.89**	0.78
9	9	0.82	**0.87**
SmallDegree	28	**0.92**	0.76
FullDegree	91	**0.90**	0.89
mean		**0.88** ($\pm$0.05)	0.83 ($\pm$0.08)

Method For each method, network and subject, we infer the mean value of the coupling matrix **A** posterior. For each coefficient, we then compute a t-score across subjects. We then feed that score to a binary logistic regression classifier. We report the area under the curve (AUC) of the classifier.

MDSI-h-VB connection detection accuracy is maintained as the number of nodes augments Table 9.1 reports the connection detection AUC of MDSI-h-VB as the number of nodes in the network augments. Both the SMallDegree and FullDegree cases feature several dozen nodes. In addition, their ground truth connections are based on axonal connectivity derived from tracer injection studies (Sanchez-Romero et al., 2018). As a result, the FullDegree setup is a good proxy for the performance of MDSI-h-VB on a full brain analysis as in Sections 9.3.4 and 9.3.5. In this challenging setup, MDSI-h-VB maintains an AUC of 0.90.

9.3.4 Full-brain directional coupling estimation in human working memory

This section scales up our method to the full brain. We consider a dataset of 737 subjects from the HCP (Van Essen et al., 2012). Subjects perform around 5 minutes-

long working memory tasks, that test their ability to temporarily memorize events. Presented with a series of pictures, subjects need to determine whether the current picture matches the first picture in the series (0-back task) or the picture 2 items before (2-back task). There thus are $C = 3$ experimental conditions (baseline, 0-back task, and 2-back task).

To study directional coupling, a previous instance of MDS identification (MDSI) extracted the data from 11 pre-determined regions (Cai, Ryali, et al., 2021). Those working-memory-associated regions were determined using prior expert knowledge. In particular, Cai, Ryali, et al. (2021) underlined the driving role of the right anterior insula (r-AI) as a causal outflow hub during working memory loads (Cai, T. Chen, et al., 2016; A. C. Chen et al., 2013). A limitation of this analysis is the restriction to pre-selected regions:

- important unobserved regions could be missed;
- potentially, those unobserved regions could drive the activity of the observed regions, and confound the results;
- the pathways through which the r-AI modulates the activity of other regions cannot be investigated.

In this work, we tackle these issues by reproducing the analysis from Cai, Ryali, et al. (2021) on the full brain. To this end, we use a brain parcellation: the `Brainnetome` (Fan et al., 2016) —see Chapter 7 for a definition of parcellation. The `Brainnetome` divides the brain into 246 regions. We extract the time series of those parcels and apply the MDSI-h-VB. This represents a coupling matrix of around 60,000 coefficients, which we can robustly estimate thanks to the scalability of MDSI-h-VB, as exposed in Section 9.2.2.

Full-brain estimation confirms the driving role of the r-AI in a data-driven manner

To assess the driving role of regions in working memory, we compute each region's *directed outflow* (Cai, Ryali, et al., 2021).

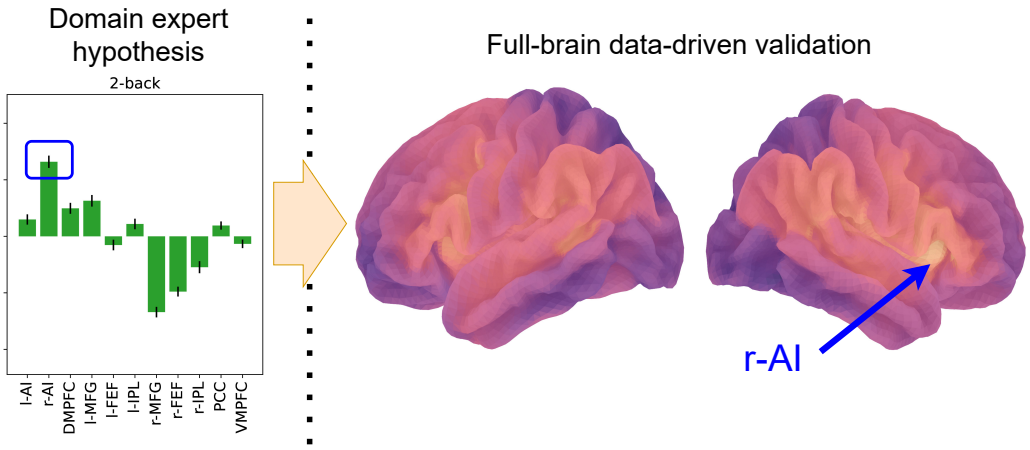

Fig. 9.7.: Full-brain directed outflow outlines the r-AI as a driving hub in working memory *On the left:* 2-back directed outflow analysis on 11 pre-selected regions. We use the same expert-selected regions as in Cai, Ryali, et al. (2021). *On the right:* 2-back full-brain directed outflow. The r-AI is hypothesized to be a driving region in the 11-regions analysis (blue rectangle). The full-brain analysis confirms this analysis: the r-AI (blue arrow) appears as a hot spot of the directed outflow. *Error bars in the 11 regions case represent the standard error across subjects. Only the mean outflow is represented in the full brain case.*

Method We define the directed outflow for a region m as the sum of the outwards coefficients minus the sum of the inwards coefficients:

$$\forall c = 1..\mathrm{C}\ \forall m_1 = 1..\mathrm{M}: \quad \text{directed\_outflow}_{c,m_1} = \sum_{m_2=1..\mathrm{M}, m_2 \neq m_1} a_{c,m_2,m_1} - \sum_{m_2=1..\mathrm{M}, m_2 \neq m_1} a_{c,m_1,m_2} \tag{9.11}$$

where the a denote the coupling values from the **A** matrix. The directed outflow quantifies whether a region activates the rest of the brain more strongly than the rest of the brain activates it. For every subject, we compute the directed outflow:

1. in the 11 regions case, using the same data as Cai, Ryali, et al. (2021);

2. in the 246 regions case.

Results Figure 9.7 displays the directed outflow obtained in both region decompositions. The full-brain analysis confirms the findings of the pre-selected regions analysis, in a data-driven manner, while mitigating the risk of confounds.

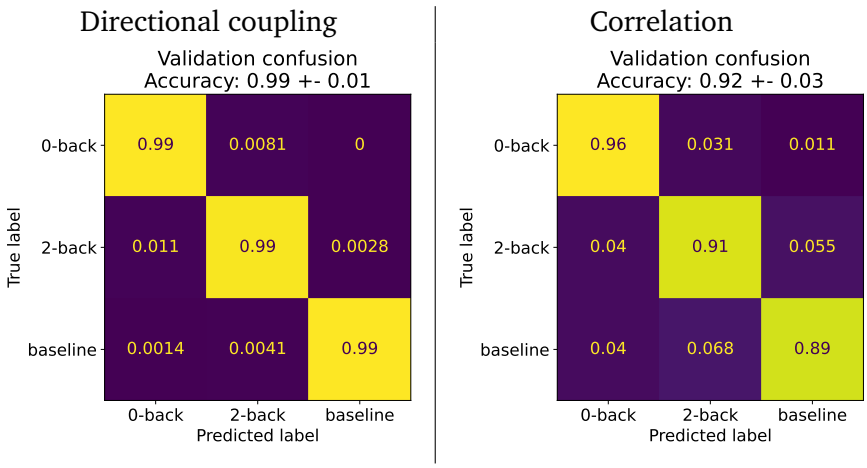

Fig. 9.8.: Task classification using directional coupling and correlation Confusion matrices for task classification (baseline, 0-back or 2-back). Directional coupling (on the left) appears more task-specific than correlation (on the right). *Performance averaged across a 10-fold population bootstraps. We report the average confusion across folds, the average accuracy, and the standard deviation of the accuracy across folds.*

Directional coupling exhibits higher task-specificity compared to correlation

We predict which experimental condition the subjects undergo (baseline, 0-back, or 2-back) using the connectivity as feature. Can we predict which activity a subject is doing based on the patterns of connectivity in the brain?

Method To compute a task classification accuracy, we use the matrices $\mathbf{A}_c$ as features and the associated conditions c as labels. We perform a group K-fold cross-validation using separate subjects in the training and validation sets. We report the mean classification confusion across 10 splits (each time leaving around 73 subjects out). As classifier, we use a logistic regression with L2 regularization. As baseline, we reproduce the same analysis, instead using the correlation instead of the directional coupling as feature. For each subject, considering the time series of each Brainnetome region $m = 1..\mathrm{M}$ during each condition $c = 1..\mathrm{C}$, we computed a $\mathrm{C} \times \mathrm{M} \times \mathrm{M}$ Pearson correlation matrix. This baseline feature matrix has the same dimensionality as our coupling matrix $\mathbf{A}$.

Results Figure 9.8 reports the classification confusion using the coupling and the correlation as feature. Directional coupling appears more task-specific, allowing to almost perfectly predict the experimental condition out-of-sample subjects undergo.

Directional coupling estimation is noisier than correlation, leading to lower inter-session stability

We test out the stability of the subject-level directional coupling across different measurement sessions. The directional coupling is computed at a latent level (not directly at the BOLD level). This creates an uncertainty in its estimation. How much does this uncertainty affect the stability of the recovered coefficients compared to a simpler measure computed at the BOLD level (the correlation)?

Method Following Frässle and Stephan (2022), we implement the intraclass correlation coefficient (ICC)(3,1) type. The ICC(3, 1) quantifies the ratio between the within-subject variability across the two sessions (σ_w) and the between-subject variability (σ_b). For a given condition c and coupling coefficient from region m_1 to m_2:

$$\forall c = 1..\mathrm{C} \quad \forall m_1 = 1..\mathrm{M} \quad \forall m_2 = 1..\mathrm{M} \quad m_2 \neq m_1 : \quad \mathrm{ICC}_{c,m_1,m_2} = \frac{\sigma_b^2 - \sigma_w^2}{\sigma_b^2 + \sigma_w^2} \tag{9.12}$$

In the ICC analysis, we ignore the diagonal of the coupling matrix **A** (self-coupling). As a baseline, we reproduce the same analysis using instead the Pearson correlation as input.

Results and discussion Figure 9.9 shows the lower inter-session stability of the directional coupling compared to the correlation. This is an expected result: partial correlation is harder to estimate than standard correlation (Van Den Heuvel and Pol, 2010; Varoquaux and Craddock, 2013). We interpret this as further motivation for hierarchical modeling: estimating coupling in multi-session, multi-subject models could help overcome noise in the estimation.

Directional coupling analysis unveils activation pathways in working memory

In this section, we underline the ability of directional coupling to recover directed functional pathways. We tackle the example illustrated in Figure 9.1 (left). In the case of an $A \rightarrow C \rightarrow B$ network, all three regions will have a correlated signal and be identified as parts of the same network. But correlation doesn't inform us about who activates whom inside this network. In contrast, we show here that the directional coupling unveils such activation pathways.

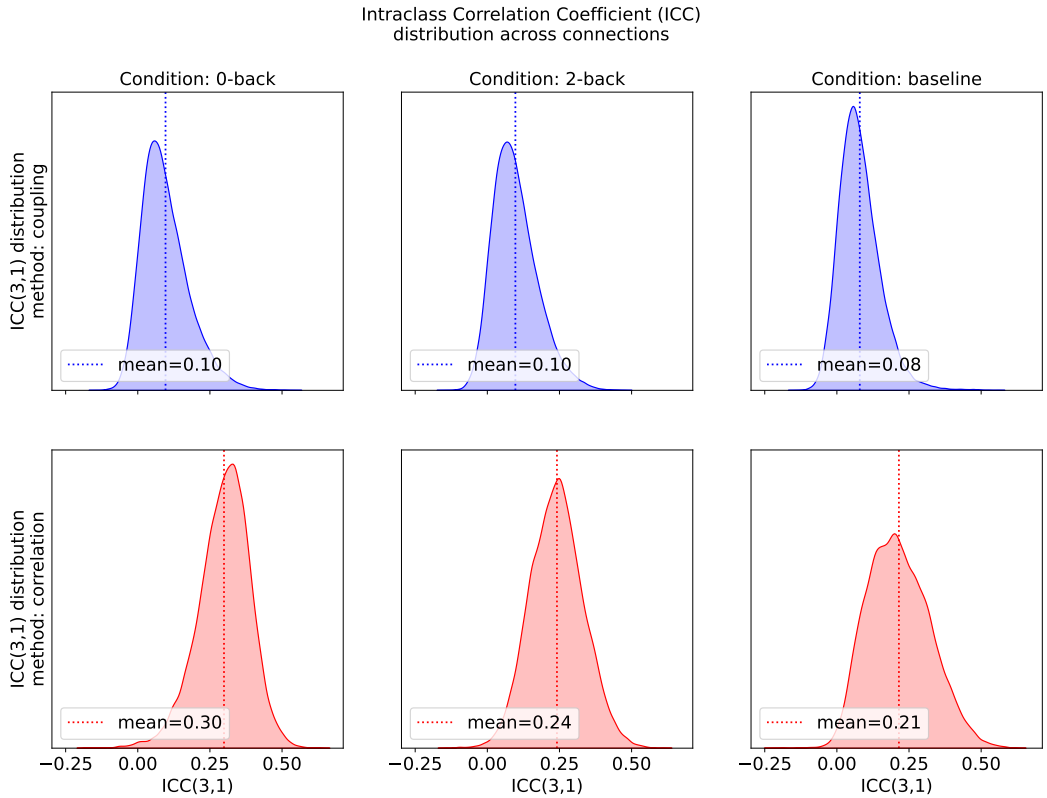

Fig. 9.9.: Intraclass correlation coefficient (ICC) using directional coupling and correlation On the top: inter-session stability using the coupling as feature. On the bottom, inter-session stability using the correlation. The directional coupling is akin to a *partial* correlation and is more technically involved to derive. This results in noisier estimates and lower inter-session stability. *ICC computed across 2 measurement sessions, for 737 subjects. We report the ICC distribution across the 60,000 coupling/correlation coefficients.*

Method We consider the $\mathbf{A} \in \mathbb{R}^{246\times246}$ coupling matrix as a matrix of direct connections. $\mathbf{A}^{\text{ind}} = \mathbf{A} \times \mathbf{A}$ represents the matrix of pairwise region couplings after one "hop" through other regions (we set the diagonal of $\mathbf{A}$ to zero to remove the effect of delayed direct coupling). We construct a filtered coupling matrix $\mathbf{A}^{\text{ind,filt}}$, keeping only the pairs of regions with a stronger indirect connection than their direct counterpart:

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} a_{0,0} & \dots & a_{0,\text{M}} \\ \vdots & \ddots & \vdots \\ a_{\text{M},0} & \dots & a_{\text{M},\text{M}} \end{bmatrix} \\
\mathbf{A}^{\text{ind}} = \mathbf{A} \times \mathbf{A} &= \begin{bmatrix} a^{\text{ind}}_{0,0} & \dots & a^{\text{ind}}_{0,\text{M}} \\ \vdots & \ddots & \vdots \\ a^{\text{ind}}_{\text{M},0} & \dots & a^{\text{ind}}_{\text{M},\text{M}} \end{bmatrix} \\
\mathbf{A}^{\text{ind,filt}} &= \begin{bmatrix} a^{\text{ind,filt}}_{0,0} & \dots & a^{\text{ind,filt}}_{0,\text{M}} \\ \vdots & \ddots & \vdots \\ a^{\text{ind,filt}}_{\text{M},0} & \dots & a^{\text{ind,filt}}_{\text{M},\text{M}} \end{bmatrix} \\
\forall m_1 = 1..\text{M} \quad \forall m_2 = 1..\text{M} : \qquad a^{\text{ind,filt}}_{m_1,m_2} &= a^{\text{ind}}_{m_1,m_2} \times \mathbb{1}_{a^{\text{ind}}_{m_1,m_2} \gg a_{m_1,m_2}}
\end{aligned} \tag{9.13}$$

where the condition of $a^{\text{ind}}_{m_1,m_2} \gg a_{m_1,m_2}$ is measured via a t-test over the subjects. To interpret the $\mathbf{A}^{\text{ind,filt}} \in \mathbb{R}^{246\times246}$ matrix, we aggregate it over the Yeo 7 functional networks (Thomas Yeo et al., 2011). To do so, we consider a source region m_1, and a target network $\mathcal{N}_{\text{target}}$. We compute the sum of the indirect coupling from m_1 to $\mathcal{N}_{\text{target}}$ as $\sum_{m_2\in\mathcal{N}_{\text{target}}} a^{\text{ind,filt}}_{m_1,m_2}$. We compare this total indirect coupling to the indirect coupling from m_1 to other networks. By decomposing $\mathbf{A}^{\text{ind}} = \mathbf{A} \times \mathbf{A}$, we can also compute the proportion of the indirect coupling of m_1 to $\mathcal{N}_{\text{target}}$ that passes through each possible mediating network $\mathcal{N}_{\text{mediator}}$.

Results Figure 9.10 illustrates the indirect coupling analysis over a constitutive region of the r-AI: the dorsal agranular insula. The analysis reveals an indirect influence of the r-AI over the default mode network, mediated by the frontoparietal network. This result reinforces the hypothesis that the r-AI acts as a "network switch" in working memory, as suggested by its role in stop signal, Flanker, and oddball tasks (Cai, T. Chen, et al., 2016; A. C. Chen et al., 2013).

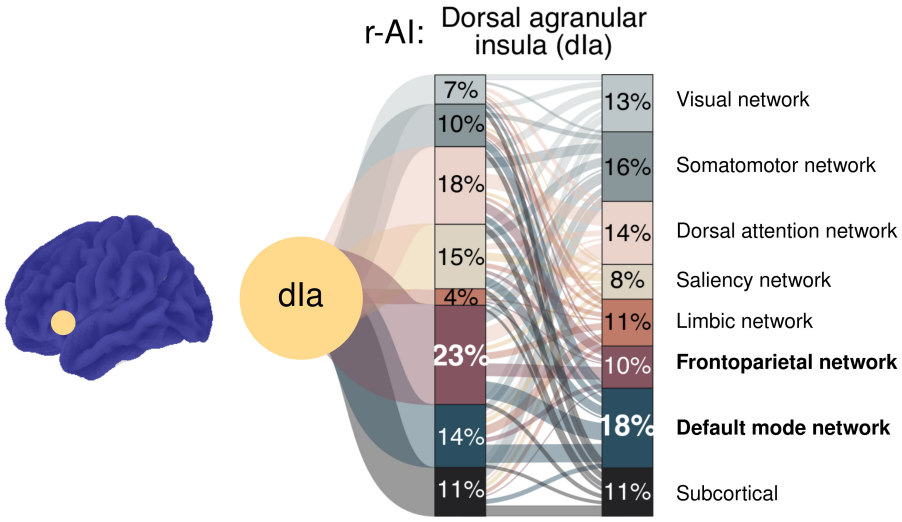

Fig. 9.10.: The r-AI drives the default mode network through the frontoparietal network We average the 2-back directional coupling coefficients across major networks (Thomas Yeo et al., 2011). We then perform a betweenness analysis, testing for networks with an indirect coupling —"hopping" through other regions— stronger than a direct coupling. This analysis confirms a strong influence of the r-AI over the default mode network, mediated by the frontoparietal network (Cai, Ryali, et al., 2021). The directional coupling unveils complex functional pathways that would be missed using the correlation.

9.3.5 Full-brain directional coupling estimation in human resting state

While Section 9.3.4 investigated human working memory, this section studies resting state fMRI —as introduced in Chapter 7. By targeting the resting state, we investigate the "default" functional organization of the brain, which can be modulated as part of various tasks (Simon B. Eickhoff et al., 2018b). To this end, we consider time series from 1,089 subjects from the HCP (Van Essen et al., 2012). Each time series is 15 minutes long, during which subjects are at rest, and do not perform any specific activity —so there is only a single condition $C = 1$. We apply MDSI-h-VB to these time series, unveiling the baseline directional coupling of the human brain.

Directional coupling opens the box of known functional networks

In this exploratory section, we shed some light on directional coupling's potential for parcellation —the target application from Chapter 7. From the coupling matrix, we aggregate parcels into macroscopic networks and put forward the similarity with correlation-based networks. This suggests that directional coupling complements

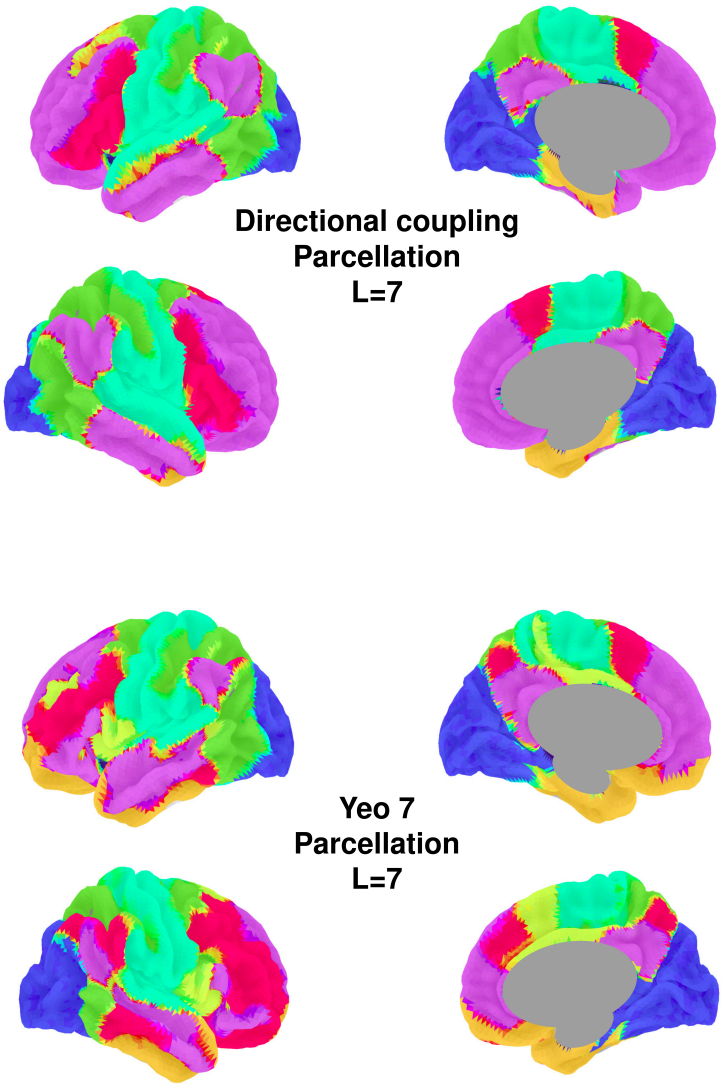

Fig. 9.11.: **Directional coupling confirms known functional networks, while unveiling the finer role of regions as part of those networks** *On top:* functional networks obtained by hierarchically clustering the directional coupling matrix. *On the bottom:* Yeo 7 networks (Thomas Yeo et al., 2011). Directional coupling yields similar major networks, with main differences located on the temporal lobes and in the frontoparietal network (in yellow). Both clusterings over the Brainnetome parcels have a Fowlkes-Maslow score of 0.47. Using directional coupling, it is possible to unveil the precise role of sub-regions as part of those macroscopic networks. *Note:* the directional coupling networks also have strong similarities with the population parcellation from Figure 7.3.

existing knowledge about the functional organization of the brain. We recover known integrative structures while uncovering the finer roles of sub-regions constituting those structures.

Method We first compute a directed, weighted adjacency matrix A using the coupling as input:

$$A \propto \max(0, \mathbf{A}) \tag{9.14}$$

Where we ignore negative coupling, as negative correlation is typically ignored in functional connectivity (Van Den Heuvel and Pol, 2010; Varoquaux and Craddock, 2013). We cluster the Brainnetome parcels hierarchically using a Louvain algorithm (Dugué and Perez, 2015). This agglomerative clustering yields a dendrogram, allowing us to group parcels in any number of disjoint clusters from 1 (all components grouped) to 246 (1 component per cluster).

Results Inspired by Thomas Yeo et al. (2011), we parcel the brain into 7 directional-coupling-based networks. Results are visible in Figure 9.11. Directed coupling yields a macroscopic functional organization close to the one based on correlation. At the same time, directed coupling allows to "open the box" of those networks, and to investigate the precise function of their constitutive sub-regions.

Resting-state directional coupling yields superior cognition/behavior prediction compared to correlation

This experiment tests out the predictive power of directional coupling on behavior and cognition. Can functional connectivity inform us about the cognitive ability of a subject?

Method We perform a 20-fold cross-validation across subjects. We use a ridge regression, predicting the scores of held-out subjects (54 subjects). As in Section 7.3.2, we report the correlation of the predicted score with the true score. We repeat this process 10 times and report the distribution (across repetitions) of the average performance (across folds).

Results Results are visible in Figure 9.12. Directional coupling yields significantly better cognitive/behavioral scores prediction. This suggests that the coupling is a richer description of an individual subject's functional connectivity.

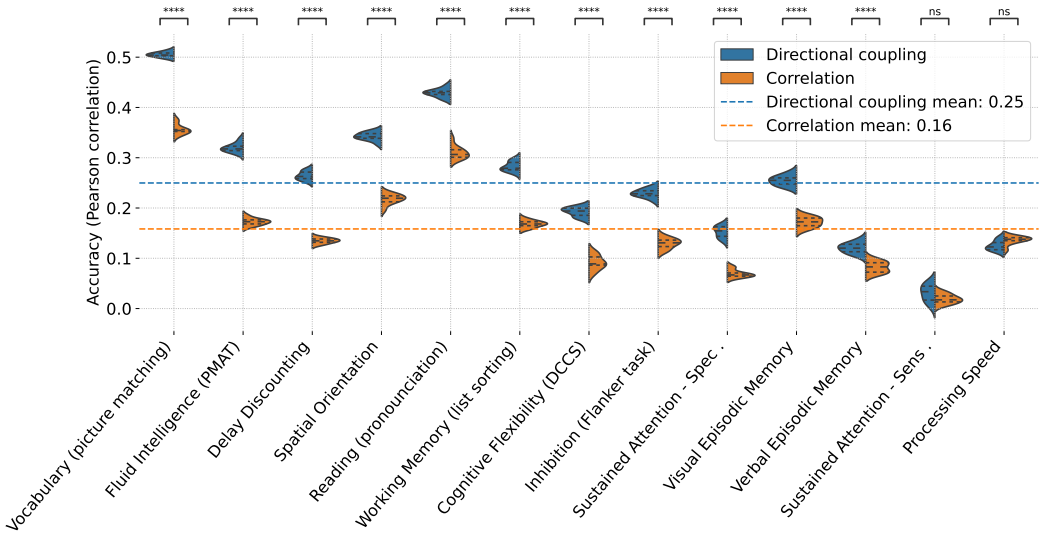

Fig. 9.12.: Directional coupling yields superior cognitive/behavioral scores prediction We predict the cognitive/behavioral scores of held-out subjects, either using the directional coupling or the correlation as feature. We use compare over the same reference 13 scores as in Table 7.1. Correlation yields a similar performance (0.16) as the baselines in Table 7.1 (Kong, J. Li, et al., 2019; Kong, Q. Yang, et al., 2021; Calhoun and Adali, 2012; E. M. Gordon et al., 2017; Danhong Wang et al., 2015; Rouillard, Bris, et al., 2023). In contrast, directional coupling yields a +0.9 increased correlation. *Reported accuracy is the correlation of the predicted scores with the true score. Performance averaged across 10 independent 20-fold cross-validations. We report the distribution of the performance across the independent cross-validations. Statistical confidence is measured via t-tests with Bonferroni correction.*

9.4 Summary of contributions

This chapter summarized our contributions in fMRI-based directional coupling estimation:

- We design a novel *hybrid* inference scheme for the MDS model: MDSI-h-VB. We inject a plate-amortized, f-KL-trained estimator inside a r-KL-trained variational family as a means to prevent mode collapse. In doing so, we properly marginalize the uncertainty in the HRF when estimating the directional coupling across regions. The uncertainty in the HRF is considered a major challenge in the community (Handwerker et al., 2004; Rangaprakash, Barry, et al., 2023; Rangaprakash, G.-R. Wu, et al., 2018).

- We demonstrate the robustness of MDSI-h-VB against baselines, in a variety of synthetic scenarios. Critically, we show that our method reliably captures the ground truth coupling value and that its accuracy does not degrade when applied to hundreds of regions.

- Leveraging modern GPU-accelerated ADVI, we scale up MDSI to the full brain (full-brain causal modeling was already introduced by Frässle, Lomakina, et al. (2017)). Compared to the estimation over a dozen pre-selected regions, full-brain analysis:

  1. does not risk ignoring important unknown regions,

  2. does not incur confounding by unobserved regions.

- Using a full-brain analysis, we confirm the driving role of the r-AI in working memory, in a data-driven manner. We also uncover novel activation pathways, showing the influence of the r-AI over the default mode network mediated by the frontoparietal network.

- Compared to correlation, we demonstrate the superior predictive power of directional coupling in:

  1. task classification,

  2. cognitive/behavioral scores prediction.

  In that regard, we obtain state-of-the-art out-of-sample cognition prediction (to the best of our knowledge, using functional connectivity as a feature).

- We illustrate the ability of directional coupling to recover known functional networks in the brain. At the same time, we allow to open the box of those networks, to unveil the finer coupling between regions with a correlated signal.

Part conclusion

This part reviewed our contributions.

Chapter 6 presented our methodological inputs. We designed variational families that are both expressive (to minimize the approximation gap) and scalable (to tackle large-scale inference). To this end, we leverage the NF/encoding couple at the core of SBI into structured, stochastic VI. We then applied this methodology to three Neuroimaging problems, each a different instantiation of the parameters/model/signal tryptic.

In Chapter 7, we applied a multi-hierarchy GM (the model) over functional connectomes (the signal) to yield individual parcellations (the parameters). For this application, we trained a PAVI variational family using the r-KL, illustrating the applicability of AVI over a complex, large-scale problem. We obtained individual probabilistic parcellations and demonstrated their predictive power over cognition.

In Chapter 8, we applied a combination of the SM with a latent GM (the model) over dMRI (the signal) to estimate microstructure (the parameters). In this application, we went one step further in leveraging the interplay between PAVI and SBI. We used NPE to build a composite model, plugging an explicit prior on top of a surrogate to the implicit distribution defined by a simulator. We then utilized the f-KL-learned, single-voxel NF/encoding couple as a powerful initialization for the full-brain variational family. Next, we trained this variational family using the scalable r-KL. In doing so, we reduced the uncertainty in microstructure estimation by injecting a meaningful hypothesis.

In Chapter 9, we applied the MDS (the model) over fMRI time series (the signal) to estimate the directional coupling matrix (the parameters). In this application, we repurposed the tools from SBI to make inference more reliable. Even in the presence of an explicit HBM, we used NPE over a set of key hyperparameters. We then plate amortized this f-KL-trained estimator inside a scalable, r-KL-trained variational family. This prevented mode collapse in the coupling estimation, illustrating the usefulness of the f-KL beyond implicit distribution learning. Our hybrid inference method allowed us to both robustify and scale up MDSI to the full brain. We uncovered novel functional pathways and demonstrated the relevance of directional coupling over cognition.

Throughout our applications, we combined tools from SBI, hierarchical modeling and VI. In doing so, we built hybrid inference methods, constructing composite HBMs P and variational families $\mathcal{Q}$. In the next and final part, we discuss some challenges that we encountered across our different applications.

Part IV

Open questions

10 Discussion: large-scale hierarchical Bayesian inference

In this short chapter, we take a look back at some of the challenges encountered during this thesis. Our goal is to provide a general discussion on the methods to tackle those hurdles. We build up this way to this thesis' conclusion.

10.1 Leveraging the f-KL in large-scale AVI

In Chapter 9, we illustrated a failure mode of the r-KL. Due to mode collapse, a VI baseline misses entire parts of the solution space (in this case, the coupling matrix), resulting in variance under-estimation and in biased results —as illustrated in Figure 9.5. Taking a step back from the directional coupling application, our methodology to tackle this issue is to:

1. Identify the set of key parameters responsible for this mode collapse;
2. Train a f-KL amortized estimator to infer those key hyper-parameters;
3. Inject this estimator into a scalable r-KL scheme.

Critically, the second step was possible because the key parameters (the HRF and noise levels) were low-dimensional —or rather, we made some independence assumption (across regions) to bring ourselves to this low-dimensional case.

Though generalizable to other situations, this method requires significant domain knowledge. The key parameters need to be identified, as well as the associated repeated structure —such as the independence across regions. As a consequence, **our methodology does not natively fit with automatic variational inference (AVI)**.

What avenues exist to benefit from the f-KL in a more systematic fashion?

Enrich the model declaration In Section 3.4.3, we sketched the contours of AVI APIs that would simply require declaring the inference problem —while its resolution would be automated. One possibility to leverage the f-KL would be for the experimenter to "flag" the problematical parameters as part of declaring the model. Such parameters are not necessarily complicated to guess: as an example, second-order moments are typically more troublesome to infer than first-order moments (Bishop, 2006). Then, systematically leveraging the model's plate structure would yield a reduced parameter space over which to train using the r-KL, while sampling the model itself would yield a synthetic dataset —providing all the ingredients necessary to reproduce the steps from Chapter 9.

Hybridising MCMC and VI in hierarchical models Another interesting avenue of research is "hybrid" inference methods combining MCMC and VI (Matthew D. Hoffman, Sountsov, et al., 2019; Grenioux et al., 2023). As already introduced in Section 3.2, those methods rely on a Markov chain providing samples to train a variational family using the f-KL. The variational family itself is used as part of an efficient kernel to explore the parameter space. However, to our knowledge, those methods have only ever been used on the entirety of the parameter space at once, and scale poorly with the dimensionality of Θ.

Could similar methods be envisioned in large-scale, hierarchical cases?

One sketch would be to combine SVI with Gibbs sampling (Andrieu et al., 2003). Both methods rely on inferring only over sub-parts of the parameter space at once. In this scheme, every ground RV would be associated with its own Markov chain, updated as part of Gibbs sampling. In addition, every ground RV would also be associated with a conditional density estimator —the NF-based $q_{i,n}$ described in Section 6.1.2. This density estimator could be trained using the f-KL, using samples from the Markov chain. Conversely, the NF-based estimator could be used as part of a kernel to propose the next sample of the ground RV's Markov chain.

Note: This training scheme is compatible with the r-KL training described in Section 6.2.1. Both could be combined as part of a *symmetrical* KL training.

Leveraging sequential methods In Section 4.2.2, we introduced the *sequential* variant of NPE (Papamakarios and Murray, 2016; Greenberg et al., 2019). The core idea is to only *locally* amortize a f-KL-trained estimator. We could leverage those techniques in the context of HBMs. An amortized PAVI-E family (see Section 6.2.3)

could be sequentially trained, using a combination of both the f-KL —using samples from the (conditional) model— and the r-KL.

A first open question is the computational efficiency of this sequential scheme. As pointed out in Section 3.3.2, amortization can impede inference in large dimensions. It is unclear whether the addition of the r-KL loss improves the training's sample efficiency. Another open question is *how* exactly to use the model to generate new "posterior predictive" samples —as described in Section 4.2.2. Contrary to the black box simulators in SBI, in the case of a HBM, the entirety of the parameter space is both inferred upon and represented using (conditional) distributions. In this context, what does it mean to "sample new synthetic signal from the simulator"? Which RVs should be considered fixed? Which RVs should be resampled?

10.2 Leveraging Bayesian theory in downstream analysis

Through this thesis, we strive to obtain full probability distributions for latent parameters and to reduce uncertainty via hierarchical modeling. Yet we do not fully exploit the richness of these elements. This section reviews some avenues for improvement of our work to shed light on practices in the community.

Held-out data validation and hierarchical modeling How to assess statistical confidence over results? Historically, this issue has been addressed through frequentist null hypothesis testing or Bayesian information criteria (Kruschke and Liddell, 2018). Those methods have in common that they consider the entirety of the available data at once. Yet, traditionally "good" in-sample measures such as low p-values can be misleading in data-rich, high-dimensional settings (Bzdok and Yeo, 2017; Efron and Hastie, 2021). In particular, models that provide a good *explanation* of the observed data will not necessarily provide good *generalization* over unseen data. Such out-of-sample generalization has become the de-facto standard in the machine learning community (Efron and Hastie, 2021; Bishop, 2006). One hurdle encountered during this thesis is the apparent gap between classical modeling —tied to classical statistical confidence evaluation— and out-of-sample generalization.

In particular, consider inference using a hierarchical model covering the entirety of the available data $\mathbf{X}^{\text{train}} \cup \mathbf{X}^{\text{val}}$ as a way to extract subject-specific features $\Theta^{\text{train}} \cup \Theta^{\text{val}}$. Training, in a second step, a machine learning model over Θ^{train} and validating over Θ^{val} induces *leakage*. Indeed, due to the hierarchical structure, the subject features

are *not* i.i.d: Θ^{val} has been used as part of computing Θ^{train}. How to circumvent this issue?

One solution is to infer twice: once over the $\mathbf{X}^{\text{train}}$ and once over $\mathbf{X}^{\text{val}}$, train a machine learning model over Θ^{train}, and evaluate it over Θ^{val}. Yet, in doing so, the features Θ^{val} —extracted over a smaller number of subjects— will be significantly noisier than Θ^{train}. This can lead to poor generalization performance.

In Chapter 7, we started experimenting with ways to improve out-of-sample generalization. By transferring the $\mathbf{X}^{\text{train}}$-fitted parameter posterior as a prior to infer over $\mathbf{X}^{\text{val}}$, we leverage the "learning" over the training data to reduce the noise in the validation features. This method builds upon the parallel between inference and machine learning (Jospin et al., 2022; Bishop, 2006). Using this analogy, "training" consists both in obtaining a prior for feature inference, *and* in fitting a machine learning model over the train-set features. We argue that similar methods could drastically improve performance in applications such as directional coupling estimation —described in Chapter 9. In Chapter 9, we eventually did not leverage hierarchical modeling due to the difficulty in developing downstream task validation. This underlines gaps in methodology and avenues for research: *how to rapidly repeat inference over data bootstraps?*

Probabilistic output and downstream tasks In this thesis, we derive full parameter probability distributions: $P(\Theta|X)$. Yet, in our downstream analysis, we oftentimes "reduce" those distributions into point estimates. For instance in Chapter 9, we use the coupling matrix posterior mean —see Sections 9.3.4 and 9.3.5. In Chapter 8, we display the microstructure parameter posterior mean in Figures 8.6 and 8.7. The reason for resorting to point estimates is that downstream tasks —be it hypothesis testing or the fitting of machine learning models— use points as inputs, not distributions. Does comparing to the existing literature —using broadly accepted metrics— require cutting down the richness of probabilistic outputs?

One counterpoint to that view is that full probability distributions offer the possibility to derive post-hoc a variety of point estimates, with different statistical properties. Where traditional optimization would usually target a ML or MAP estimate, the full distribution offers the possibility to compute the minimum mean square error (MMSE) estimator $\mathbb{E}(\Theta|X)$. In addition, second-order statistics such as the standard deviation can also be computed from the posterior, to inform on the uncertainty of point estimates —as illustrated in Figure 8.5. Finally, integrating distributions allows computing probabilities: taking Chapter 9 as an example, we could report the

probability of a positive coupling from a region A to another region B. Even reduced to point statistics, full distributions are thus worth deriving.

Note on PAVI encodings: As an opening, in Section 10.3, we underline the interaction in PAVI between the variational family $\mathcal{Q}$ and the signal encodings $\mathbf{E}_{i,n}$ (see Section 6.1.2). The encodings act as summary statistics, embedding all the information relevant to the inference over e.g. a given subject. Instead of using distribution statistics, could those encodings be used as input features for a machine-learning model?

Beyond point estimates, could better ways to leverage the posterior $P(\Theta|X)$ be explored? As an example, given the subject "data" $(y^{\text{subject}}, P(\Theta|X = \mathbf{X}^{\text{subject}}))$, where y denotes some covariate like the age, we could sample repeatedly from $P(\Theta|X = \mathbf{X}^{\text{subject}})$ to generate multiple data points $\{(y^{\text{subject}}, \Theta_1^{\text{subject}}), \dots, (y^{\text{subject}}, \Theta_N^{\text{subject}})\}$. Would the training of a machine learning model f over this "augmented" dataset — marginalizing over the feature posterior— yield interesting regularization? Similarly, we could marginalize prediction over the posterior to yield integrated estimators: $\mathbb{E}_{\Theta \sim P(\Theta|X=\mathbf{X}^{\text{subject}})}[f(\Theta)]$.

More generally, we argue there is an interest in *bridging the gap in the community's practices between full Bayesian inference and downstream task validation*.

Model selection and Bayes factor Another improvement in our analysis would be the addition of Bayesian model selection (Gelman, Carlin, et al., 2004; Kruschke and Liddell, 2018; K. J. Friston, 2009). Any model $P(X, \Theta)$, fitted over a data $\mathbf{X}$, will yield a posterior $P(\Theta|X = \mathbf{X})$. This does not mean that the model P is a good representation of the data $\mathbf{X}$. To validate P, our methodology of choice has been to correlate the model parameters Θ with some external covariates y —as an example cognitive scores in Chapters 7 and 9. The underlying assumption is that if Θ is predictive of y —which was not considered during the model fit— then Θ contains some meaningful biological signal, making P a "good" model.

A complementary analysis is the systematic comparison of different models $P_1, \dots, P_N$ over the same data $\mathbf{X}$. To compare two models P_1 and P_2, one can rely on the Bayes factor:

$$\begin{aligned} \mathcal{B}(P_1/P_2) &= \frac{P_1(X = \mathbf{X})}{P_2(X = \mathbf{X})} \\ &= \frac{\texttt{Pr}(P_1|\mathbf{X})}{\texttt{Pr}(P_2|\mathbf{X})} \div \frac{\texttt{Pr}(P_1)}{\texttt{Pr}(P_2)} \end{aligned} \tag{10.1}$$

where $\texttt{Pr}$ denotes a probability. The Bayes factor $\mathcal{B}$ quantifies the change from the prior model odds $\texttt{Pr}(P_1)/\texttt{Pr}(P_2)$ to the posterior model odds $\texttt{Pr}(P_1|\mathbf{X})/\texttt{Pr}(P_2|\mathbf{X})$

due to observing the data. Said differently, $\mathcal{B}$ denotes how much the observed data $\mathbf{X}$ weighs in favor of a model P_1 versus another model P_2 in the light of the evidence both models provide for $\mathbf{X}$. If the models P_1 and P_2 translate different hypotheses over the data $\mathbf{X}$, the Bayes factor provides a way to perform Bayesian hypothesis testing (Kruschke and Liddell, 2018).

Consider directional coupling estimation in Chapter 9 as an example. P_1 could be the MDS model —described in Section 9.2.1— while P_2 would be a similar model, but forcing the coupling matrix $\mathbf{A}$ to be diagonal —no coupling across different regions. $\mathcal{B}$ would then quantify how much evidence there is in favor of a non-null coupling across regions.

To compute $\mathcal{B}$, we need to compute both model evidences $P_1(X = \mathbf{X})$ and $P_2(X = \mathbf{X})$. In theory, this would require marginalizing over the entirety of their respective parameter spaces —which, as Section 3.1 underlines, is usually infeasible. However, VI provides a solution for this problem. Provided that the variational distribution matches the true posterior, the evidence lower bound (ELBO) approximately equals the model evidence. Computing the Bayes factor could thus be a byproduct of VI's optimization. As an alternative in the case where P_1 and P_2 share the same parameter support, K. Friston et al. (2018) propose to fit once a model using a non-informative prior $P_3(\Theta)$, and to integrate the prior ratio $P_1(\Theta)/P_2(\Theta)$ over the resulting posterior $P_3(\Theta|X = \mathbf{X})$. This allows estimating the evidence ratio between P_1 and P_2.

This thesis focused on out-of-sample generalization as a validation strategy. *Bayesian model selection appears as a computationally feasible addition that would further bridge the gap with traditional statistics.*

10.3 HBMs as parametric generative models

This thesis concentrated on the inference of the posterior distribution $P(\Theta|X)$. A byproduct of inference is the learning of the distribution of the observed signal. After inference, one can draw latent parameters Θ from the posterior, and use those to generate a new synthetic signal X. This corresponds to sampling the posterior predictive distribution:

$$P(X, \Theta|\mathbf{X}^{\text{observed}}) = P(X|\Theta) \times P(\Theta|X = \mathbf{X}^{\text{observed}}) \tag{10.2}$$

HBMs are thus generative models, that can learn from observed data, and generate new data (Bond-Taylor et al., 2022), with a few perks and caveats:

- HBMs are *parametric* models. Though their number of parameters can adapt to the quantity of available data, HBMs rely on experimenters specifying distributions. This property is both a blessing and a curse:

  - parametric models can suffer from *model misspecification*. Making a parallel with the approximation gap described in Section 3.3.1, the parametric distribution $P(X)$ is limited in its expressivity, and can more or less closely match the *true* distribution of the observed data $\mathbf{X}$. As a result, HBMs fall far from the non-parametric state-of-the-art when it comes to generating good copies of existing data (Bond-Taylor et al., 2022)

  - on the other hand, HBMs are *interpretable* models, that link the observed signal to meaningful parameters. As an example consider microstructure estimation in Chapter 8, which extracts a tissue parcellation. In Chapter 9, sampling from the MDS model would yield time series far away from realistic fMRI data. But, *per se*, a perfect generative model of fMRI time series would bring us no closer to understanding cognition (Bzdok and Yeo, 2017). Understanding the prediction of fitted neural network models is, in fact, a thriving research topic (Linardatos et al., 2020).

- HBMs are *explicit* distributions: they yield a density over a given signal $\mathbf{X}$. As such, HBMs are amenable to applications such as outlier detection (Hodge and Austin, 2004). What's more, explicit distributions can be trained using a variety of effective methods, including the f-KL and the r-KL. In contrast, implicit-distribution models (such as generative adversarial networks (GANs)) must rely on more complex strategies (e.g. adversarial training) because they cannot simply maximize their density over the observed data (Bond-Taylor et al., 2022).

- HBMs are *transparent hypotheses*. Training arbitrary models using optimization involves many —sometimes subtle— regularizations. In contrast, a HBM encapsulates *all* the hypotheses made by the experimenter. An HBM defines a unique theoretical posterior, that is then approximated via inference. Due to their graphical structure, HBMs notably make transparent assumptions about the dependencies between parameters. We argue that transparency is an interesting property in a field where reproducibility has been pointed as a collective challenge (Botvinik-Nezer et al., 2020). As a counter-point, one

could argue that the burden of reproducibility in Bayesian modeling is moved onto the inference method itself. We see this as further motivation for AVI.

Note on model misspecification: The issue with model misspecification goes beyond the ability of HBMs as generative models. When applied over some signal $\mathbf{X}$ that is *not* distributed according to $P(X)$ —that is, most of the time— inference is not well mathematically defined (Box, 1976). In extreme cases, model misspecification can yield non-sensical results. As an illustration, consider Figure 3.4 (right). Fitting a bimodal target distribution using a monomodal parametric approximation, the parametric distribution's mode falls into a low-density region of the target distribution! Beyond cases of extreme misspecification, major sources of unmodeled noise can also yield degenerate results. The fact that HBMs are parametric models thus has implications in terms of numerical stability.

Beyond those general considerations, we discuss below two connections between HBMs and non-parametric modeling.

Inference, data geometry and encodings Plate amortization, defined in Section 6.1.2, leverages the conditional i.i.d structure of a HBM. Taking a step back, plate amortization can be viewed as the amortization of density approximators across different sub-structures of a problem. This general concept could have applications in other highly-structured problem classes such as graphs or sequences (Z. Wu et al., 2020; Salehinejad et al., 2018).

To perform this amortization, our contribution PAVI adjoins an encoding structure to the HBM (Rouillard, Bris, et al., 2023). The encodings $\mathbf{E}_{i,n}$ embed all the "individualized" information in the problem, while the shared conditional density estimators $\mathcal{F}_i$ translate those encodings into distributions. We argue there is potential in externalizing part of the inference problem into an embedding. For instance, embeddings could be learned separately from density estimators, opening up the possibility for contrastive and transfer learning (Le-Khac et al., 2020; Jaiswal et al., 2020; Weiss et al., 2016). Or the encodings could integrate some known symmetry of the problem, such as convolution-based encodings in the case of random fields (Bishop, 2006).

The derivation of encodings also means that a "companion" non-parametric encoder could be systematically adjoined to the HBM. This draws a parallel between PAVI-E and the encoder/decoder architectures found in generative models such as variational auto encoders (VAEs) (Bond-Taylor et al., 2022).

Meta-learning, model-free approaches and amortization Supervised learning can be interpreted as the mapping from a given context set $\mathcal{C} = \{(x, y)\}$ to a predictive function f such that $f(x) = y$ (Bishop, 2006). Meta-learning —or "learning to learn"— instead recovers this mapping $\mathcal{C} \mapsto f$ in the general case (Ravi and Beatson, 2019; Iakovleva et al., 2020; Yao et al., 2019). Once the meta-training is complete, a predictive function f conditioned by an unseen context $\mathcal{C}$ can be obtained in a single forward pass —without any training done on $\mathcal{C}$. As an instance of meta-learning, the neural process family (NPF) encodes the context $\mathcal{C}$ via a deep set encoder (Garnelo et al., 2018; Dubois et al., 2020; Zaheer et al., 2018). The encoded context, along with the data point x is then used to condition an estimator for the density $q(y|x, \mathcal{C})$.

Taking a step back, NPFs are instances of fully implicit, or "model-free" Bayesian methods. A model P between the signal X and a covariate y is never explicitly defined, and sometimes, the model is actually *marginalized* (Müller et al., 2021; Hollmann et al., 2022). As an example, through prior-fitted networks, Müller et al. (2021) target the posterior predictive distribution:

$$\texttt{Pr}(y|x, \mathcal{C}) \propto \int_P P(y|x, \mathcal{C})\texttt{Pr}(P|\mathcal{C})dP \tag{10.3}$$

where `Pr` denotes a probability. Applied in the context of tabular data, the posterior predictive distribution of the covariate y given the feature x marginalizes all the possible causal structures (i.e. the models) that could have generated the context $\mathcal{C} = \{(x^{\text{context}}, y^{\text{context}})\}$ (Hollmann et al., 2022). Such "model-free" Bayesian methods have two defining characteristics:

1. They are instances of meta-learning, over all the possible contexts $\mathcal{C}$. Drawing a parallel with inference, this is similar to sample-amortized inference as defined in Section 3.3.2

2. They rely on the f-KL. In the case of prior-fitted networks: the KL from a variational approximation to the true posterior predictive distribution in Equation (10.3) (Müller et al., 2021). In practice, those methods are trained using input-output pairs $((x, \mathcal{C}), y)$ from a meta distribution (the distribution of models), without explicitly evaluating the density of the models P. Incidentally, this moves the burden of statistical specification from the distribution P to the meta-distribution of models.

This analogy —connecting Bayesian "model-free" methods to SBI— allows us to consider fully-non-parametric Bayes under a critical lens. Amortization complicates inference, so we can question the ability of model-free methods to approximate well the true posterior predictive distribution, especially in high-dimensional contexts.

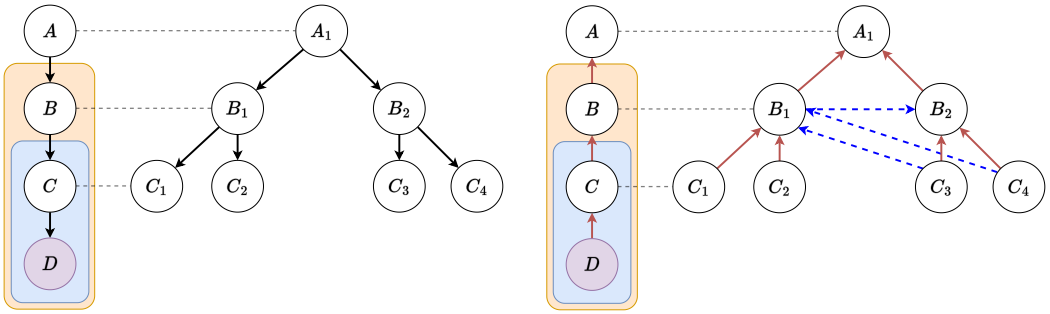

Fig. 10.1.: Faithful graph inversion induces horizontal dependencies (Webb et al., 2018) On the left, we illustrate a graph template and the corresponding parameter ground graph. On the right, we illustrate the faithful inversion of the graph template, and of the ground graph. Red arrows indicate the dependencies covered by a "naive" template inversion. The blue dashed arrows indicate the required dependencies that would be missed by template inversion.

Following up on that parallel, postulating a *specific* model regularizes the solution space and is necessary to work in the non-amortized context, which appears more suited for large-scale problems. Said differently, to our knowledge: "One cannot be both non-parametric and non-amortized".

10.4 Towards AVI

As introduced in Section 3.4, this thesis investigates the applicability of AVI to hierarchical, high-dimensional problems. To what extent do the methods and applications pursued in this thesis contribute to that debate?

Approximation gap, SVI and conditional dependencies In Section 5.2.1, we introduced *structured* VI, which combines different approximators through a causal structure. The expressivity of structured variational families depends on two factors:

1. The expressivity of the density estimators on each "node" of the graph. How closely can each estimator approximate the true (conditional) posterior of the associated ground RV?

2. The causal structure that links those estimators. How closely do the variational dependencies match the true causal structure arising from inference? In general, this structure is *not* the same as the original model's structure, due to effects such as colliders (Ambrogioni, Silvestri, et al., 2021; Webb et al., 2018).

This thesis primarily tackled the first of those two items. We showed that it was possible to combine expressive NFs into a variational family without compromising its scalability (Rouillard and Wassermann, 2022; Rouillard, Bris, et al., 2023). Critically, if such expressive "nodes" were to be combined following the *true* causal structure of the posterior, **the resulting variational family's approximation gap would be virtually null**. Combined with efficient optimization, this would make VI a fast, scalable and asymptotically *unbiased* inference method.

How feasible would such a scheme be in the context of large-scale inference?

The problem is that arbitrary dependencies are not compatible with SVI. In the case of a single plate —the 2-level case— Agrawal and Domke (2021) show that modeling only the *forward* dependencies (the same as the model's) does not reduce expressivity compared to the modeling of the full dependencies. Yet this result does not hold in the n-level case: Webb et al. (2018) show that faithful model inversion features conditional dependencies between ground RVs *of the same template* —as illustrated in Figure 10.1. We can dub those dependencies as *horizontal* dependencies —across RVs in the same plate. In practice, this means that even faithfully inverting the graph template is *not* sufficient to faithfully inverse the ground graph.

In practice, horizontal dependencies are difficult to inject into SVI. In the PAVI design, the use of a common density estimator across the ground RVs of the same template (Section 6.1.2) and the stochastic training over batches of those RVs (Section 6.2.1) prevent the modeling of *horizontal* dependencies. Put differently, the fact that we consider the inference over different ground RVs as conditionally independent inference problems is central to the PAVI design and adverse to the modeling of *horizontal* dependencies. This opens up promising research directions: *how could arbitrary conditional dependencies be modeled in the variational posterior in the context of stochastic training?*

Expressivity is not sufficient Though a necessary feature, even the unbiased variational family hypothesized above would not be sufficient for reliable inference. As already presented in Section 10.1, VI also has a lot to do with the way the variational family is trained. To this end, the r-KL is not enough. The design of structured variational families must thus be *complemented by the development of robust training methods*.

Our applications also underline the importance of *initialization* in VI. In Section 7.3.3, we show advantages in warm-starting the weights of the variational family. In doing so, we significantly improve downstream task performance, more so than

by designing informative Bayesian priors. This suggests that r-KL inference is a complex optimization problem, with potentially multiple strong local minima. Similarly, in Section 8.2.3 we describe a principled but complex and multi-layered initialization scheme. We initialize hierarchical posteriors using their independent-case counterparts. This strong initialization scheme was necessary for inference to succeed in practice.

Initialization and training methods can be seen as two sides of the same coin. Roughly, a perfect optimization method would converge to the "true" loss minimum, no matter the initialization (assuming a rather convex loss, which is hardly the case in VI). In our opinion, however, it may be overly optimistic to ignore initialization in AVI. Automated schemes could systematically compute signal summary statistics associated with canonical distributions (e.g. the mean and covariance for a Gaussian distribution). Those summary statistics could in turn be used to initialize variational approximators over ML points. Systematic initialization could both help robustify inference —multiple inferences would not risk falling into different minima— while also speeding up inference.

Thesis conclusion

This thesis explored the usage of VI to approach large-scale, hierarchical Neuroimaging problems. To this end, we leveraged NFs, structured and stochastic VI into novel variational families. In doing so, we worked towards making AVI an unbiased and scalable inference method.

We applied those techniques to a variety of Neuroscience problems, each time tackling concrete issues in the community:

1. obtaining more stable individual parcellations through transfer learning;
2. reducing uncertainty in microstructure estimation through hierarchical modeling;
3. scaling up directional coupling estimation to the full brain, while ensuring its reliability.

Through these applications, we shed light on the perks and limitations of large-scale VI. In particular, we underlined the interplay between the reliable f-KL and the scalable r-KL as complementary tools for inference.

We underline the potential of AVI to accelerate the experimenter's research cycle: from hypothesis-making through HBMs down to inferring parameters explaining the observed data.

We believe in the capabilities of those techniques to foster model-driven, interpretable and computationally efficient Neuroscience.

Acronyms

ADAVI automatic dual amortized variational inference 75

ADVI automatic differentiation variational inference 40

ANOVA analysis of variance 22

API application programming interface 43

AUC area under the curve 156

AVI automatic variational inference ix

BOLD blood-oxygen-level dependent x

CNN convolutional neural network 57

DAG directed acyclic graph 18

DCM dynamic causal modeling 140

DiFuMo dictionary of functional modes 110

dMRI diffusion magnetic resonance imaging 3

DNN deep neural network 26

HRF hemodynamic response function 10

ICA independent component analysis 25

ICC intraclass correlation coefficient 160

i.i.d independent and identically distributed 20

KL Kullback Leibler divergence 35

MAF masked autoregressive flow viii

MAP maximum a posteriori 23

MCMC Markov chain Monte Carlo 30

MDS multivariate dynamical system x

MDSI MDS identification 157

MDSI-h-VB MDSI using hybrid variational Bayes x

MF mean-field dependency scheme 66

ML maximum likelihood 22

MLP multi-layer perceptron 56

MMSE minimum mean square error 176

Summary of publications (published and in preparation)

Related to Chapter 6:

- *ADAVI: Automatic Dual Amortized Variational Inference Applied To Pyramidal Bayesian Models* (Rouillard and Wassermann, 2022) accepted at ICLR 2022
- *PAVI: Plate-Amortized Variational Inference* (Rouillard, Bris, et al., 2023) accepted at TMLR

Related to Chapter 7:

- *Improving Behavioral Prediction From Individual Functional Connectivity Through Transfer Learning* (Alexandre Le Bris, Louis Rouillard, Demian Wassermann) in preparation

Related to Chapter 8:

- *Hierarchical-uGUIDE: fast and robust Bayesian hierarchical modeling using deep learning simulation-based inference* (Rouillard, Wassermann, et al., 2024) accepted (Oral) at ISMRM 2024
- (a journal publication with an extension to multi-subject, multi-session hierarchical modeling is in preparation)

Related to Chapter 9:

- *Brain-wide modeling of causal circuit dynamics in human working memory* (Byeongwook Lee*, Louis Rouillard*, Luca Ambrogioni, Srikanth Ryali, Nicholas Branigan, Percy Mistry, Demian Wassermann', Vinod Menon') in preparation
- *Reliable estimation of effective coupling from time-shifted signals using hybrid variational Bayes* (Louis Rouillard, Luca Ambrogioni, Byeongwook Lee, Srikanth Ryali, Nicholas Branigan, Percy Mistry, Vinod Menon, Demian Wassermann) in preparation
- (a journal publication on the HCP resting-state data analysis is in preparation)

Bibliography

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, et al. (2014). "Machine learning for neuroimaging with scikit-learn". In: *Frontiers in Neuroinformatics* 8 (cit. on p. 110).

Agrawal, Abhinav and Justin Domke (2021). "Amortized Variational Inference for Simple Hierarchical Models". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., pp. 21388–21399 (cit. on pp. 86, 183).

Aitchison, Laurence (2019). "Tensor Monte Carlo: Particle Methods for the GPU era". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, et al. Vol. 32. Curran Associates, Inc. (cit. on p. 43).

Alexander, Daniel C, Tim B Dyrby, Markus Nilsson, and Hui Zhang (2019). "Imaging brain microstructure with diffusion MRI: practicality and applications". In: *NMR in Biomedicine* 32.4, e3841 (cit. on pp. 12, 123, 137).

Ali, S. M. and S. D. Silvey (1966). "A General Class of Coefficients of Divergence of One Distribution from Another". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 28.1. Publisher: [Royal Statistical Society, Wiley], pp. 131–142 (cit. on p. 38).

Ambrogioni, Luca, Umut Güçlü, Julia Berezutskaya, et al. (16–18 Apr 2019). "Forward Amortized Inference for Likelihood-Free Variational Marginalization". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 777–786 (cit. on p. 146).

Ambrogioni, Luca, Umut Güçlü, Yağmur Güçlütürk, et al. (2018). "Wasserstein Variational Inference". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, et al. Vol. 31. Curran Associates, Inc. (cit. on p. 38).

Ambrogioni, Luca, Kate Lin, Emily Fertig, et al. (13–15 Apr 2021). "Automatic structured variational inference". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 676–684 (cit. on pp. 5, 6, 48, 66, 68, 86).

Ambrogioni, Luca, Gianluigi Silvestri, and Marcel van Gerven (18–24 Jul 2021). "Automatic variational inference with cascading flows". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 254–263 (cit. on pp. 6, 33, 48, 66, 68, 78, 86, 96, 99, 102, 103, 182).

Amunts, Katrin, Hartmut Mohlberg, Sebastian Bludau, and Karl Zilles (2020). "Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture". In: *Science* 369.6506, pp. 988–992. eprint: https://www.science.org/doi/pdf/10.1126/science.abb4588 (cit. on p. 9).

Andrieu, Christophe, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan (Jan. 2003). "An Introduction to MCMC for Machine Learning". en. In: *Machine Learning* 50.1, pp. 5–43 (cit. on pp. 30, 174).

Arab, Fahimeh, AmirEmad Ghassami, Hamidreza Jamalabadai, Megan AK Peters, and Erfan Nozari (2023). "Whole-Brain Causal Discovery Using fMRI". In: *bioRxiv*, pp. 2023–08 (cit. on p. 141).

Ardalan, Zaniar and Vignesh Subbian (Feb. 2022). "Transfer Learning Approaches for Neuroimaging Analysis: A Scoping Review". In: *Frontiers in Artificial Intelligence* 5, p. 780405 (cit. on p. 108).

Baydin, Atilim Güneş, Lei Shao, Wahid Bhimji, et al. (2019). "Etalumis: Bringing probabilistic programming to scientific simulators at scale". In: *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pp. 1–24 (cit. on p. 58).

Beaumont, Mark A, Wenyang Zhang, and David J Balding (2002). "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4, pp. 2025–2035 (cit. on p. 58).

Bihan, Denis Le and Mami Iima (July 2015). "Diffusion Magnetic Resonance Imaging: What Water Tells Us about Biological Tissues". en. In: *PLOS Biology* 13.7. Publisher: Public Library of Science, e1002203 (cit. on pp. 3, 12).

Bijsterbosch, Janine, Stephen M Smith, and Christian Beckmann (2017). *An introduction to resting state fMRI functional connectivity*. Oxford University Press (cit. on p. 106).

Bijsterbosch, Janine Diane, Mark W Woolrich, Matthew F Glasser, et al. (2018). "The relationship between spatial configuration and functional connectivity of brain regions". In: *elife* 7, e32992 (cit. on p. 108).

Bingham, Eli, Jonathan P. Chen, Martin Jankowiak, et al. (2019). "Pyro: Deep Universal Probabilistic Programming". In: *J. Mach. Learn. Res.* 20, 28:1–28:6 (cit. on pp. 40, 43, 58).

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag (cit. on pp. 3, 17, 27, 28, 38, 39, 174–176, 180, 181, 213).

Bishop, Christopher M., Geoffrey Hinton, Christopher M. Bishop, and Geoffrey Hinton (Nov. 1995). *Neural Networks for Pattern Recognition*. Oxford, New York: Oxford University Press (cit. on pp. 40, 54, 57, 63, 64, 70).

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (Apr. 2017). "Variational Inference: A Review for Statisticians". en. In: *Journal of the American Statistical Association* 112.518. arXiv: 1601.00670, pp. 859–877 (cit. on pp. 5, 31, 36, 45, 46, 66).

Bond-Taylor, Sam, Adam Leach, Yang Long, and Chris G. Willcocks (2022). "Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11, pp. 7327–7347 (cit. on pp. 46, 179, 180).

Bottou, Léon and Olivier Bousquet (2007). "The Tradeoffs of Large Scale Learning". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc. (cit. on p. 32).

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F Camerer, et al. (2020). "Variability in the analysis of a single neuroimaging dataset by many teams". In: *Nature* 582.7810, pp. 84–88 (cit. on p. 179).

Box, George EP (1976). "Science and statistics". In: *Journal of the American Statistical Association* 71.356, pp. 791–799 (cit. on p. 180).

Brodmann, Korbinian and Margarete Brodmann (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth (cit. on p. 106).

Bzdok, Danilo and B. T. Thomas Yeo (July 2017). "Inference in the age of big data: Future perspectives on neuroscience". In: *NeuroImage* 155, pp. 549–564 (cit. on pp. 23–26, 175, 179).

Cai, Weidong, Tianwen Chen, Srikanth Ryali, et al. (2016). "Causal interactions within a frontal-cingulate-parietal network during cognitive control: Convergent evidence from a multisite–multitask investigation". In: *Cerebral cortex* 26.5, pp. 2140–2153 (cit. on pp. 157, 162).

Cai, Weidong, Srikanth Ryali, Ramkrishna Pasumarthy, Viswanath Talasila, and Vinod Menon (2021). "Dynamic causal brain circuits during working memory and their functional controllability". In: *Nature Communications* 12.1, p. 3314 (cit. on pp. 157, 158, 163, 220, 221).

Calhoun, Vince D. and Tülay Adali (2012). "Multisubject Independent Component Analysis of fMRI: A Decade of Intrinsic Networks, Default Mode, and Neurodiagnostic Discovery". en. In: *IEEE Reviews in Biomedical Engineering* 5, pp. 60–73 (cit. on pp. 114, 116, 166, 222).

Callaghan, Ross, Daniel C Alexander, Marco Palombo, and Hui Zhang (2020). "ConFiG: Contextual Fibre Growth to generate realistic axonal packing for diffusion MRI simulation". In: *Neuroimage* 220, p. 117107 (cit. on p. 12).

Chen, Ashley C, Desmond J Oathes, Catie Chang, et al. (2013). "Causal interactions between fronto-parietal central executive and default-mode networks in humans". In: *Proceedings of the National Academy of Sciences* 110.49, pp. 19944–19949 (cit. on pp. 157, 162).

Chollet, François et al. (2015). *Keras*. `https://keras.io` (cit. on p. 48).

Churchland, Patricia S and Terrence J Sejnowski (1988). "Perspectives on cognitive neuroscience". In: *Science* 242.4879, pp. 741–745 (cit. on p. 106).

Coelho, Santiago, Steven H Baete, Gregory Lemberskiy, et al. (2022). “Reproducibility of the Standard Model of diffusion in white matter on clinical MRI systems”. In: *NeuroImage* 257, p. 119290 (cit. on pp. 133, 134, 218).

Cranmer, Kyle, Johann Brehmer, and Gilles Louppe (May 2020). “The frontier of simulation-based inference”. en. In: *Proceedings of the National Academy of Sciences*, p. 201912789 (cit. on pp. 5, 53, 58–61, 102, 213).

Cremer, Chris, Xuechen Li, and David Duvenaud (May 2018). “Inference Suboptimality in Variational Autoencoders”. en. In: *arXiv:1801.03558 [cs, stat]*. arXiv: 1801.03558 (cit. on pp. 32–34, 212).

Dadi, Kamalaker, Gaël Varoquaux, Antonia Machlouzarides-Shalit, et al. (2020). “Fine-grain atlases of functional modes for fMRI analysis”. In: *NeuroImage* 221, p. 117126 (cit. on pp. 10, 106, 110).

Dao, Viet-Hung, David Gunawan, Minh-Ngoc Tran, et al. (2021). *Efficient Selection Between Hierarchical Cognitive Models: Cross-validation With Variational Bayes* (cit. on p. 40).

Deng, Jia, Wei Dong, Richard Socher, et al. (2009). “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255 (cit. on p. 108).

Deshpande, Gopikrishna, Krish Sathian, and Xiaoping Hu (2010). “Effect of hemodynamic variability on Granger causality analysis of fMRI”. In: *Neuroimage* 52.3, pp. 884–896 (cit. on p. 141).

Devonshire, Ian M, Nikos G Papadakis, Michael Port, et al. (2012). “Neurovascular coupling is brain region-dependent”. In: *Neuroimage* 59.3, pp. 1997–2006 (cit. on pp. 10, 141).

Dillon, Joshua V., Ian Langmore, Dustin Tran, et al. (Nov. 2017). “TensorFlow Distributions”. en. In: *arXiv:1711.10604 [cs, stat]*. arXiv: 1711.10604 (cit. on pp. 40, 43, 58).

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (2016). “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803* (cit. on p. 64).

Donoho, David (Jan. 2000). “High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality”. In: *AMS Math Challenges Lecture*, pp. 1–32 (cit. on pp. 31, 46).

Duane, Simon, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth (Sept. 1987). “Hybrid Monte Carlo”. In: *Physics Letters B* 195.2, pp. 216–222 (cit. on p. 31).

Dubois, Yann, Jonathan Gordon, and Andrew YK Foong (Sept. 2020). *Neural Process Family*. `http://yanndubs.github.io/Neural-Process-Family/` (cit. on p. 181).

Dugué, Nicolas and Anthony Perez (2015). “Directed Louvain: maximizing modularity in directed networks”. PhD thesis. Université d’Orléans (cit. on p. 165).

Efron, Bradley and Trevor Hastie (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*. Vol. 6. Cambridge University Press (cit. on pp. 26, 175).

Eickenberg, Michael, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion (2017). "Seeing it all: Convolutional network layers map the function of the human visual system". In: *NeuroImage* 152, pp. 184–194 (cit. on p. 26).

Eickhoff, Simon B, R Todd Constable, and BT Thomas Yeo (2018a). "Topographic organization of the cerebral cortex and brain cartography". In: *Neuroimage* 170, pp. 332–347 (cit. on p. 106).

Eickhoff, Simon B., B. T. Thomas Yeo, and Sarah Genon (Nov. 2018b). "Imaging-based parcellations of the human brain". en. In: *Nature Reviews Neuroscience* 19.11, pp. 672–686 (cit. on pp. 10, 11, 105, 106, 108, 137, 140, 163).

Fan, Lingzhong, Hai Li, Junjie Zhuo, et al. (2016). "The human brainnetome atlas: a new brain atlas based on connectional architecture". In: *Cerebral cortex* 26.8, pp. 3508–3526 (cit. on pp. 106, 157).

Fayaz, A, P Croft, RM Langford, LJ Donaldson, and GT Jones (2016). "Prevalence of chronic pain in the UK: a systematic review and meta-analysis of population studies". In: *BMJ open* 6.6, e010364 (cit. on p. 22).

Foreman-Mackey, Daniel, David W. Hogg, Dustin Lang, and Jonathan Goodman (Mar. 2013). "emcee: The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925. arXiv:1202.3665 [astro-ph, physics:physics, stat], pp. 306–312 (cit. on p. 44).

Frässle, Stefan, Ekaterina I. Lomakina, Adeel Razi, et al. (2017). "Regression DCM for fMRI". In: *NeuroImage* 155, pp. 406–421 (cit. on pp. 5, 141, 153–155, 167, 220).

Frässle, Stefan and Klaas E. Stephan (Feb. 2022). "Test-retest reliability of regression dynamic causal modeling". en. In: *Network Neuroscience* 6.1, pp. 135–160 (cit. on pp. 153, 155, 160).

Friston, Karl, Thomas Parr, and Peter Zeidman (2018). "Bayesian model reduction". In: *arXiv preprint arXiv:1805.07092* (cit. on p. 178).

Friston, Karl J (2009). "Modalities, modes, and models in functional neuroimaging". In: *Science* 326.5951, pp. 399–403 (cit. on pp. 139, 155, 177).

Friston, Karl J, Lee Harrison, and Will Penny (2003). "Dynamic causal modelling". In: *Neuroimage* 19.4, pp. 1273–1302 (cit. on pp. 12, 140).

Gao, Yufei, Yameng Zhang, Hailing Wang, Xiaojuan Guo, and Jiacai Zhang (2019). "Decoding Behavior Tasks From Brain Activity Using Deep Transfer Learning". In: *IEEE Access* 7, pp. 43222–43232 (cit. on p. 109).

Garnelo, Marta, Jonathan Schwarz, Dan Rosenbaum, et al. (2018). *Neural Processes* (cit. on p. 181).

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2004). *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC (cit. on pp. 3, 17, 27, 36, 96, 177).

Gelman, Andrew and Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press (cit. on p. 22).

Ginsburger, Kévin, Felix Matuschke, Fabrice Poupon, et al. (2019). “MEDUSA: A GPU-based tool to create realistic phantoms of the brain microstructure using tiny spheres”. In: *NeuroImage* 193, pp. 10–24 (cit. on p. 12).

Giordano, Ryan J, Tamara Broderick, and Michael I Jordan (2015). “Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. (cit. on p. 46).

Glöckler, Manuel, Michael Deistler, and Jakob H. Macke (2022). “Variational methods for simulation-based inference”. In: *International Conference on Learning Representations* (cit. on pp. 5, 128, 137).

Glover, Gary H (1999). “Deconvolution of impulse response in event-related BOLD fMRI1”. In: *Neuroimage* 9.4, pp. 416–429 (cit. on p. 143).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press (cit. on pp. 25, 26).

Gordon, Evan M., Timothy O. Laumann, Adrian W. Gilmore, et al. (2017). “Precision Functional Mapping of Individual Human Brains”. In: *Neuron* 95.4, 791–807.e7 (cit. on pp. 114, 116, 166, 222).

Grathwohl, Will, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud (2018). “Backpropagation through the Void: Optimizing control variates for black-box gradient estimation”. In: *International Conference on Learning Representations* (cit. on pp. 43, 112).

Greenberg, David, Marcel Nonnenmacher, and Jakob Macke (2019). “Automatic posterior transformation for likelihood-free inference”. In: *International Conference on Machine Learning*. PMLR, pp. 2404–2414 (cit. on pp. 59, 60, 102, 174).

Grenioux, Louis, Alain Durmus, Éric Moulines, and Marylou Gabrié (2023). “On Sampling with Approximate Transport Maps”. In: *arXiv preprint arXiv:2302.04763* (cit. on p. 174).

Handwerker, Daniel A, John M Ollinger, and Mark D’Esposito (2004). “Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses”. In: *Neuroimage* 21.4, pp. 1639–1651 (cit. on pp. 10, 141, 167).

Hodge, Victoria and Jim Austin (2004). “A survey of outlier detection methodologies”. In: *Artificial intelligence review* 22, pp. 85–126 (cit. on p. 179).

Hoffman, Matthew D and David M Blei (2015). “Structured stochastic variational inference”. In: *Artificial Intelligence and Statistics*, pp. 361–369 (cit. on pp. 5, 66, 70).

Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley (2013). “Stochastic Variational Inference”. In: *Journal of Machine Learning Research* 14.40, pp. 1303–1347 (cit. on pp. 5, 68–70, 83, 84, 97).

Hoffman, Matthew D., Pavel Sountsov, Joshua V. Dillon, et al. (2019). “NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport”. In: *arXiv: Computation* (cit. on pp. 31, 174).

Hollmann, Noah, Samuel Müller, Katharina Eggensperger, and Frank Hutter (2022). "Tabpfn: A transformer that solves small tabular classification problems in a second". In: *arXiv preprint arXiv:2207.01848* (cit. on p. 181).

Howard, Amy FD, Istvan N Huszar, Adele Smart, et al. (2022). "The BigMac dataset: an open resource combining multi-contrast MRI and microscopy in the macaque brain". In: *bioRxiv*. eprint: `https://www.biorxiv.org/content/early/2022/09/10/2022.09.08.506363.full.pdf` (cit. on p. 9).

Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren, eds. (2019). *Automated Machine Learning: Methods, Systems, Challenges*. en. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing (cit. on p. 48).

Iakovleva, Ekaterina, Jakob Verbeek, and Karteek Alahari (13–18 Jul 2020). "Meta-Learning with Shared Amortized Variational Inference". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4572–4582 (cit. on p. 181).

Ianuş, Andrada, Noam Shemesh, Daniel C Alexander, and Ivana Drobnjak (2017). "Double oscillating diffusion encoding and sensitivity to microscopic anisotropy". In: *Magnetic resonance in medicine* 78.2, pp. 550–564 (cit. on pp. 58, 125).

Jaiswal, Ashish, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon (2020). "A survey on contrastive self-supervised learning". In: *Technologies* 9.1, p. 2 (cit. on p. 180).

Jallais, Maëliss and Marco Palombo (2023). "uGUIDE: a framework for microstructure imaging via generalized uncertainty-driven inference using deep learning". In: *arXiv preprint arXiv:2312.17293* (cit. on pp. 5, 14, 34, 126, 129, 217).

Jallais, Maëliss, Pedro Luiz Coelho Rodrigues, Alexandre Gramfort, and Demian Wassermann (2021). "Inverting brain grey matter models with likelihood-free inference: a tool for trustable cytoarchitecture measurements". In: *arXiv preprint arXiv:2111.08693* (cit. on p. 14).

Jang, Hojin, Sergey M Plis, Vince D Calhoun, and Jong-Hwan Lee (2017). "Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks". In: *NeuroImage* 145, pp. 314–328 (cit. on p. 26).

Jasra, Ajay, Chris C Holmes, and David A Stephens (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling". In: (cit. on pp. 65, 102).

Jelescu, Ileana O, Marco Palombo, Francesca Bagnato, and Kurt G Schilling (2020). "Challenges for biophysical modeling of microstructure". In: *Journal of Neuroscience Methods* 344, p. 108861 (cit. on pp. 4, 12, 14, 123, 137, 211).

Jelescu, Ileana O, Alexandre de Skowronski, Françoise Geffroy, Marco Palombo, and Dmitry S Novikov (2022). "Neurite Exchange Imaging (NEXI): A minimal model of diffusion in gray matter with inter-compartment water exchange". In: *NeuroImage* 256, p. 119277 (cit. on p. 123).

Jospin, Laurent Valentin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun (2022). "Hands-on Bayesian neural networks—A tutorial for deep learning users". In: *IEEE Computational Intelligence Magazine* 17.2, pp. 29–48 (cit. on p. 176).

Justin, Alsing, Charnock Tom, Feeney Stephen, and Wandelt Benjamin (2019). "Fast likelihood-free cosmology with neural density estimators and active learning". In: *Mon. Not. Roy. Astron. Soc* 488, pp. 4440–4458 (cit. on p. 58).

Le-Khac, Phuc H, Graham Healy, and Alan F Smeaton (2020). "Contrastive representation learning: A framework and review". In: *Ieee Access* 8, pp. 193907–193934 (cit. on p. 180).

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun (cit. on p. 40).

Kingma, Durk P and Prafulla Dhariwal (2018). "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31 (cit. on pp. 54, 62, 64).

Koller, Daphne and Nir Friedman (2009). *Probabilistic graphical models: principles and techniques*. en. Adaptive computation and machine learning. Cambridge, MA: MIT Press (cit. on pp. 3, 17, 18, 20, 29, 76).

Kong, Ru, Jingwei Li, Csaba Orban, et al. (2019). "Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion." In: *Cerebral cortex* 29 6, pp. 2533–2551 (cit. on pp. 3–5, 40, 106, 108, 110, 114, 116, 120, 166, 211, 222, 223).

Kong, Ru, Qing Yang, Evan Gordon, et al. (2021). "Individual-specific areal-level parcellations improve functional connectivity prediction of behavior". In: *Cerebral Cortex* 31.10, pp. 4477–4500 (cit. on pp. 114, 116, 122, 166, 222).

Krishnan, Rahul, Uri Shalit, and David Sontag (2017). "Structured inference networks for nonlinear state space models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1 (cit. on p. 119).

Kruschke, John K. and Torrin M. Liddell (Feb. 2018). "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". en. In: *Psychonomic Bulletin & Review* 25.1, pp. 178–206 (cit. on pp. 23, 24, 175, 177, 178).

Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei (Mar. 2016). "Automatic Differentiation Variational Inference". en. In: *arXiv:1603.00788 [cs, stat]*. arXiv: 1603.00788 (cit. on pp. 40, 48).

Le Bihan, D and E Breton (1985). "In vivo magnetic resonance imaging of diffusion". In: *Comptes Rendus des Seances de l'Academie des Sciences. Serie 2* 301.15, pp. 1109–1112 (cit. on p. 124).

Lee, Juho, Yoonho Lee, Jungtaek Kim, et al. (Sept. 2019a). “Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3744–3753 (cit. on p. 86).

– (Sept. 2019b). “Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 3744–3753 (cit. on p. 87).

Li, Yingzhen and Richard E Turner (2016). “Rényi Divergence Variational Inference”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. (cit. on p. 38).

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). “Explainable ai: A review of machine learning interpretability methods”. In: *Entropy* 23.1, p. 18 (cit. on p. 179).

Liu, Runjing, Jeffrey Regier, Nilesh Tripuraneni, Michael Jordan, and Jon Mcauliffe (Sept. 2019). “Rao-Blackwellized Stochastic Gradients for Discrete Distributions”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 4023–4031 (cit. on p. 42).

Lueckmann, Jan-Matthis, Pedro J Goncalves, Giacomo Bassetto, et al. (2017). “Flexible statistical inference for mechanistic models of neural dynamics”. In: *Advances in neural information processing systems* 30 (cit. on p. 59).

Mandonnet, Emmanuel (2011). “Intraoperative electrical mapping: advances, limitations and perspectives”. In: *Brain Mapping: From Neural Basis of Cognition to Surgical Applications*. Ed. by Hugues Duffau. Vienna: Springer Vienna, pp. 101–108 (cit. on pp. 11, 109).

Martín Abadi, Ashish Agarwal, Paul Barham, et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org (cit. on p. 43).

McGlothlin, Anna E. and Kert Viele (Dec. 2018). “Bayesian Hierarchical Models”. In: *JAMA* 320.22, pp. 2365–2366. eprint: `https://jamanetwork.com/journals/jama/articlepdf/2718053/jama\_mcglothlin\_2018\_gm\_180005.pdf` (cit. on p. 22).

Modi, Chirag, Charles Margossian, Yuling Yao, et al. (July 2023). *Variational Inference with Gaussian Score Matching*. arXiv:2307.07849 [cs, stat] (cit. on p. 38).

Mohamed, Shakir and Balaji Lakshminarayanan (2016). “Learning in implicit generative models”. In: *arXiv preprint arXiv:1610.03483* (cit. on p. 59).

Moral, Pierre Del, Arnaud Doucet, and Ajay Jasra (July 2007). “Sequential Monte Carlo for Bayesian Computation”. In: *Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting June 2–6, 2006*. Ed. by J M Bernardo, M J Bayarri, J O Berger, et al. Oxford University Press, p. 0 (cit. on p. 31).

Müller, Samuel, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter (2021). “Transformers can do bayesian inference”. In: *arXiv preprint arXiv:2112.10510* (cit. on p. 181).

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press (cit. on p. 23).

Novikov, Dmitry S., Els Fieremans, Sune N. Jespersen, and Valerij G. Kiselev (2019). “Quantifying brain microstructure with diffusion MRI: Theory and parameter estimation”. en. In: *NMR in Biomedicine* 32.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nbm.3998, e3998 (cit. on pp. 3, 4, 12–14, 124, 125, 137, 211, 217).

Obermeyer, Fritz, Eli Bingham, Martin Jankowiak, et al. (Sept. 2019). “Tensor Variable Elimination for Plated Factor Graphs”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 4871–4880 (cit. on p. 43).

Ogawa, Seiji and Tso-Ming Lee (1990). “Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulation”. In: *Magnetic resonance in medicine* 16.1, pp. 9–18 (cit. on p. 10).

Orton, Matthew R, David J Collins, Dow-Mu Koh, and Martin O Leach (2014). “Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling”. In: *Magnetic resonance in medicine* 71.1, pp. 411–420 (cit. on p. 127).

Palombo, Marco, Andrada Ianus, Michele Guerreri, et al. (July 2020). “SANDI: A compartment-based model for non-invasive apparent soma and neurite imaging by diffusion MRI”. In: *NeuroImage* 215, p. 116835 (cit. on pp. 12–14, 124, 137, 211).

Papamakarios, George and Iain Murray (2016). “Fast \epsilon -free Inference of Simulation Models with Bayesian Conditional Density Estimation”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. (cit. on pp. 5, 59, 60, 126, 129, 174, 217).

Papamakarios, George, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan (Dec. 2019). “Normalizing Flows for Probabilistic Modeling and Inference”. en. In: *arXiv:1912.02762 [cs, stat]*. arXiv: 1912.02762 (cit. on pp. 5, 42, 53, 55, 64, 213).

Papamakarios, George, Theo Pavlakou, and Iain Murray (2017). “Masked autoregressive flow for density estimation”. In: *Advances in neural information processing systems* 30 (cit. on pp. 55, 126).

Papamakarios, George, David Sterratt, and Iain Murray (2019). “Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 837–848 (cit. on pp. 59, 60).

Paszke, Adam, Sam Gross, Francisco Massa, et al. (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 8024–8035 (cit. on p. 43).

Plis, Sergey M, Devon R Hjelm, Ruslan Salakhutdinov, et al. (2014). “Deep learning for neuroimaging: a validation study”. In: *Frontiers in neuroscience* 8, p. 229 (cit. on p. 26).

Poldrack, Russell A, Jeanette A Mumford, and Thomas E Nichols (2011). *Handbook of functional MRI data analysis*. Cambridge University Press (cit. on pp. 3, 10, 22, 106, 139).

Powell, Elizabeth, Matteo Battocchio, Christopher S Parker, and Paddy J Slator (2021). "Generalised Hierarchical Bayesian Microstructure Modelling for Diffusion MRI". In: *Computational Diffusion MRI: 12th International Workshop, CDMRI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 12*. Springer, pp. 36–47 (cit. on pp. 5, 14, 127, 137).

Power, Jonathan D, Alexander L Cohen, Steven M Nelson, et al. (2011). "Functional network organization of the human brain". In: *Neuron* 72.4, pp. 665–678 (cit. on p. 106).

Ranganath, Rajesh, Sean Gerrish, and David M. Blei (Dec. 2013). "Black Box Variational Inference". en. In: *arXiv:1401.0118 [cs, stat]*. arXiv: 1401.0118 (cit. on pp. 38, 40).

Rangaprakash, D, Robert L Barry, and Gopikrishna Deshpande (2023). "The confound of hemodynamic response function variability in human resting-state functional MRI studies". In: *Frontiers in Neuroscience* 17 (cit. on pp. 141, 167).

Rangaprakash, D, Guo-Rong Wu, Daniele Marinazzo, Xiaoping Hu, and Gopikrishna Deshpande (2018). "Hemodynamic response function (HRF) variability confounds resting-state fMRI functional connectivity". In: *Magnetic resonance in medicine* 80.4, pp. 1697–1713 (cit. on pp. 141, 167).

Ravi, Sachin and Alex Beatson (2019). "Amortized Bayesian Meta-Learning". In: *International Conference on Learning Representations* (cit. on p. 181).

Rezende, Danilo and Shakir Mohamed (2015). "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR, pp. 1530–1538 (cit. on pp. 53, 102).

Roebroeck, Alard, Elia Formisano, and Rainer Goebel (2005). "Mapping directed influence over the brain using Granger causality and fMRI". In: *Neuroimage* 25.1, pp. 230–242 (cit. on p. 140).

Rosen-Zvi, Michal, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers (2010). "Learning author-topic models from text corpora". In: *ACM Transactions on Information Systems (TOIS)* 28.1, pp. 1–38 (cit. on p. 25).

Rouillard, Louis, Alexandre Le Bris, Thomas Moreau, and Demian Wassermann (2023). "PAVI: Plate-Amortized Variational Inference". In: *Transactions on Machine Learning Research*. Reproducibility Certification (cit. on pp. 5, 75, 78, 81, 102, 148, 166, 180, 183, 195, 222).

Rouillard, Louis and Demian Wassermann (2022). "ADAVI: Automatic Dual Amortized Variational Inference Applied To Pyramidal Bayesian Models". In: *International Conference on Learning Representations* (cit. on pp. 5, 75, 78, 86, 87, 102, 103, 183, 195).

Rouillard, Louis, Demian Wassermann, Marco Palombo, and Maëliss Jallais (May 2024). "Hierarchical-uGUIDE: fast and robust Bayesian hierarchical modelling using deep learning simulation-based inference". In: (cit. on pp. 128, 137, 195).

Rubin, Donald B (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician". In: *The Annals of Statistics*, pp. 1151–1172 (cit. on p. 58).

Ryali, Srikanth, Kaustubh Supekar, Tianwen Chen, and Vinod Menon (Jan. 2011). "Multivariate dynamical systems models for estimating causal interactions in fMRI". en. In: *NeuroImage* 54.2, pp. 807–823 (cit. on pp. 4, 5, 10, 12, 142, 211).

Salehinejad, Hojjat, Julianne Baarbe, Sharan Sankar, et al. (2018). "Recent Advances in Recurrent Neural Networks". In: *CoRR* abs/1801.01078. arXiv: 1801.01078 (cit. on p. 180).

Sanchez-Romero, R, JD Ramsey, K Zhang, et al. (2018). *Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. Network Neuroscience, 3 (2), 274–306* (cit. on pp. 155, 156).

Schaefer, Alexander, Ru Kong, Evan M Gordon, et al. (2018). "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI". In: *Cerebral cortex* 28.9, pp. 3095–3114 (cit. on p. 106).

Stangor, Charles and Jennifer Walinga (2014). *Introduction to psychology-1st Canadian edition* (cit. on pp. 3, 13, 108, 211).

Steffener, Jason, Matthias Tabert, Aaron Reuben, and Yaakov Stern (2010). "Investigating hemodynamic response variability at the group level using basis functions". In: *Neuroimage* 49.3, pp. 2113–2122 (cit. on p. 143).

Suder, Piotr M., Jason Xu, and David B. Dunson (2023). "Bayesian Transfer Learning". In: (cit. on p. 109).

Sudlow, Cathie, John Gallacher, Naomi Allen, et al. (Mar. 2015). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". eng. In: *PLoS medicine* 12.3, e1001779 (cit. on pp. 15, 22).

Taylor, Amanda J., Jung Hwan Kim, and David Ress (June 2018). "Characterization of the hemodynamic response function across the majority of human cerebral cortex". eng. In: *NeuroImage* 173, pp. 322–331 (cit. on pp. 10, 141, 143, 150).

Tejero-Cantero, Alvaro, Jan Boelts, Michael Deistler, et al. (2020). "sbi: A toolkit for simulation-based inference". In: *Journal of Open Source Software* 5.52, p. 2505 (cit. on pp. 58, 60).

Thomas Yeo, B. T., Fenna M. Krienen, Jorge Sepulcre, et al. (2011). "The organization of the human cerebral cortex estimated by intrinsic functional connectivity". In: *Journal of Neurophysiology* 106.3. PMID: 21653723, pp. 1125–1165. eprint: https://doi.org/10.1152/jn.00338.2011 (cit. on pp. 3, 38, 106–108, 115, 119–121, 162–165, 215, 216, 221).

Tietz, Marian, Thomas J. Fan, Daniel Nouri, Benjamin Bossan, and skorch Developers (July 2017). *skorch: A scikit-learn compatible neural network library that wraps PyTorch* (cit. on p. 48).

Titsias, Michalis K. and Francisco Ruiz (16–18 Apr 2019). "Unbiased Implicit Variational Inference". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 167–176 (cit. on pp. 33, 99, 103).

Tononi, Giulio, Olaf Sporns, and Gerald M Edelman (1994). “A measure for brain complexity: relating functional segregation and integration in the nervous system.” In: *Proceedings of the National Academy of Sciences* 91.11, pp. 5033–5037 (cit. on p. 106).

Towsley, Kayle, Michael I Shevell, Lynn Dagenais, Repacq Consortium, et al. (2011). “Population-based study of neuroimaging findings in children with cerebral palsy”. In: *European journal of paediatric neurology* 15.1, pp. 29–35 (cit. on p. 22).

Tucker, George, Andriy Mnih, Christopher Maddison, and Jascha Sohl-Dickstein (Mar. 2017). “REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models”. In: (cit. on p. 42).

Valdes-Sosa, Pedro A, Alard Roebroeck, Jean Daunizeau, and Karl Friston (2011). “Effective connectivity: influence, causality and biophysical modeling”. In: *Neuroimage* 58.2, pp. 339–361 (cit. on p. 140).

Valverde, Juan Miguel, Vandad Imani, Ali Abdollahzadeh, et al. (Apr. 2021). “Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review”. In: *Journal of Imaging* 7.4, p. 66 (cit. on p. 108).

Van Den Heuvel, Martijn P and Hilleke E Hulshoff Pol (2010). “Exploring the brain network: a review on resting-state fMRI functional connectivity”. In: *European neuropsychopharmacology* 20.8, pp. 519–534 (cit. on pp. 10, 11, 105, 106, 139, 140, 144, 160, 165).

Van Essen, D. C., K. Ugurbil, E. Auerbach, et al. (Oct. 2012). “The Human Connectome Project: a data acquisition perspective”. In: *Neuroimage* 62.4, pp. 2222–2231 (cit. on pp. 15, 108–110, 112, 116, 156, 163).

Varoquaux, Gaël and R. Cameron Craddock (2013). “Learning and comparing functional connectomes across subjects”. In: *NeuroImage* 80, pp. 405–415 (cit. on pp. 140, 144, 160, 165).

Wang, Bo and D. M. Titterington (Jan. 2005). “Inadequacy of interval estimates corresponding to variational Bayesian approximations”. en. In: *International Workshop on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, pp. 373–380 (cit. on p. 46).

Wang, Danhong, Randy L Buckner, Michael D Fox, et al. (Dec. 2015). “Parcellating cortical functional networks in individuals”. en. In: *Nature Neuroscience* 18.12, pp. 1853–1860 (cit. on pp. 114, 116, 166, 222).

Wang, Dilin, Hao Liu, and Qiang Liu (2018). “Variational Inference with Tail-adaptive f-Divergence”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, et al. Vol. 31. Curran Associates, Inc. (cit. on p. 38).

Webb, Stefan, Adam Golinski, Rob Zinkov, et al. (2018). “Faithful Inversion of Generative Models for Effective Amortized Inference”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, et al. Vol. 31. Curran Associates, Inc. (cit. on pp. 68, 182, 183, 222).

Wehenkel, Antoine and Gilles Louppe (2021). “Graphical normalizing flows”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 37–45 (cit. on p. 65).

Weilbach, Christian, Boyan Beronov, Frank Wood, and William Harvey (26–28 Aug 2020). "Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 4441–4451 (cit. on pp. 33, 65).

Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang (May 2016). "A survey of transfer learning". In: *Journal of Big Data* 3.1, p. 9 (cit. on pp. 108, 109, 180).

Wu, Zonghan, Shirui Pan, Fengwen Chen, et al. (2020). "A Comprehensive Survey on Graph Neural Networks". en. In: *IEEE Transactions on Neural Networks and Learning Systems*. arXiv: 1901.00596, pp. 1–21 (cit. on p. 180).

Yao, Huaxiu, Ying Wei, Junzhou Huang, and Zhenhui Li (Sept. 2019). "Hierarchically Structured Meta-learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 7045–7054 (cit. on p. 181).

Zaheer, Manzil, Satwik Kottur, Siamak Ravanbakhsh, et al. (Apr. 2018). "Deep Sets". en. In: *arXiv:1703.06114 [cs, stat]*. arXiv: 1703.06114 (cit. on pp. 86, 87, 181).

Zanitti, Gaston E, Yamil Soto, Valentin Iovene, et al. (2022). "Scalable Query Answering under Uncertainty to Neuroscientific Ontological Knowledge: The NeuroLang Approach". In: *Neuroinformatics* (cit. on p. 48).

Zhang, Cheng, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt (Aug. 2019). "Advances in Variational Inference". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8, pp. 2008–2026 (cit. on p. 33).

Zhang, Fan, Alessandro Daducci, Yong He, et al. (2022). "Quantitative mapping of the brain's structural connectivity using diffusion MRI tractography: A review". In: *NeuroImage* 249, p. 118870 (cit. on pp. 3, 12).

Zhang, Hejia, Po Hsuan Chen, and Peter Jeffrey Ramadge (Jan. 2018). "Transfer learning on fMRI datasets". In: pp. 595–603 (cit. on p. 109).

Zhang, Hui, Torben Schneider, Claudia A Wheeler-Kingshott, and Daniel C Alexander (2012). "NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain". In: *Neuroimage* 61.4, pp. 1000–1016 (cit. on pp. 12–14, 124, 137, 211).

List of Figures

List of Tables