HAL
open science

Monotonic Gaussian Process for Spatio-Temporal Disease Progression Modeling in Brain Imaging Data

Clement Abi Nader, Nicholas Ayache, Philippe Robert, Marco Lorenzi

To cite this version:

Clement Abi Nader, Nicholas Ayache, Philippe Robert, Marco Lorenzi. Monotonic Gaussian Process for Spatio-Temporal Disease Progression Modeling in Brain Imaging Data. 2019. hal-02051843v2

HAL Id: hal-02051843

https://hal.science/hal-02051843v2

Preprint submitted on 7 Jun 2019 (v2), last revised 10 Oct 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Monotonic Gaussian Process for Spatio-Temporal Disease Progression Modeling in Brain Imaging Data

Clément Abi Nader[a,*], Nicholas Ayache[a], Philippe Robert[b], Marco Lorenzi[a], for the Alzheimer's Disease Neuroimaging Initiative[**]

[a] *Université Nice Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France.*
[b] *Université Nice Côte d'Azur, CoBTeK lab, MNC3 program*

Abstract

We introduce a probabilistic generative model for disentangling spatio-temporal disease trajectories from collections of high-dimensional brain images. The model is based on spatio-temporal matrix factorization, where inference on the sources is constrained by anatomically plausible statistical priors. To model realistic trajectories, the temporal sources are defined as monotonic and time-reparameterized Gaussian Processes. To account for the non-stationarity of brain images, we model the spatial sources as sparse codes convolved at multiple scales. The method was tested on synthetic data favourably comparing with standard blind source separation approaches. The application on large-scale imaging data from a clinical study allows to disentangle differential temporal progression patterns mapping brain regions key to neurodegeneration, while revealing a disease-specific time scale associated to the clinical diagnosis.

Keywords: Alzheimer's disease, Disease progression modeling, Gaussian Process, Bayesian modeling, Stochastic variational inference, Clinical trials

1. Introduction

Neurodegenerative disorders such as Alzheimer's disease (AD) are characterized by morphological and molecular changes of the brain, ultimately leading to cognitive and behavioral decline. Clinicians suggested hypothetical models of the disease evolution, showing how different types of biomarkers interact and lead to the final dementia stage [14]. In the past

*Corresponding author at: Epione Research Project, INRIA Sophia-Antipolis, 2004, route des Lucioles, 06902 Sophia-Antipolis, France, clement.abi-nader@inria.fr.

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

Email addresses: `clement.abi-nader@inria.fr` (Clément Abi Nader), `nicholas.ayache@inria.fr` (Nicholas Ayache), `probert@unice.fr` (Philippe Robert), `marco.lorenzi@inria.fr` (Marco Lorenzi)

years, efforts have been made in order to collect large databases of imaging and clinical measures, hoping to obtain more insights about the disease progression through data-driven models describing the trajectory of the disease over time. This kind of models are of critical importance for understanding the pathological progression in large scale data, and would represent a valuable reference for improving the individual diagnosis. Within this context, we propose a spatio-temporal generative model of disease progression, aimed at disentangling and quantifying the independent dynamics of changes observed in datasets of multi-modal data. With this term we indicate data acquired via different imaging modalities such as Magnetic Resonance Imaging (MRI) or Positron-Emission Tomography (PET), as well as non-imaging data such as clinical scores assessed by physicians. Moreover, we aim at automatically inferring the disease severity of a patient with respect to the estimated trajectory. Defining such a disease progression model raises a number of methodological challenges.

AD spreads over decades with a temporal mismatch between the onset of the disease and the moment where the clinical symptoms appear. Either age of diagnosis, or the chronological age, are therefore not suitable as a temporal reference to describe the disease progression in time. Moreover, as the follow-up of patients doesn't exceed a few years, the development of a model of long-term pathological changes requires to integrate cross-sectional data from different individuals, in order to consider a longer period of time. In virtue of the lack of a well defined temporal reference, observations from different individuals are characterized by large and unknown variability in the onset and speed of the disease. It is therefore necessary to account for a time-reparameterization function, mapping each individuals' observations to a common temporal axis associated to the absolute disease trajectory [15, 33]. This would allow to estimate an absolute time-reference related to the natural history of the pathology.

The analysis of MRI and PET data, requires to account for spatio-temporally correlated features (voxels, i.e. volumetric pixels) defined over arrays of more than a million entries. The development of inference schemes jointly considering these correlation properties thus raises scalability issues, especially when accounting for the non-stationarity of the image signal. Furthermore, the brain regions involved in AD exhibit various dynamics in time, and evolve at different speeds [35]. From a modeling perspective, accounting for differential trajectories over space and time raises the problem of source identification and separation. This issue has been widely addressed in neuroimaging via Independent Component Analysis (ICA) [8], especially on functional MRI (fMRI) data [7]. Nevertheless, while fMRI time-series are usually defined over a few hundreds of time points acquired per subject, our problem consists in jointly analyzing short-term and cross-sectional data observations with respect to an unknown time-line. This problem cannot be tackled with standard ICA, as time is generally an independent variable on which inference is not required. Moreover, ICA retrieves spatial sources based on the assumption of statistical independence. This assumption does not necessarily lead to clinically interpretable findings. Indeed, dependency across temporal patterns can be still highly relevant to the pathology, for example when modeling temporal delay across similar sources.

The problem of providing a realistic description of the biological processes is critical when analyzing biomedical data, such as medical images. For example, to describe a plausible evolution of AD from normal to pathological stages, smoothness and monotonicity are commonly assumed for the temporal sources. It is also necessary to account for the non-stationarity of changes affecting the brain from global to localized spatio-temporal processes. As a result, spatial sources need to account for different resolutions at which these changes take place. While several multi-scale analysis approaches have been proposed to model spatio-temporal signals [23, 6, 13], extending this type of methods to the high-dimension of medical images is generally not trivial due to scalability issues. Finally, the noisy nature of medical images, along with the large signal variability across observations, requires a modeling framework robust to bias and noise.

In this work, we propose to jointly address these issues within a Bayesian framework for the spatio-temporal analysis of large-scale collections of multi-modal brain data. We show that this framework allows us to naturally encode plausibility constraints through clinically-inspired priors, while accounting for the uncertainty of the temporal profiles and brain structures we wish to estimate. Similarly to the ICA setting, we formulate the problem of trajectory modeling through matrix factorization across temporal and spatial sources. This is done for each modality by inferring their specific spatio-temporal sources. To promote smoothness in time and avoid any unnecessary hypothesis on the temporal trajectories, we rely on non-parametric modeling based on Gaussian Process (GP). We account for a plausible evolution from healthy to pathological stages thanks to a monotonicity constraint applied on the GP. Moreover, individuals' observations are temporally re-aligned on a common scale via a time-warping function. In the case of imaging data, to model the non-stationarity of the spatial signal, the spatial sources are defined as sparse activation maps convolved at different scales. We show that our framework can be efficiently optimized through stochastic variational inference, allowing to exploit automatic differentiation and GPU support to speed up computations.

The paper is organized as follows: Section 2 analyzes related work on spatio-temporal modeling of neurodegeneration, while Section 3 details our method. In Section 4 we present experiments on synthetic data in which we compare our model to standard blind source separation approaches. We finally provide a demonstration of our method on the modeling of imaging data from a large scale clinical study. Prospects for future work and conclusions are drawn in section 5. Derivations that we could not fit in the paper are detailed in the Appendices.

2. Related Work in Neurodegeneration Modeling

To deal with the uncertainty of the time-line of neurodegenerative pathologies, the concept of time-reparameterization of imaging-derived features has been used in several works. The underlying principle consists in estimating an absolute time-scale of disease progression by temporally re-aligning data from different subjects. For instance, in [36] the time-evolution

was approximated as a sequence of events which need to be re-ordered for each patient. This approach thus considers the evolution of neurodegenerative diseases as a collection of transitions between discrete stages. This hypothesis is however limiting, as it doesn't reflect the continuity of changes affecting the brain along the course of the pathology.

To address this limitation, we rely on a continuous parameterization of the time-axis as in [21, 10]. In particular, individuals' observations are time-realigned on a common temporal scale via a time-warping function. Using a set of relevant scalar biomarkers, this kind of approach allows to learn a time-scale describing the pathology evolution, and to estimate a data-driven time-line markedly correlated with the decline of cognitive abilities. Similarly, in [4] a disease progression score was estimated using biomarkers from molecular imaging. These methods are however based on the analysis of low-dimensional measures, such as collections of clinical variables. Therefore, they do not allow to scale to the high dimension of multi-modal medical images. Our work tackles this shortcoming thanks to a scalable inference scheme based on stochastic variational inference.

Concerning the spatio-temporal representation of neurodegeneration, a mixed-effect model was proposed by [19] to learn an average spatio-temporal trajectory of brain evolution on cortical thickness data. The fixed-effect describes the average trajectory, while random effects are estimated through individual spatio-temporal warping functions, modeling how each subject differs from the global progression. Still, the extension of this approach to image volumes raises scalability issues. It has also to be noted that, to allow computational tractability, the brain evolution was assumed to be stationary both in space and time, thus limiting the ability of the model to disentangle the multiple dynamics of the brain structures involved in AD.

An attempt to source separation is proposed in [24], through the decomposition of cortical thickness measurements as a mixture of spatio-temporal processes. This is performed by associating to each cortical vertex a temporal progression modeled by a sigmoid function, which may be however too simplistic to describe the progression of AD temporal processes. We propose to overcome this issue by non-parametric modeling of the temporal sources through GPs. Moreover, due to the lack of an explicit spatial correlation model, the approach of [24] may be potentially sensitive to spatial variation and noise, thus leading to poor interpretability. We address this problem by modeling the spatial sources through convolution of sparse maps at multiple resolutions, allowing to deal with signal non-stationarity and robustness to noise.

3. Methods

In the following sections a matrix will be denoted by an uppercase letter $\boldsymbol{X}$, its n-th row will be given by $\boldsymbol{X}_{n:}$ and its n-th column by $\boldsymbol{X}_{:n}$. A column vector will be denoted by a lowercase letter $\boldsymbol{x}$. Subscript indices will be used to index the elements of matrices, vectors or sets of scalars. Superscipt indices will allow to index the blocks of block diagonal matrices.

3.1. Individual time-shift

To account for the uncertainty of the time-line of individual measurements, we assume that the observations are defined with respect to an absolute temporal reference τ. This is performed through a time-warping function $t_p = \boldsymbol{f}_p(\tau)$, that models the individual time-reparameterization. We choose an additive parameterization such that:

$$\boldsymbol{f}_p(\tau) = \tau + \delta_p. \tag{1}$$

Within this setting the individual time-shift δ_p encodes the temporal position of subject p, which in our application can be interpreted as the disease stage of subject p with respect to the long-term disease trajectory. We denote by $\boldsymbol{\delta} = \{\delta_p\}_{p=0}^{P}$ the set of time-shift parameters.

3.2. Data modeling

We represent the spatio-temporal data $\boldsymbol{D}$ by a block diagonal matrix in which we differentiate two main blocks $\boldsymbol{Y}$ and $\boldsymbol{V}$ as illustrated in Figure 1. Each sub-block $\boldsymbol{Y}^m$ is a matrix

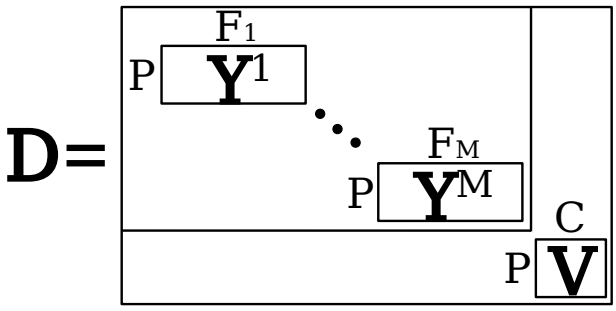

Figure 1: The block matrices $\boldsymbol{Y}^m$ contain the data from a specific imaging modality for each subjects. The matrix $\boldsymbol{V}$ contains the data from all the scalar modalities for each subjects.

containing the data represented by one of the M imaging modalities we wish to consider. These matrices have dimensions $P \times F_m$, where P denotes the number of subjects and F_m the number of imaging features for modality m, which in our case is the number of voxels. The matrix $\boldsymbol{V}$ accounts for non-imaging or scalar data such as clinical scores and has dimensions $P \times C$, where C is the number of scalar features considered. We postulate a generative model and decompose the data as shown in Figure 2. For each sub-block $\boldsymbol{Y}^m$, the data

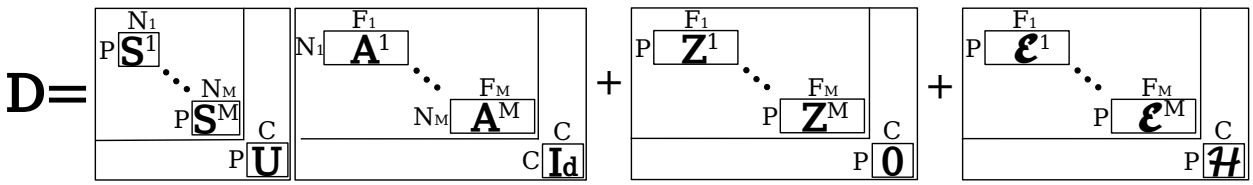

Figure 2: Spatio-temporal decomposition of each data block.

is factorized in a set of N_m spatio-temporal sources $\boldsymbol{Y}^m = \boldsymbol{S}^m \boldsymbol{A}^m$. The columns of the matrix $\boldsymbol{S}^m$ describe the temporal evolution of the corresponding spatial maps contained in

the rows of $\boldsymbol{A}^m$. The data in matrix $\boldsymbol{V}$ is modelled by a matrix $\boldsymbol{U}$ whose columns depict the temporal trajectories of the different scalar scores. In the case of imaging data, we also consider a constant term modeling brain areas which don't exhibit any intensity changes over time. This is done by including constant matrix terms $\boldsymbol{Z}^m$ that we need to estimate. We assume for a given modality m that the vectors $\boldsymbol{Z}^m_{p:}$ are common to every subjects. Finally, for each modality m, scalar score c, and subject p, we assume Gaussian observational noise $\boldsymbol{\mathcal{E}}^m_{p:} \sim \mathcal{N}(\boldsymbol{0}, \sigma_m^2 \boldsymbol{I})$, and $\boldsymbol{\mathcal{H}}_{p,c} \sim \mathcal{N}(0, \nu_c^2)$ for respectively imaging and scalar information.

Therefore, if we consider the data from modality m and scalar c of patient p observed at time $\boldsymbol{f}_p(\tau)$ we have:

$$\begin{aligned} \boldsymbol{Y}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m, \psi_m) &= \boldsymbol{S}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m)\boldsymbol{A}^m(\psi_m) + \boldsymbol{Z}^m_{p:} + \boldsymbol{\mathcal{E}}^m_{p:}, \\ \boldsymbol{V}_{p,c}(\boldsymbol{f}_p(\tau), \theta_c) &= \boldsymbol{U}_{p,c}(\boldsymbol{f}_p(\tau), \theta_c) + \boldsymbol{\mathcal{H}}_{p,c}. \end{aligned} \tag{2}$$

We denote by θ_m and θ_c the temporal parameters related respectively to the modality m and scalar feature c, while ψ_m represents the set of spatial parameters of modality m. We assume conditional independence across modalities and scalar scores given the time-shift information:

$$p(\boldsymbol{Y}, \boldsymbol{V}|\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{\delta}, \sigma, \nu) = \Big(\prod_m p(\boldsymbol{Y}^m|\boldsymbol{A}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)\Big)\Big(\prod_c p(\boldsymbol{V}_{:c}|\boldsymbol{U}_{:c}, \boldsymbol{\delta}, \nu_c)\Big),$$

and according to the generative model, the data likelihood for the imaging modality m is:

$$\begin{aligned} p(\boldsymbol{Y}^m|\boldsymbol{A}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)\Big) = \Big(\prod_m \prod_p \frac{1}{(2\pi\sigma_m^2)^{\frac{F_m}{2}}} \exp(-\frac{1}{2\sigma_m^2}||\boldsymbol{Y}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m, \psi_m) \\ - \boldsymbol{S}^m_{p:}(\boldsymbol{f}_p(\tau), \theta_m)\boldsymbol{A}^m(\psi_m) - \boldsymbol{Z}^m_{p:}||^2)\Big). \end{aligned} \tag{3}$$

Naturally, a similar equation holds for $p(\boldsymbol{V}_{:c}|\boldsymbol{U}_{:c}, \boldsymbol{\delta}, \nu_c)$.

Within a Bayesian modeling framework, we wish to maximize the marginal log-likelihood $\log(p(\boldsymbol{Y}, \boldsymbol{V}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu))$, to obtain posterior distributions for the spatio-temporal processes. Since the derivation of this quantity in a closed-form is not possible, we tackle this optimization problem through stochastic variational inference. Based on this formulation, in what follows we illustrate our model by detailing the variational approximations imposed on the spatio-temporal sources, along with the priors and constraints we impose to represent the data (Sections 3.3 and 3.4). Finally, we detail the variational lower bound and optimization strategy in Section 3.5.

For ease of notation we will drop the m and c indexes in Sections 3.3 and 3.4. As a result the matrix $\boldsymbol{S}$ will indistinctly refer to either any $\boldsymbol{S}^m$ or $\boldsymbol{U}$, while matrix $\boldsymbol{A}$ will refer to any $\boldsymbol{A}^m$, and $\boldsymbol{Y}$ to any $\boldsymbol{Y}^m$. For a given modality m, the number of patients P will be indexed

by p, the number of sources N^m or the number of scalar scores C will be indexed by n, and finally f will index the number of imaging features F^m.

3.3. Spatio-temporal processes

3.3.1. Temporal sources

In order to flexibly account for non-linear temporal patterns, the temporal sources are encoded in a matrix $\boldsymbol{S}$ in which each column $\boldsymbol{S}_{:n}$ is a GP representing the evolution of source n and is independent from the other sources. To allow computational tractability within a variational setting, we rely on the GP approximation proposed in [9], through kernel approximation via random feature expansion [29]. Within this framework, a GP can be approximated as a Bayesian Neural Network with form: $\boldsymbol{S}_{:n}(\boldsymbol{t}) = \phi(\boldsymbol{t}(\boldsymbol{\omega}^n)^T)\boldsymbol{w}^n$. For example, in the case of the Radial Basis Function (RBF) covariance, $\boldsymbol{\omega}^n$ is a linear projection in the spectral domain. It is equipped with a Gaussian distributed prior $p(\boldsymbol{\omega}^n) \sim \mathcal{N}(\boldsymbol{0}, l_n\boldsymbol{I})$ with a zero-mean and a covariance parameterized by a scalar l_n, acting as the length-scale parameter of the RBF covariance. The non-linear basis functions activation is defined by setting $\phi(\cdot) = (\cos(\cdot), \sin(\cdot))$, while the regression parameter $\boldsymbol{w}^n$ is given with a standard normal prior. The GP inference problem can be conveniently performed by estimating approximated variational distributions for all the $\boldsymbol{\omega}^n$ and $\boldsymbol{w}^n$ (Section 3.5). We will respectively denote by $\boldsymbol{\Omega}$ and $\boldsymbol{W}$ the block diagonal matrices whose blocks are the $(\boldsymbol{\omega}^n)^T$ and $\boldsymbol{w}^n$. Considering the N temporal sources, we can write $p(\boldsymbol{\Omega}) = \prod_n p(\boldsymbol{\omega}^n)$ and $p(\boldsymbol{W}) = \prod_n p(\boldsymbol{w}^n)$.

We wish also to account for a steady evolution of the temporal processes, hence constraining the temporal sources to monotonicity. This is relevant in the medical case, where one would like to model the steady progression of a disease from normal to pathological stages. In our case, we want to constrain the space of the temporal sources to the set of solutions $\mathcal{C}_n = \{\boldsymbol{S}_{:n}(\boldsymbol{t}) \mid \boldsymbol{S}'_{:n}(\boldsymbol{t}) \geq 0 \quad \forall\, \boldsymbol{t}\}$. This can be done done consistently within the regression setting of [30], and in particular with the GP random feature expansion framework as shown in [20]. In that work, the constraint is introduced as a second likelihood term on the temporal sources dynamics:

$$p(\mathcal{C}|\boldsymbol{S}', \gamma) = \prod_{p,n}(1 + \exp(-\gamma \boldsymbol{S}'_{p,n}(\boldsymbol{t})))^{-1}, \tag{4}$$

where $\boldsymbol{S}'$ contains every derivatives $\boldsymbol{S}'_{:n}$, γ controls the magnitude of the monotonicity constraint, and $\mathcal{C} = \bigcap_n \mathcal{C}_n$. According to [20] this constraint can be specified through the parametric form for the derivative of each $\boldsymbol{S}_{:n}$:

$$\boldsymbol{S}'_{:n}(t) = \frac{d\phi(\boldsymbol{t}(\boldsymbol{\omega}^n)^T)}{d\boldsymbol{t}}\boldsymbol{w}^n. \tag{5}$$

This setting leads to an efficient scheme for estimating the temporal sources through stochastic variational inference (Section 3.5).

3.3.2. Spatial sources.

According to the model introduced in Section 3.2, each observation $\boldsymbol{Y}_{p:}$ is obtained as the linear combination at a specific time-point between the temporal and spatial sources. In

order to deal with the multi-scale nature of the imaging signal, we propose to represent the spatial sources at multiple resolutions. To this end, we encode the spatial sources in a matrix $\boldsymbol{A}$ whose rows $\boldsymbol{A}_{n:}$ represent a specific source at a given scale. The scale is prescribed by a convolution operator $\boldsymbol{\Sigma}^n$, which is a applied to a map $\boldsymbol{B}_{n:}$ that we wish to infer. This problem can be specified by defining $\boldsymbol{A}_{n:} = \boldsymbol{B}_{n:}\boldsymbol{\Sigma}^n$, where $\boldsymbol{\Sigma}^n$ is an $F \times F$ Gaussian kernel matrix imposing a specific spatial resolution. The length-scale parameter λ_n of the Gaussian kernel is fixed for each source, to force the model to pick details at that specific scale. Due to the high-dimension of the data we are modeling, performing stochastic variational inference in this setting raises scalability issues. For instance, if we assume a Gaussian distribution $\mathcal{N}(\mu_{\boldsymbol{B}_{n:}}, diag(\boldsymbol{\Lambda}))$ for $\boldsymbol{B}_{n:}$, the distribution of the spatial signal would be $p(\boldsymbol{A}_{n:}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{B}_{n:}}\boldsymbol{\Sigma}^n, \boldsymbol{\Sigma}^n diag(\boldsymbol{\Lambda})(\boldsymbol{\Sigma}^n)^T)$. As a result, sampling from $p(\boldsymbol{A}_{n:})$ is not computationally tractable due to the size of the covariance matrix, which prevents the use of standard inference schemes on $\boldsymbol{B}_{n:}$. This can be overcome thanks to the separability of the Gaussian convolution kernel [25, 22], according to which the 3D convolution matrix $\boldsymbol{\Sigma}^n$ can be decomposed into the Kronecker product of 1D matrices, $\boldsymbol{\Sigma}^n = \boldsymbol{\Sigma}^n_x \otimes \boldsymbol{\Sigma}^n_y \otimes \boldsymbol{\Sigma}^n_z$. This decomposition allows to efficiently perform standard operations such as matrix inversion, or matrix-vector multiplication [32]. Thanks to this choice, we recover tractability for the inference of $\boldsymbol{B}_{n:}$ through sampling, as required by stochastic inference methods [18].

3.4. Sparsity

In order to detect specific brain areas involved in neurodegeneration, we propose to introduce a sparsity constraint over the maps (or codes) $\boldsymbol{B}_{n:}$. Consistently with our variational inference scheme, we induce sparsity via *Variational Dropout* as proposed in [17]. This approach leverages on an improper log-scale uniform prior $p(|\boldsymbol{B}_{n:}|) \propto \prod_f \ 1/|\boldsymbol{B}_{n,f}|$, along with an approximate posterior distribution:

$$q_1(\boldsymbol{B}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{M}_{n:}, diag(\alpha_{n,1}\boldsymbol{M}^2_{n,1}...\alpha_{n,F}\boldsymbol{M}^2_{n,F})). \tag{6}$$

In this formulation, the dropout parameter $\alpha_{n,f}$ is related to the individual dropout probability $p_{n,f}$ of each weight by $\alpha_{n,f} = p_{n,f}(1 - p_{n,f})^{-1}$. When the parameter $\alpha_{n,f}$ exceeds a fixed threshold, the dropout probability $p_{n,f}$ is considered high enough to ignore the corresponding weight $\boldsymbol{M}_{n,f}$ by setting it to zero. However, this framework raises stability issues affecting the inference of the dropout parameters due to large-variance gradients, thus limiting $p_{n,f}$ to values smaller than 0.5. To tackle this problem, we leverage on the extension of *Variational Dropout* proposed in [26]. In this setting, the variance parameter is encoded in a new independent variable $\boldsymbol{P}_{n,f} = \alpha_{n,f}\boldsymbol{M}^2_{n,f}$, while the posterior distribution is optimized with respect to $(\boldsymbol{M}, \boldsymbol{P})$. Therefore, in order to minimize the cost function for large variance $\boldsymbol{P}_{n,f} \to \infty$ ($\alpha_{n,f} \to \infty$ i.e $p_{n,f} \to 1$), the value of the weight's magnitude must be controlled by setting to zero the corresponding parameter $\boldsymbol{M}_{n,f}$. As a result, by dropping out weights in the code, we sparsify the estimated spatial maps, thus better isolating relevant spatial sub-structures. Spatial correlations in the images are obtained thanks to the convolution operation detailed in Section 3.3.2.

3.5. Variational inference

We detailed in the previous sections the choices of priors and constraints that we apply to the spatio-temporal processes in order to plausibly model the data. To illustrate the overall formulation of the method, we provide in Figure 3 the graphical model over the M modalities in the case of imaging data. Naturally, this graph simplifies when we deal with scalar data as we don't need to account for any spatial dependence. To infer the time-shift parameter $\boldsymbol{\delta}$,

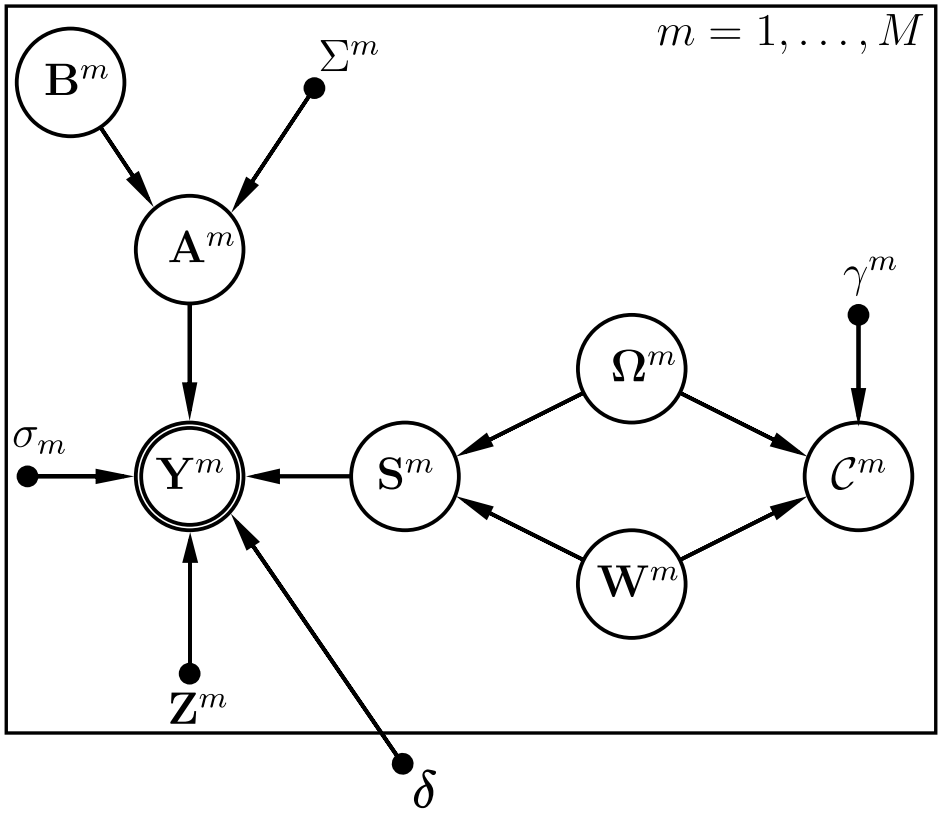

Figure 3: Graphical model for imaging data, $\boldsymbol{Y} = \{\boldsymbol{Y}^m\}$.

the sets of parameters θ_m, θ_c, and ψ_m, as well as $\boldsymbol{Z}$, σ and ν, we need to jointly optimize the data evidence according to priors and constraints:

$$\log(p(\boldsymbol{Y}, \boldsymbol{V}, \mathcal{C}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu, \gamma)) = \sum_m \log(p(\boldsymbol{Y}^m, \mathcal{C}^m|\boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) + \sum_c \log(p(\boldsymbol{V}_{:c}, \mathcal{C}^c|\boldsymbol{\delta}, \nu_c, \gamma_c)). \tag{7}$$

We tackle the optimization of Equation (7) via stochastic variational inference. Following [9] and [20] we introduce approximations, $q_2(\boldsymbol{\Omega}^m)$ and $q_3(\boldsymbol{W}^m)$ in addition to $q_1(\boldsymbol{B}^m)$ in order to derive a lower bound $\boldsymbol{\mathcal{L}}_m$ for each modality. We recall that the temporal trajectories $\boldsymbol{S}^m$ and $\boldsymbol{U}$ are treated similarly as described in Section 3.3.1. We also note that the choice of distributions q_1, q_2 and q_3 is the same across modalities, while their parameters will be

inferred independently. This leads to:

$$\begin{aligned}
\log(p(\boldsymbol{Y}^m, \mathcal{C}^m|\boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) \geqslant & \mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m))] \\
& + \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^m|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m))] \\
& - \mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)] - \mathcal{D}[q_2(\boldsymbol{\Omega}^m)||p(\boldsymbol{\Omega}^m)] - \mathcal{D}[q_3(\boldsymbol{W}^m)||p(\boldsymbol{W}^m)], \\
\log(p(\boldsymbol{V}_{c:}, \mathcal{C}^c|\boldsymbol{\delta}, \nu_c, \gamma_c)) \geqslant & \mathbb{E}_{q_2,q_3}[\log(p(\boldsymbol{V}_{c:}|\boldsymbol{\Omega}^c, \boldsymbol{W}^c, \boldsymbol{\delta}, \sigma_c))] \\
& + \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^c|\boldsymbol{\Omega}^c, \boldsymbol{W}^c, \boldsymbol{\delta}, \gamma_c))] \\
& - \mathcal{D}[q_2(\boldsymbol{\Omega}^c)||p(\boldsymbol{\Omega}^c)] - \mathcal{D}[q_3(\boldsymbol{W}^c)||p(\boldsymbol{W}^c)]
\end{aligned} \tag{8}$$

Where $\mathcal{D}$ refers to the Kullback-Leibler (KL) divergence. Combining the lower bounds of the different modalities we obtain:

$$\log(p(\boldsymbol{Y}, \boldsymbol{V}, \mathcal{C}|\boldsymbol{Z}, \boldsymbol{\delta}, \sigma, \nu, \gamma)) \geqslant \sum_m \boldsymbol{\mathcal{L}}_m + \sum_c \boldsymbol{\mathcal{L}}_c. \tag{9}$$

A detailed derivation of the lower bound is given in Appendix A.
The approximated distributions $q_2(\boldsymbol{\Omega}^m)$ and $q_3(\boldsymbol{W}^m)$ are factorized across GPs such that:

$$\begin{aligned}
q_2(\boldsymbol{\Omega}^m) &= \prod_{n=1}^{N_m} q_2(\boldsymbol{\omega}^n)^m = \prod_{n=1}^{N_m}\prod_{j=1}^{N_{rf}} \mathcal{N}(\boldsymbol{R}_{n,j}, \boldsymbol{Q}_{n,j}^2)^m, \\
q_3(\boldsymbol{W}^m) &= \prod_{n=1}^{N_m} q_3(\boldsymbol{w}^n)^m = \prod_{n=1}^{N_m}\prod_{j=1}^{N_{rf}} \mathcal{N}(\boldsymbol{T}_{n,j}, \boldsymbol{V}_{n,j}^2)^m,
\end{aligned} \tag{10}$$

where N_{rf} is the number of random features used for the projection in the spectral domain. Using Gaussian priors and approximations we introduced above, we can obtain a closed-form formula for the KL divergence. Moreover, the choice of prior and approximate posterior distribution for the maps of $\boldsymbol{B}^m$ leads to an approximation for the divergence $\mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)]$ detailed in [26]. This allows to analytically compute all the KL terms in our cost function. Formulas for the KL divergences are detailed in Appendix B.

Finally, we optimize the individual time-shifts $\boldsymbol{\delta} = \{\delta_p\}_{p=0}^P$, $\boldsymbol{Z}$, σ, ν as well as the overall sets of spatio-temporal parameters $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^M \cup \{\theta_c\}_{c=1}^C$ and $\boldsymbol{\psi} = \{\psi_m\}_{m=1}^M$.

$$\begin{aligned}
\boldsymbol{\theta} &= \{\boldsymbol{R}_{n:}^m, \boldsymbol{Q}_{n:}^m, \boldsymbol{T}_{n:}^m, \boldsymbol{V}_{n:}^m, l_n, n \in [1, N_m]\}_{m=1}^M \cup \{\boldsymbol{R}_{n:}^c, \boldsymbol{Q}_{n:}^c, \boldsymbol{T}_{n:}^c, \boldsymbol{V}_{n:}^c, l_n, n \in [1, N_c]\}_{c=1}^C, \\
\boldsymbol{\psi} &= \{\boldsymbol{M}_{n:}^m, \boldsymbol{P}_{n:}^m, n \in [1, N_m]\}_{m=1}^M.
\end{aligned} \tag{11}$$

Following [18] and using the reparameterization trick, we can efficiently sample from the approximated distributions q_1, q_2 and q_3 to compute the two expectation terms from (8) for each modality. We chose to alternate the optimization between the spatio-temporal parameters and the time-shift. We empirically set the γ_m to the minimum value that gives monotonic sources. The threshold for the dropout probability above which we set a weight

$\boldsymbol{B}^m_{n,f}$ to zero was fixed at 95% (i.e $\alpha = 19$), while the σ_m and ν_m were optimized during training along with the spatio-temporal parameters. The model is implemented and trained using the Pytorch library [27]. The complete experimental setting is detailed in Appendix C. In the following sections we will refer to our method as Monotonic Gaussian Process Analysis (MGPA).

4. Experiments and Results

In this section we first benchmark MGPA on synthetic data to demonstrate its reconstruction and separation properties while comparing it to standard sources separation methods. We finally apply our model on a large set of medical data from a publicly available clinical study, demonstrating the ability of our method to retrieve spatio-temporal processes relevant to AD, along with a time-scale describing the course of the disease.

4.1. Synthetic tests on spatio-temporal trajectory separation

For the synthetic tests we considered the case where the data is associated to a single imaging modality only. We tested MGPA on synthetic data generated as a linear combination of temporal functions and 3D activation maps at prescribed resolutions. The goal was to assess the method's ability to identify the spatio-temporal sources underlying the data. We benchmarked our method with respect to ICA, Non-Negative Matrix Factorization (NMF), and Principal Component Analysis (PCA), which were applied from the standard implementation provided in the Scikit-Learn library [28].

The benchmark was specified by defining a 10-folds validation setting, generating the data at each fold as a linear combination of temporal sources $\widetilde{\boldsymbol{S}}(\boldsymbol{t}) = [\widetilde{\boldsymbol{S}}_{:0}(\boldsymbol{t}), \widetilde{\boldsymbol{S}}_{:1}(\boldsymbol{t})]$, and spatial maps $\widetilde{\boldsymbol{A}} = [\widetilde{\boldsymbol{A}}_{0:}, \widetilde{\boldsymbol{A}}_{1:}]$. The data was defined as $\boldsymbol{Y}_{p:} = \widetilde{\boldsymbol{S}}_{p:}(\boldsymbol{t}_p)\widetilde{\boldsymbol{A}} + \boldsymbol{\mathcal{E}}_{p:}$ over 50 time points $\boldsymbol{t}_p$, where $\boldsymbol{t}_p$ was uniformly distributed in the range $[0, 0.7]$, and $\boldsymbol{\mathcal{E}}_{p:} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The temporal sources were specified as sigmoid functions $\widetilde{\boldsymbol{S}}_{p,i}(\boldsymbol{t}_p) = 1/(1 + \exp(-\boldsymbol{t}_p + \alpha_i))$, while the spatial structures had dimensions $(30 \times 30 \times 30)$ such that $\widetilde{\boldsymbol{A}}_{i:} = \widetilde{\boldsymbol{B}}_{i:}\widetilde{\boldsymbol{\Sigma}}^i$. The $\widetilde{\boldsymbol{\Sigma}}^i$ were chosen as Gaussian convolution matrices with respective length-scale of $\lambda = 2$ mm and $\lambda = 1$ mm. The $\widetilde{\boldsymbol{B}}_{i:}$ were randomly sampled sparse 3D maps.

Variable selection. We applied our method by specifying an over-complete set of six sources with respective spatial length-scale of $\lambda = \{2, 2, 1, 1, 0.5, 0.5\ mm\}$. Figure 4 shows an example of the sparse maps obtained for a specific fold. The model prunes the signal for most of the maps, while retaining two sparse maps, $\boldsymbol{B_{0:}}$ and $\boldsymbol{B_{4:}}$, whose length-scale are $\lambda = 2$ mm and $\lambda = 1$ mm, thus correctly estimating the right number of sources and their spatial resolution. As it can be qualitatively observed in Figure 4, we notice that the estimated sparse code convolved with a Gaussian kernel matrix with $\lambda = 1$ mm is closer to its ground truth than the one convolved with a length-scale $\lambda = 2$ mm. According to our tests, sparse codes associated to high resolution details (low λ) are indeed more identifiable. On the contrary, the identifiability of images obtained via a convolution operator with larger kernels (large λ) is lower, since these maps can be equivalently obtained through the convolution of

different sparse codes.

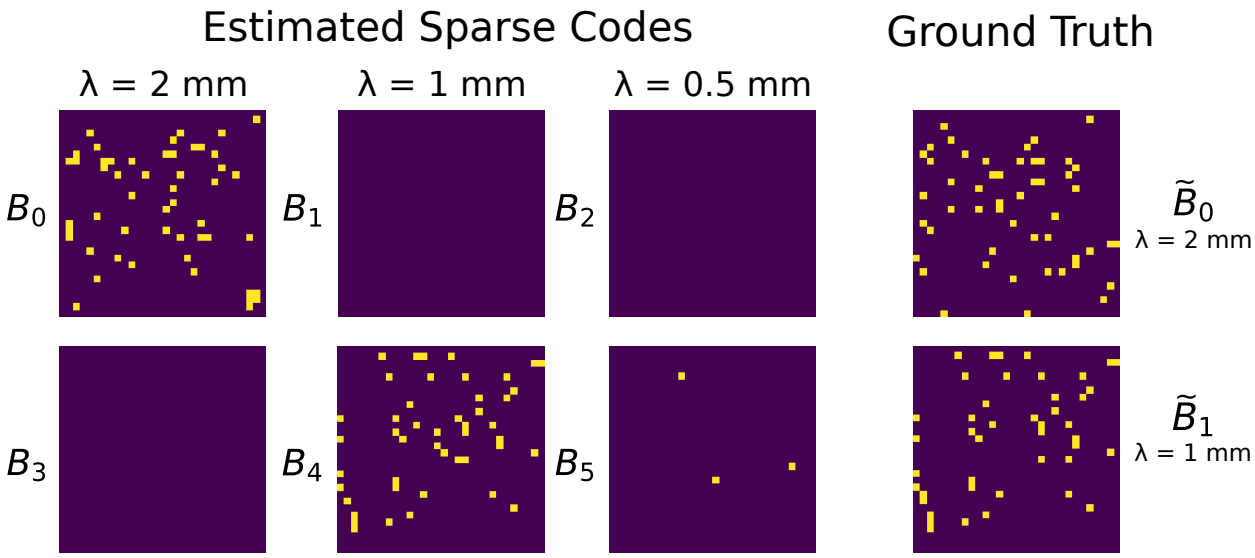

Figure 4: Slices extracted from the six sparse codes and the ground truth. Blue: Rejected points. Yellow: Retained points.

Sources separation. We observe in Table 1 that the lowest Mean-Squared Error (MSE) for the temporal sources reconstruction is obtained by MGPA, closely followed by ICA. Similarly, our model and ICA show the highest Structural Similarity (SSIM) score [34], which quantifies the image reconstruction accuracy with respect to the ground truth maps, while accounting for the inter-dependencies between neighbouring pixels. An example of image reconstruction from a sample fold is illustrated in Figure 5.

Table 1: MSE and SSIM between respectively the ground truth temporal and spatial sources with respect to the ones estimated by the different methods.

	Temporal (MSE)	Spatial (SSIM)
MGPA	$(\mathbf{8} \pm \mathbf{4}).\mathbf{10^{-5}}$	$\mathbf{98\%} \pm \mathbf{1}$
ICA	$(6 \pm 3).10^{-4}$	$97\% \pm 2$
NMF	$(3 \pm 2).10^{-2}$	$40\% \pm 17$
PCA	0.44 ± 10^{-3}	$15\% \pm 1$

4.2. Synthetic tests on trajectory separation and time-reparameterization

In this test, we modify the experimental benchmark by introducing a further element of variability associated to the time-axis. The temporal and spatial sources were modelled following the same procedure as in Section 4.1, however the observations were mixed along

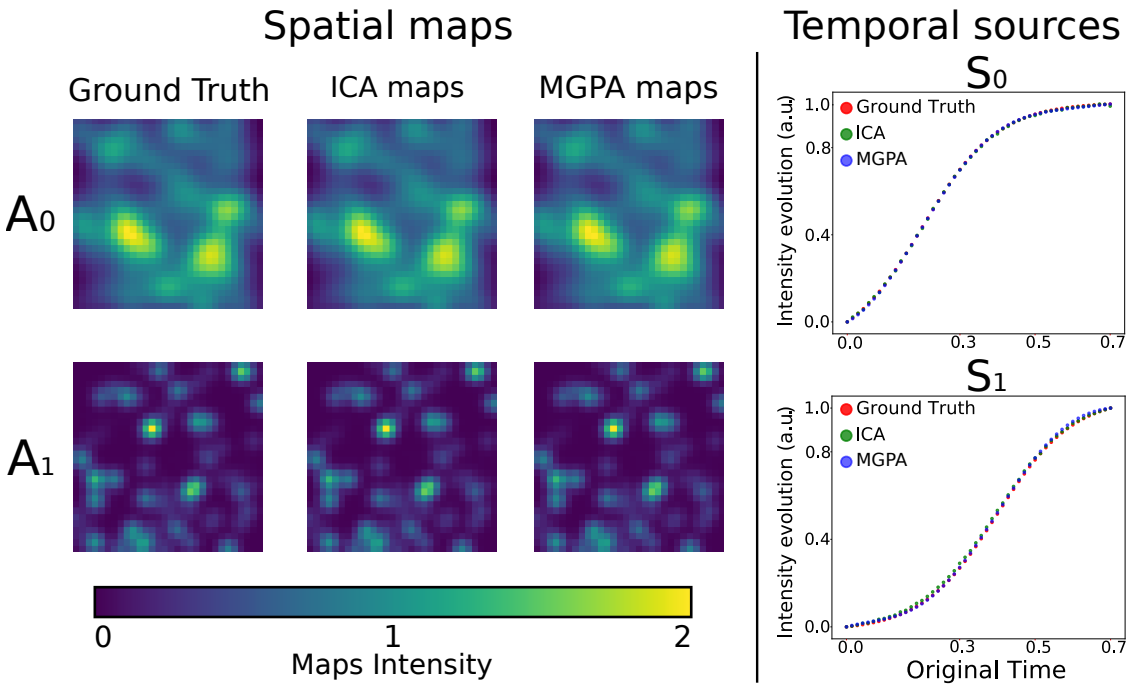

Figure 5: Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm), the maps estimated by ICA, and the ones estimated by MGPA. Temporal sources: Ground truth temporal sources (red) along with sources estimated by ICA (green) and MGPA (blue).

the temporal axis. To do so we generated longitudinal data as $\boldsymbol{Y}_{p,j,:} = \widetilde{\boldsymbol{S}}_{p:}(\boldsymbol{t})\widetilde{\boldsymbol{A}} + \boldsymbol{\mathcal{E}}_{j:}$, by sampling between 1 and 10 images per time-point and randomly re-arranging them along the time-axis (cf. time-shift t_p of each observation at initialization in Figures 6 and 7, panel "Time-Shift"). As a result, the method needs to estimate the spatio-temporal sources, while reconstructing the original time-ordering. Since the model is agnostic of a time-scale, we note that the time-shift may have a different range than the original time-axis. However, its relative ordering should be consistent with the original time points. We fitted a linear regression model over the 10 folds between the original time and the estimated time-shift parameter, and obtained an average R^2 coefficient of 0.98 with a standard deviation of 0.005 (cf. Table 2). This is illustrated for two different folds in the Time-Shift panel of Figures 6

Table 2: MSE and SSIM between respectively the ground truth temporal and spatial sources with respect to the ones estimated by MGPA. R^2 coefficient of the linear regression between the original time-line and the estimated time-shift.

	Temporal (MSE)	Spatial (SSIM)	R^2
MGPA	$(\mathbf{2} \pm \mathbf{0.8}).\mathbf{10^{-2}}$	$\mathbf{95\% \pm 4}$	$\mathbf{0.98 \pm 0.005}$

and 7, where we observe a strong linear correlation with the original time-line, meaning that the algorithm correctly re-ordered the data with respect to the original time-axis. However,

we notice in Table 2 that the MSE of the temporal sources significantly increased, due to the additional difficulty brought by the time-shift estimation. Indeed, in order to reconstruct the temporal signal we need to perfectly re-align hundreds of observations. This is the case in Figure 6 (optimal reconstruction result), where the time-shift is highly correlated with the original time-line, allowing to distinguish every single observation and reconstruct the original temporal profiles. Whereas in Figure 7 (sub-optimal reconstruction result), the estimated time-shift doesn't exhibit a perfect fit, and generally underestimates the time-reparameterization for the later and earlier time points. This is related to the challenging setting of reconstructing the time-line identified by the original temporal sources. Indeed, we observe that $\boldsymbol{S}_{:0}$ reaches a plateau for early time points, while $\boldsymbol{S}_{:1}$ is flat for later ones. This behaviour increases the difficulty of differentiating time points with low signal differences. As a result, it impacts the time-shift optimization and adds variability to the time-shift estimation performances, thus deteriorating the reconstruction of the temporal sources over the 10 folds compared to the previous benchmark. The spatial sources estimation remains comparable to the one without time-shift both quantitatively, with an average SSIM of 95%, and qualitatively, as shown in Figures 6 and 7. Within this setting, ICA, NMF and PCA poorly perform as they can't reconstruct the time-line. Interestingly, spatial ICA correctly estimated the spatial processes without however associating them to the corresponding temporal profile.

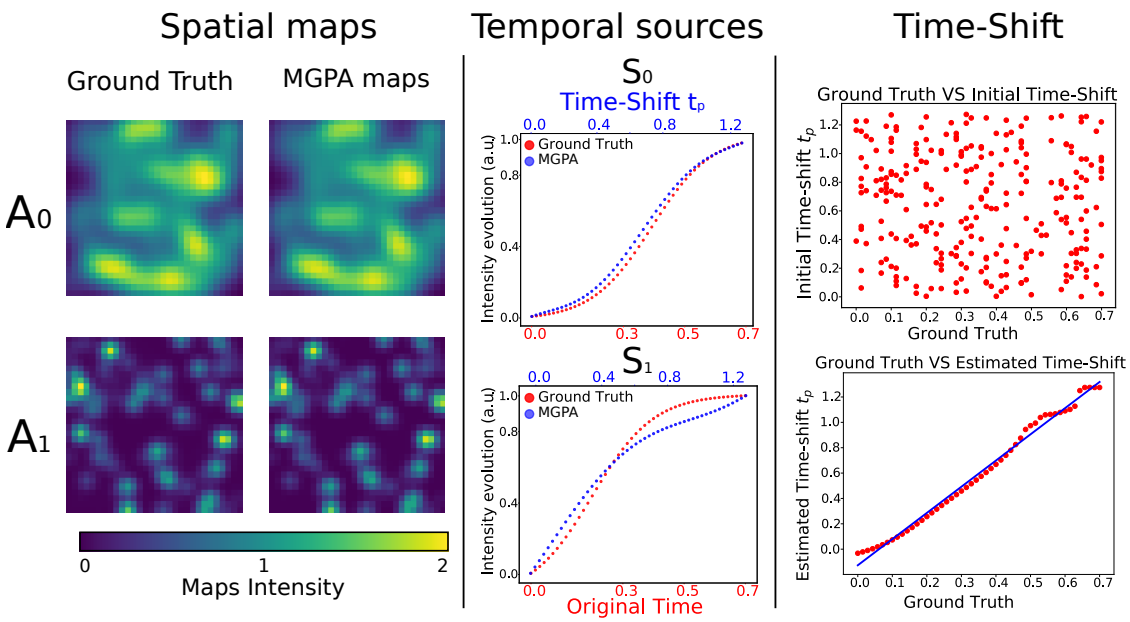

Figure 6: Optimal reconstruction result on synthetic data. Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm) and estimated spatial sources. Temporal sources: In red the original temporal sources, in blue the estimated temporal sources. Time-Shift: Time-shift t_p of each image at initialization (top), and after estimation (bottom). In blue, linear fit with the ground truth.

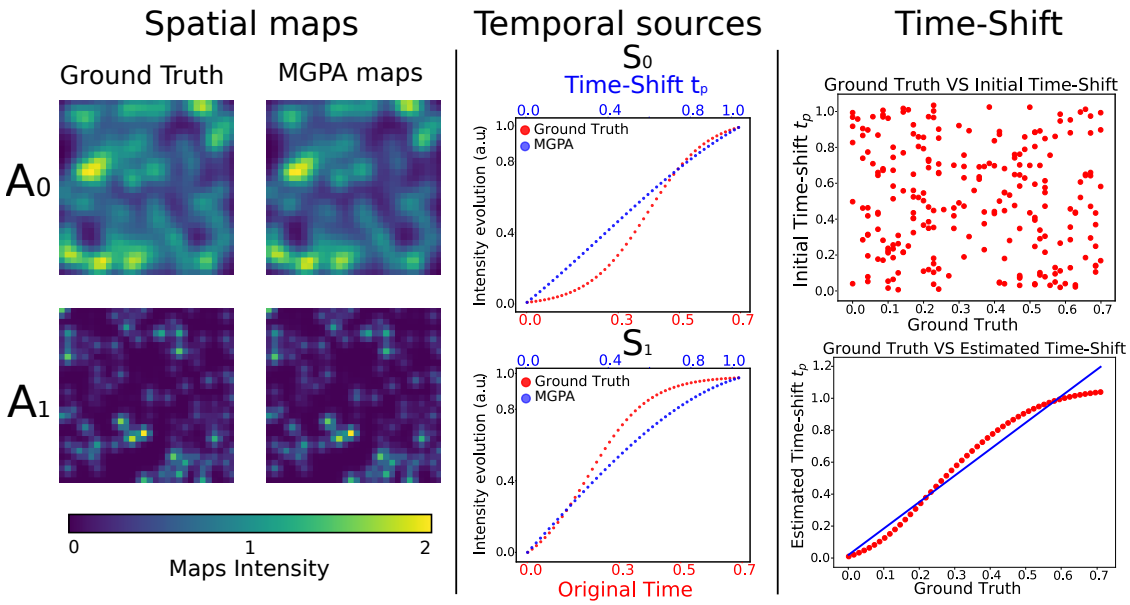

Figure 7: Sub-optimal reconstruction result on synthetic data. Spatial maps: Sample slice from ground truth images (A_0 $\lambda = 2$ mm, A_1 $\lambda = 1$ mm) and estimated spatial sources. Temporal sources: In red the original temporal sources, in blue the estimated temporal sources. Time-Shift: Time-shift t_p of each image at initialization (top), and after estimation (bottom). In blue, linear fit with the ground truth.

4.3. Application to spatio-temporal brain progression modeling

4.3.1. Data processing

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

We selected a cohort of 544 amyloid positive subjects of the ADNI database composed of 103 controls (NL), 164 Mild Cognitive Impairment (MCI), 114 AD patients, 34 healthy individuals converted to MCI or to AD (NL converter) and 129 MCI converted to AD (MCI converter). The term amyloid positive refers to subjects whose amyloid level in the cerebrospinal fluid (CSF) is below the nominal cutoff of 192 pg/ml. Conversion to MCI or AD was determined using the last follow-up available information. We provide in Table 3 socio-demographic and clinical information across the different groups.

The MRI, FDG-PET and AV45-PET of each individual were processed in order to obtain respectively, volumes of gray matter density, glucose uptake, and amyloid load in a standard anatomical space.

MRI processing protocol. Baseline MRI images were analyzed according to the SPM12 processing pipeline [2]. Each image was initially segmented into grey, white matter and CSF probabilistic maps. Grey matter images were used for the following analysis and normalized

Table 3: Baseline socio-demographic and clinical information for study cohort. Average values and standard deviation in parenthesis. NL: normal individuals, NL converter: normal subjects who converted to MCI or to AD, MCI: mild cognitive impairment, MCI converter: MCI subjects who converted to AD, AD: Alzheimer's patients. ADAS13: Alzheimer's Disease Assessment Scale-cognitive subscale, 13 items. FAQ: Functional Assessment Questionnaire. FDG: (18)F-fluorodeoxyglucose Positron Emission Tomography (PET) imaging. AV45: (18)F-florbetapir Amyloid PET imaging.

Group	NL	NL converter	MCI	MCI converter	AD
N	103	34	164	129	114
Age	73 (6)	78 (5)	73 (7)	73 (7)	74 (8)
Education (yrs)	16.3 (3)	16 (3)	15.7 (3)	16 (3)	15.6 (3)
ADAS13	9.1 (4.4)	11.4 (4.3)	14.6 (5.5)	20.4 (6.5)	31.6 (8.5)
FAQ	0.3 (0.7)	0.2 (0.6)	1.9 (2.8)	5.0 (4.6)	13.5 (6.9)
Entorhinal (cm^3)	3.8 (0.5)	3.5 (0.5)	3.6 (0.6)	3.2 (0.7)	2.8 (0.6)
Hippocampus (cm^3)	7.4 (0.9)	6.9 (0.7)	6.9 (0.9)	6.4 (0.9)	5.9 (0.8)
Ventricles (cm^3)	31 (16)	42 (21)	39 (23)	40 (19)	48 (23)
Whole brain (cm^3)	1033 (104)	1019 (91)	1058 (103)	1037 (102)	1005 (115)
FDG	1.3 (0.1)	1.3 (0.1)	1.2 (0.1)	1.1 (0.1)	1.0 (0.1)
AV45	1.3 (0.2)	1.3 (0.1)	1.3 (0.2)	1.4 (0.2)	1.5 (0.2)

to a group-wise reference space via DARTEL [1]. The subsequent modeling was carried out on the normalised images at the original spatial resolution.

PET processing protocol. Individuals' baseline PET images were initially affinely aligned to the corresponding MRI. After scaling the intensities to the cerebellum, the images were normalized to the grey matter template obtained with DARTEL and smoothed with a FWHM parameter of 4.55.

The images have dimension $102 \times 130 \times 107$ before vectorization, leading to 1418820 spatial features per patient. These spatial features represent for each voxel their gray matter concentration in the case of MRI images, their glucose metabolism for FDG-PET images, or their amyloid concentration for AV45-PET images. To exploit the ability of our model to automatically adapt to different spatial scales, we chose to keep the MRI images at their native resolution for the analysis, and thus do not perform additional smoohting to equalize to the PET FWHM. In addition to the imaging data of each patient, we also integrate the ADAS13 score assessed by clinicians. High values of this score indicate a decline of cognitive abilities. We consider three matrices $\boldsymbol{Y}^{MRI}$, $\boldsymbol{Y}^{FDG}$, and $\boldsymbol{Y}^{AV45}$ of dimension (543×1418820) containing the images of all the subjects, and a matrix $\boldsymbol{V}$ of dimension (543×1) containing their ADAS13 score. From now on we will refer to the data as the block diagonal matrix containing the four matrices $\boldsymbol{Y}^{MRI}, \boldsymbol{Y}^{FDG}$, $\boldsymbol{Y}^{AV45}$, and $\boldsymbol{V}$ as described in Section 3.2. We note that the analysis is performed by only considering a single scan per imaging modality and ADAS13 score for each patient. Therefore, the temporal evolution has to be inferred solely through the analysis of relative differences between the brain morphologies, glucose

metabolisms, amyloid concentrations and cognitive abilities across individuals.

4.3.2. Model specification

We aim at showing how MGPA applied on the data extracted from the ADNI cohort is able to temporally re-align patients in order to describe AD progression in a plausible way, while detecting relevant spatio-temporal processes at stake in AD. The temporal sources $\boldsymbol{S}^{MRI}$ and $\boldsymbol{S}^{FDG}$ associated to the loss of gray matter, and to the decrease of glucose uptake, are enforced to be monotonically decreasing. On the contrary, the temporal sources $\boldsymbol{S}^{AV45}$ and $\boldsymbol{U}_{:ADAS13}$, modeling respectively the evolution of amyloid concentration, and ADAS13 score, are enforced to be monotonically increasing. Since we don't consider any information about the disease stage of each individual before applying our method, all the observations are initialized at the same time reference $\tau = 0$. Therefore, as for the tests in Section 4.2, the time-shift reparameterization describes a relative re-ordering of the subjects not related to a specific time-unit. To decompose the imaging data we apply our model by specifying an over-complete basis of six sources with $\lambda = \{8, 8, 4, 4, 2, 2\ mm\}$, to cover both different scales and the associated variety of temporal evolution. Due to the high-dimension of the data matrix, the computations were parallelized over six GPUs, and the model required eighteen hours to complete the training.

4.3.3. Estimated spatio-temporal brain dynamics

In Figure 8 we show the spatio-temporal processes retained by the model for each imaging modality. Interestingly, the model adapts to the spatial resolution of MRI and PET images. Indeed, we notice that the model accounts for the high-resolution of MRI images by retaining a source associated to the lowest length-scale ($\lambda = 2$ mm). Concerning PET data, we observe that the induced sparsity discards the highest resolution codes ($\lambda = 2$ mm) for both FDG and AV45, highlighting the ability of the model to adapt to the coarser resolution of the PET signal.

In the case of MRI data, two sources were retained at two different resolutions ($\lambda = 4$ mm and $\lambda = 2$ mm). Source $\boldsymbol{S}_4^{MRI}$ describes a gray matter loss encompassing a large extent of the brain with a focus on cortical areas (see $\boldsymbol{A}_4^{MRI}$). We note that this map also targets subcortical areas such as the hippocampi, which are key regions of the AD pathology. Source $\boldsymbol{S}_2^{MRI}$ ($\lambda = 4$ mm) indicates a mild decrease of gray matter which accelerates in the latest stages of the disease, and targets the temporal poles (see $\boldsymbol{A}_2^{MRI}$). It is interesting to notice that this differential pattern of gray matter loss also affects the parahippocampal region, whose atrophy is known to be prominent in AD [11]. These results underline the complex evolution of brain atrophy, and the ability of the model to disentangle spatio-temporal processes mapping different regions involved in the pathology [3, 12]. Concerning the spatio-temporal processes extracted from the FDG-PET data, we see on Figure 8 that the model retained two sources at the coarsest resolutions ($\lambda = 8$ mm). Source $\boldsymbol{S}_1^{FDG}$ indicates a pattern of hypometabolism that tends to plateau and which involves most of the brain regions, thus describing a global effect of the pathology on the glucose uptake. Source $\boldsymbol{S}_0^{FDG}$ describes a linear pattern of hypometabolism targeting only areas such as the precuneus and the parietal lobe, which are known to be strongly affected during the evolution of the disease [5]. Finally,

the model extracted two spatio-temporal sources from the AV45-PET data at two different resolutions ($\lambda = 8$ mm and $\lambda = 4$ mm). We observe that source $\boldsymbol{S}_2^{AV45}$ suggests an increase of amyloid deposition mapping a large extent of the brain, such as the parietal and frontal lobes as well as temporal areas, thus concurring with clinical evidence [31]. Similarly to the FDG-PET processes, we have a source $\boldsymbol{S}_0^{AV45}$ exhibiting a differential pattern of amyloid deposition targeting mostly the occipital lobe and the precuneus.

The estimated spatio-temporal processes can be combined to obtain an estimated evolution $\boldsymbol{S}^m\boldsymbol{A}^m$ of the brain along the time-shift axis for each modality. In Figure 9, we show the ratio $|\boldsymbol{S}^m_{p:}\boldsymbol{A}^m - \boldsymbol{S}^m_{0:}\boldsymbol{A}^m|/\boldsymbol{S}^m_{0:}\boldsymbol{A}^m$ between the image predicted at four time-points t_p and the image predicted at t_0 for the three imaging modalities. This allows us to visualize the trajectory of a brain going from a healthy to a pathological state in terms of atrophy, glucose metabolism and amyloid load according to our model.

4.3.4. Model Consistency

To verify the plausibility of the fitted model, we compare in Figure 10 the concentration predicted by the model and the raw concentration measures in different brain areas for the three imaging modalities. We observe a decrease of gray matter and glucose metabolism as we progress along the estimated time-line, allowing to relate large time-shift values to lower gray matter density and glucose uptake. Moreover, we notice the agreement between the predictions made by the model (in blue) and the raw concentration measures (in red). In the case of AV45 data there is only a mild increase of amyloid load according to the model, probably due to the fact that the subjects selected in the cohort are already amyloid positive. As a result, they already show a high baseline amyloid level concentration, close to plateau levels.

In Figure 11, we show the estimated GP $\boldsymbol{U}_{:ADAS13}$. We observe that the model is able to plausibly describe the evolution of this cognitive score, while demonstrating a larger variability than in the case of imaging modalities.

4.3.5. Plausibility with respect to clinical evidence

We assessed the clinical relevance of the estimated time-shift by relating it to independent medical information. To this end, we compared the estimated time-shift to ADAS11, MMSE and FAQ scores. High values of ADAS11 and FAQ or low values of MMSE indicate a decline of performances. We show in Figure 12 that the estimated time-shift correlates with a decrease of cognitive and functional abilities. Moreover, we notice a non-linear relationship between the scores and the time-shift, suggesting an acceleration of symptoms along the estimated time-course, which is characteristic of AD in its latest stages.

The box-plot of Figure 13 shows the time-shift distribution across clinical groups. We observe an increase of the estimated time-shift when going from healthy to pathological stages. The high uncertainty associated to the MCI group is due to the broad definition of this clinical category, which includes subjects not necessarily affected by dementia. We note that the MCI subjects subsequently converted to AD (MCI converter) exhibit higher

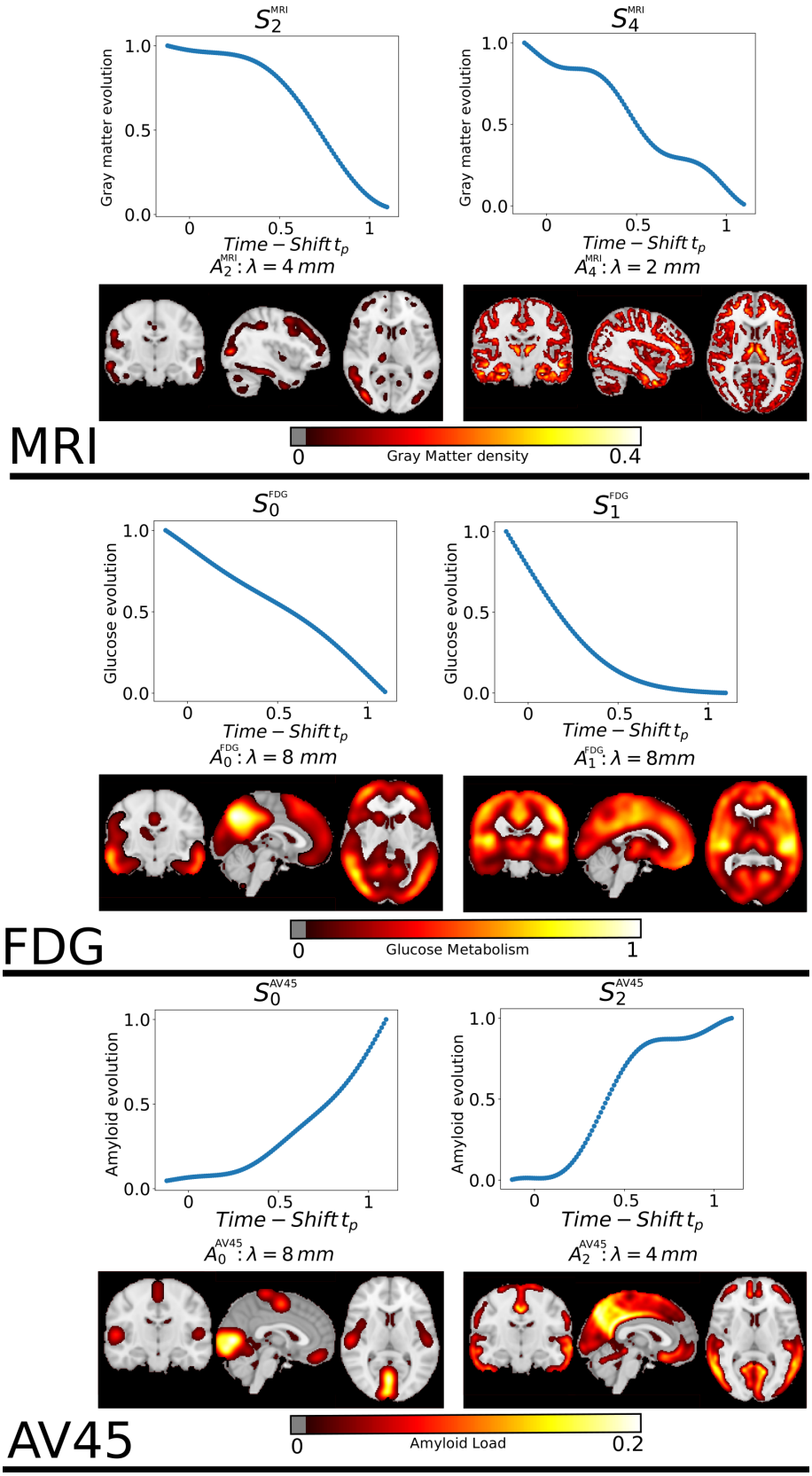

Figure 8: Estimated spatio-temporal processes for the three imaging modalities. The time-scale was rescaled to the aribtrary range [0, 1].

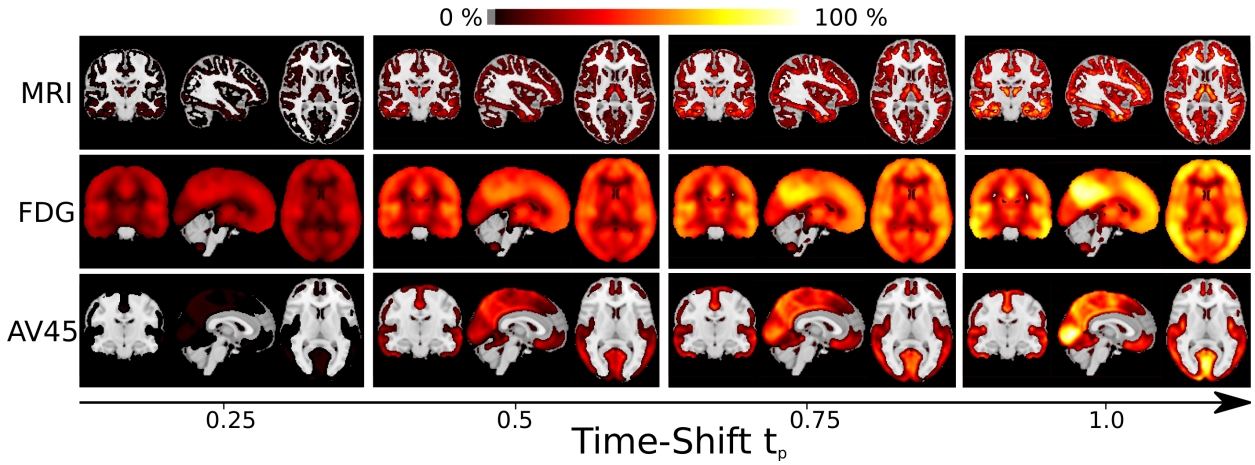

Figure 9: Ratio between the model prediction at time t_p and the prediction at t_0 for the three imaging modalities. The time-scale was rescaled to the aribtrary range [0, 1].

time-shift than the MCI group, highlighting the ability of the model to differentiate clinical diagnosis without any prior knowledge. A similar distinction can be noticed between the NL and NL converter groups. We found significant differences between median time-shift for NL-NL converter, MCI-MCI converter and MCI converter-AD (comparisons $p < 0.01$, Figure 13). It is also important to recall that this result is obtained from the analysis of a single scan per imaging modality and ADAS13 score for each patient.

5. Discussion

We presented a generative approach to spatio-temporal disease progression modeling based on matrix factorization across temporal and spatial sources. The proposed application on large set of medical images shows the ability of the model to disentangle relevant spatio-temporal processes at stake in AD, along with an estimated time-scale related to the disease evolution.

There are several avenues of improvement for the proposed approach. We found that the optimization is highly sensitive to the initialization of the spatial sources. This is typical of such complex non-convex problems, and requires further investigations to better control the algorithm convergence. More generally, the problem of source separation tackled in this work is intrinsically ill-posed, as the given data can be explained by several solutions. This was illustrated for example in our tests on synthetic data (Section 4.2), where the identification of the sources was more challenging in the case of coarse resolution codes and of flat temporal sources. We note however that this issue is general, and intrinsic to the problem of disease progression modeling. Finally, as mentioned in Section 3.4, the *Variational Dropout* framework leads to stability issues affecting inference, which are mostly due to the use of an improper prior. This problem may motivate the identification of alternative ways to induce sparsity on the spatial maps.

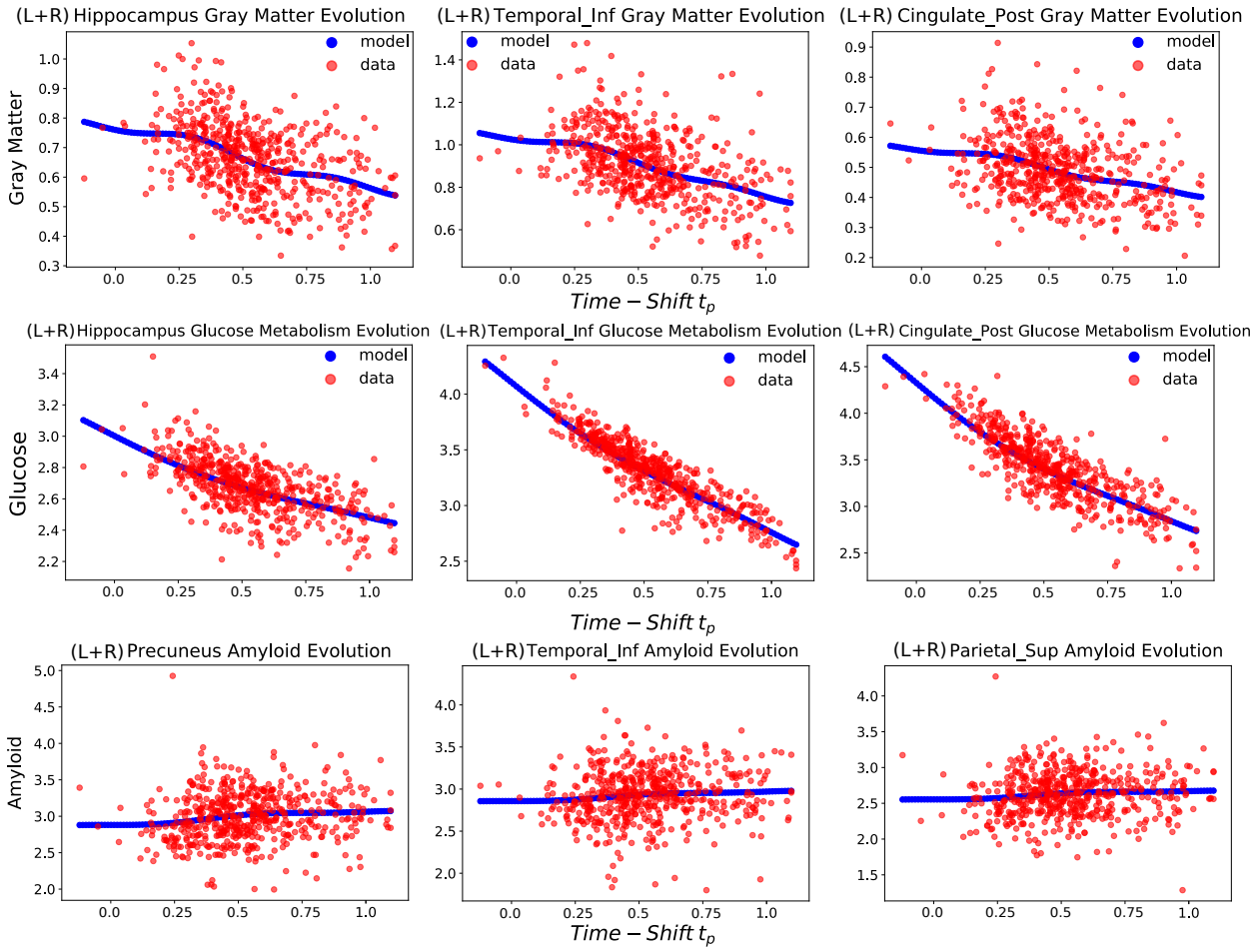

Figure 10: Model prediction averaged on specific brain areas (blue line), and observed values (red dots), along the estimated time-line for the three imaging modalities. L and R respectively stand for left and right. The time-scale was rescaled to the aribtrary range [0, 1].

In this work, we modeled the time-shift of each subject as a translation with respect to a common temporal reference. However, since pathological trajectories are different across individuals, it would be valuable to account for individual speed of progressions by introducing a scaling effect, as it has been proposed for example in [19, 33]. This was not in the scope of the current study, as we focused on the analysis of cross-sectional data, thus having only one data point per subject. Therefore, one of the main extensions of this model will be the integration of longitudinal data for each individual, which will allow a more specific time-reparameterization.

The modeling results are also sensitive to the specification of the spatio-temporal processes priors. In our case, the monotonicity constraint imposed to the GPs may be too restrictive to completely capture the complexity of the progression of neurodegeneration. From a more clinical point of view, the model could also benefit from the integration of data measuring the concentration of Tau protein via PET imaging, in order to quantify key neurobiological processes associated to AD [16].

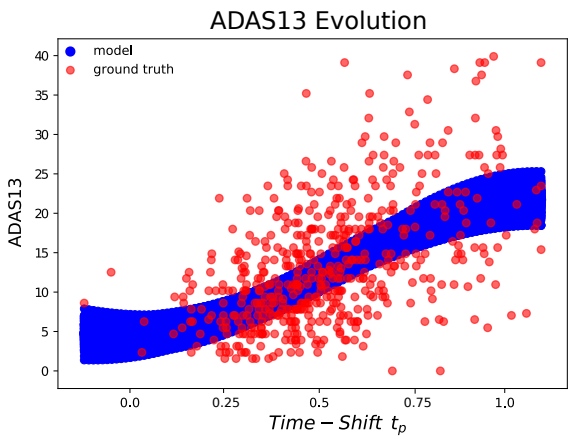

Figure 11: Model prediction of the ADAS13 score (blue line), and observed values (red dots) along the estimated time-line. The time-scale was rescaled to the aribtrary range [0, 1].

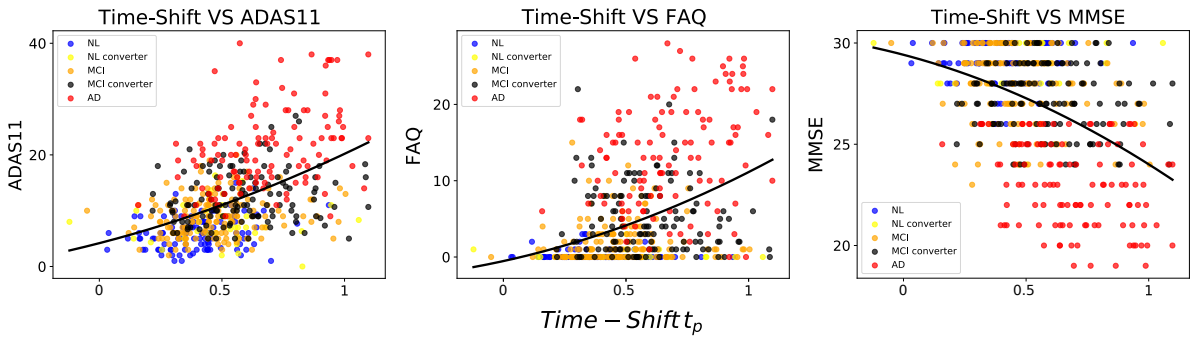

Figure 12: Evolution of the ADAS11 (left), FAQ (middle) and MMSE (right) along the estimated time-line. The time-scale was rescaled to the aribtrary range [0, 1].

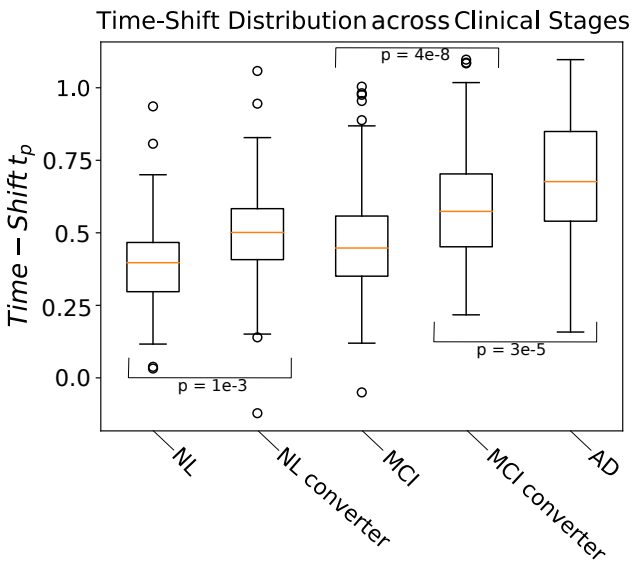

Figure 13: Distribution of the time-shift values over the different clinical stages. The time-scale was rescaled to the aribtrary range [0, 1].

The model has only been applied to a cohort of amyloid positive subjects, which restricts the dynamics of evolution that we could estimate. Indeed, only considering these subjects narrows down the time-line of the pathology, as we study patients at potentially advanced disease stages. Therefore, it would be interesting in a future work to apply the model on a cohort including amyloid negative subjects, to model the brain dynamics over the whole disease natural history.

Finally, we wish to validate the model on different cohorts to demonstrate its generalization properties. The validation for each subject could be done by finding the time-point minimizing the cost between the data of the observed subject and the model estimation. The indication given by the estimated time-shift could then be compared with the clinical diagnosis of the subject, allowing to test the reliability of our model. This validation step would ultimately allow to use the model as a diagnostic instrument of AD.

6. Acknowledgements

This work has been supported by the French government, through the UCA^{JEDI} Investments in the Future project managed by the National Research Agency (ref.n ANR-15-IDEX-01), the grant AAP Sant 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) and DOD ADNI. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company;CereSpir, Inc.;Cogstate;Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.;Lumosity;Lundbeck;Merck & Co., Inc.; Meso Scale Diagnostics, LLC.;NeuroRx Research; Neurotrack Technologies;Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging;Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95 – 113.

[2] Ashburner, J., Friston, K. J., Jun 2000. Voxel-based morphometry–the methods. NeuroImage 11 (6 Pt 1), 805–821.

[3] Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., Marcus, D. S., Cairns, N. J., Xie, X., Blazey, T. M., Holtzman, D. M., Santacruz, A., Buckles, V., Oliver, A., Moulder, K., Aisen, P. S., Ghetti, B., Klunk, W. E., McDade, E., Martins, R. N., Masters, C. L., Mayeux, R., Ringman, J. M., Rossor, M. N., Schofield, P. R., Sperling, R. A., Salloway, S., Morris, J. C., 2012. Clinical and biomarker changes in dominantly inherited alzheimer's disease. New England Journal of Medicine 367 (9), 795–804, pMID: 22784036.

[4] Bilgel, M., Jedynak, B., Wong, D. F., Resnick, S. M., Prince, J. L., 2015. Temporal Trajectory and Progression Score Estimation from Voxelwise Longitudinal Imaging Measures: Application to Amyloid Imaging. Inf Process Med Imaging 24, 424–436.

[5] Brown, R. K., Bohnen, N. I., Wong, K. K., Minoshima, S., Frey, K. A., 2014. Brain PET in suspected dementia: patterns of altered FDG metabolism. Radiographics 34 (3), 684–701.

[6] Bullmore, E., Fadili, J., Maxim, V., Sendur, L., Whitcher, B., Suckling, J., Brammer, M., Breakspear, M., 2004. Wavelets and functional magnetic resonance imaging of the human brain. NeuroImage 23 Suppl 1, S234–249.

[7] Calhoun, V. D., Liu, J., Adali, T., Mar 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. NeuroImage 45 (1 Suppl), S163–172.

[8] Comon, P., Apr. 1994. Independent Component Analysis, a new concept? Signal Processing 36, 287–314.

[9] Cutajar, K., Bonilla, E. V., Michiardi, P., Filippone, M., 06–11 Aug 2017. Random feature expansions for deep Gaussian processes. In: Precup, D., Teh, Y. W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. Vol. 70 of Proceedings of Machine Learning Research. PMLR, International Convention Centre, Sydney, Australia, pp. 884–893.

[10] Donohue, M. C., Jacqmin-Gadda, H., Goff, M. L., Thomas, R. G., Raman, R., Gamst, A. C., Beckett, L. A., Jack, C. R., Weiner, M. W., Dartigues, J.-F., Aisen, P. S., 2014. Estimating long-term multivariate progression from short-term data. Alzheimer's & Dementia 10 (5, Supplement), S400 – S410.

[11] Echavarri, C., Aalten, P., Uylings, H. B., Jacobs, H. I., Visser, P. J., Gronenschild, E. H., Verhey, F. R., Burgmans, S., Jan 2011. Atrophy in the parahippocampal gyrus as an early biomarker of Alzheimer's disease. Brain Struct Funct 215 (3-4), 265–271.

[12] Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., Thompson, P. M., Feb 2010. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 6 (2), 67–77.

[13] Hackmack, K., Paul, F., Weygandt, M., Allefeld, C., Haynes, J. D., Aug 2012. Multi-scale classification of disease using structural MRI and wavelet transform. NeuroImage 62 (1), 48–58.

[14] Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., Trojanowski, J. Q., Jan 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 9 (1), 119–128.

[15] Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., Raunig, D., Jedynak, C. P., Caffo, B., Prince, J. L., Nov 2012. A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort. NeuroImage 63 (3), 1478–1486.

[16] Kametani, F., Hasegawa, M., 2018. Reconsideration of Amyloid Hypothesis and Tau Hypothesis in Alzheimer's Disease. Front Neurosci 12, 25.

[17] Kingma, D. P., Salimans, T., Welling, M., 2015. Variational dropout and the local reparameterization trick. CoRR abs/1506.02557.

[18] Kingma, D. P., Welling, M., 2013. Auto-encoding variational bayes. CoRR abs/1312.6114.

[19] Koval, I., Schiratti, J.-B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., Durrleman, S., Sep. 2017. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In: Medical Image Computing and Computer Assisted Intervention. Medical Image Computing and Computer Assisted Intervention. Quebec City, Canada.

[20] Lorenzi, M., Filippone, M., 10–15 Jul 2018. Constraining the dynamics of deep probabilistic models.

In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. Vol. 80 of Proceedings of Machine Learning Research. PMLR, Stockholmsmssan, Stockholm Sweden, pp. 3233–3242.

[21] Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., Ourselin, S., 2017. Probabilistic disease progression modeling to characterize diagnostic uncertainty: Application to staging and prediction in alzheimer's disease. NeuroImage.

[22] Lorenzi, M., Ziegler, G., Alexander, D. C., Ourselin, S., 2015. Efficient Gaussian Process-Based Modelling and Prediction of Image Time Series. Inf Process Med Imaging 24, 626–637.

[23] Mallat, S. G., Jul. 1989. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 11 (7), 674–693.

[24] Marinescu, R. V., Eshaghi, A., Lorenzi, M., Young, A. L., Oxtoby, N. P., Garbarino, S., Shakespeare, T. J., Crutch, S. J., Alexander, D. C., 2017. A vertex clustering model for disease progression: Application to cortical thickness images. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 134–145.

[25] Marquand, A. F., Brammer, M., Williams, S. C., Doyle, O. M., May 2014. Bayesian multi-task learning for decoding multi-subject neuroimaging data. NeuroImage 92, 298–311.

[26] Molchanov, D., Ashukha, A., Vetrov, D., 06–11 Aug 2017. Variational dropout sparsifies deep neural networks. In: Precup, D., Teh, Y. W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. Vol. 70 of Proceedings of Machine Learning Research. PMLR, International Convention Centre, Sydney, Australia, pp. 2498–2507.

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

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

[29] Rahimi, A., Recht, B., 2008. Random features for large-scale kernel machines. In: Platt, J. C., Koller, D., Singer, Y., Roweis, S. T. (Eds.), Advances in Neural Information Processing Systems 20. Curran Associates, Inc., pp. 1177–1184.

[30] Riihimäki, J., Vehtari, A., 13–15 May 2010. Gaussian processes with monotonicity information. In: Teh, Y. W., Titterington, M. (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Vol. 9 of Proceedings of Machine Learning Research. PMLR, Chia Laguna Resort, Sardinia, Italy, pp. 645–652.
URL `http://proceedings.mlr.press/v9/riihimaki10a.html`

[31] Rodrigue, K. M., Kennedy, K. M., Park, D. C., Dec 2009. Beta-amyloid deposition and the aging brain. Neuropsychol Rev 19 (4), 436–450.

[32] Saatçi, Y., 2011. Scalable inference for structured gaussian process models.

[33] Schiratti, J., Allassonnière, S., Colliot, O., Durrleman, S., 2015. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In: NIPS. pp. 2404–2412.

[34] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE TRANSACTIONS ON IMAGE PROCESSING 13 (4), 600–612.

[35] Whitwell, J. L., Nov 2010. Progression of atrophy in Alzheimer's disease and related disorders. Neurotox Res 18 (3-4), 339–346.

[36] Young, A. L., Oxtoby, N. P., Huang, J., Marinescu, R. V., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., Schott, J. M., Alexander, D. C., 2015. Multiple Orderings of Events in Disease Progression. Inf Process Med Imaging 24, 711–722.

Appendix A.

In this Appendix, we detail the complete derivation of the lower bound.

$$\begin{aligned}\log(p(\boldsymbol{Y}^m, \mathcal{C}^m|\boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m)p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{S}^m, \frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}|\boldsymbol{\delta}, \gamma)d\boldsymbol{B}^m d\boldsymbol{S}^m\Big] \\ &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m)p(\boldsymbol{B}^m) \\ &\qquad p(\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}|\boldsymbol{S}^m, \boldsymbol{\delta}, \gamma)p(\boldsymbol{S}^m)d\boldsymbol{B}^m d\boldsymbol{S}^m\Big].\end{aligned}$$

By observing that $\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}$ is completely identified by $\boldsymbol{S}^m$, the equation can be written as:

$$\begin{aligned}\log(p(\boldsymbol{Y}^m, \mathcal{C}^m|\boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m)p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{S}^m)d\boldsymbol{B}^m d\boldsymbol{S}^m\Big].\end{aligned}$$

Similarly this derivation can be applied to $\log(p(\boldsymbol{V}_{:c}, \mathcal{C}^c|\boldsymbol{\delta}, \nu_c, \gamma_c))$.

$$\begin{aligned}\log(p(\boldsymbol{Y}^m, \mathcal{C}^m|\boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m, \gamma_m)) &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{S}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\frac{d\boldsymbol{S}^m}{d\boldsymbol{t}}, \boldsymbol{\delta}, \gamma_m)p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{S}^m)d\boldsymbol{B}^m d\boldsymbol{S}^m\Big] \\ &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m)p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{\Omega}^m)p(\boldsymbol{W}^m)d\boldsymbol{B}^m d\boldsymbol{\Omega}^m d\boldsymbol{W}^m\Big] \\ &= \log\Big[\int p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m)p(\boldsymbol{B}^m) \\ &\qquad p(\boldsymbol{\Omega}^m)p(\boldsymbol{W}^m)\frac{q_1(\boldsymbol{B}^m)q_2(\boldsymbol{\Omega}^m)q_3(\boldsymbol{W}^m)}{q_1(\boldsymbol{B}^m)q_2(\boldsymbol{\Omega}^m)q_3(\boldsymbol{W}^m)}d\boldsymbol{B}^m d\boldsymbol{\Omega}^m d\boldsymbol{W}^m\Big] \\ &= \log\Big[\mathbb{E}_{q_1,q_2,q_3}\frac{p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m)}{q_1(\boldsymbol{B}^m)q_2(\boldsymbol{\Omega}^m)q_3(\boldsymbol{W}^m)} \\ &\qquad \frac{p(\boldsymbol{B}^m)p(\boldsymbol{\Omega}^m)p(\boldsymbol{W}^m)}{q_1(\boldsymbol{B}^m)q_2(\boldsymbol{\Omega}^m)q_3(\boldsymbol{W}^m)}\Big] \\ &\geq \mathbb{E}_{q_1,q_2,q_3}\Big(\log\Big[\frac{p(\boldsymbol{Y}^m|\boldsymbol{B}^m, \boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{Z}^m, \boldsymbol{\delta}, \sigma_m)p(\mathcal{C}|\boldsymbol{\Omega}^m, \boldsymbol{W}^m, \boldsymbol{\delta}, \gamma_m)}{q_1(\boldsymbol{B}^m)q_2(\boldsymbol{\Omega}^m)q_3(\boldsymbol{W}^m)} \\ &\qquad \frac{p(\boldsymbol{B}^m)p(\boldsymbol{\Omega}^m)p(\boldsymbol{W}^m)}{q_1(\boldsymbol{B}^m)q_2(\boldsymbol{\Omega}^m)q_3(\boldsymbol{W}^m)}\Big]\Big)\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{q_1,q_2,q_3}[\log(p(\boldsymbol{Y}^m|\boldsymbol{B}^m,\boldsymbol{\Omega}^m,\boldsymbol{W}^m,\boldsymbol{Z}^m,\boldsymbol{\delta},\sigma_m))] \\
&\quad + \mathbb{E}_{q_2,q_3}[\log(p(\mathcal{C}^m|\boldsymbol{\Omega}^m,\boldsymbol{W}^m,\boldsymbol{\delta},\gamma_m))] \\
&\quad - \mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)] - \mathcal{D}[q_2(\boldsymbol{\Omega}^m)||p(\boldsymbol{\Omega}^m)] - \mathcal{D}[q_3(\boldsymbol{W}^m)||p(\boldsymbol{W}^m)].
\end{aligned}
$$

This derivation gives us the lower bound $\boldsymbol{\mathcal{L}}_m$ of a given modality m. The same technique can be used to derive a lower bound for $\log(p(\boldsymbol{V}_{c:},\mathcal{C}^c|\boldsymbol{\delta},\nu_c,\gamma_c))$, and by summation over m and c we obtain the lower bound of Equation 9 for $\log(p(\boldsymbol{Y},\boldsymbol{V},\mathcal{C}|\boldsymbol{Z},\boldsymbol{\delta},\sigma,\nu,\gamma))$.

Appendix B.

In this section we provide formulas for computing the three KL terms of the lower bound. The total KL divergences are:

$$
\begin{aligned}
\mathcal{D}[q_1(\boldsymbol{B})||p(\boldsymbol{B})] &= \sum_m \mathcal{D}[q_1(\boldsymbol{B}^m)||p(\boldsymbol{B}^m)], \\
\mathcal{D}[q_2(\boldsymbol{\Omega})||p(\boldsymbol{\Omega})] &= \sum_m \mathcal{D}[q_1(\boldsymbol{\Omega}^m)||p(\boldsymbol{\Omega}^m)] + \sum_c \mathcal{D}[q_1(\boldsymbol{\Omega}^c)||p(\boldsymbol{\Omega}^c)], \\
\mathcal{D}[q_3(\boldsymbol{W})||p(\boldsymbol{W})] &= \sum_m \mathcal{D}[q_3(\boldsymbol{W}^m)||p(\boldsymbol{W}^m)] + \sum_c \mathcal{D}[q_3(\boldsymbol{W}^c)||p(\boldsymbol{W}^c)].
\end{aligned}
$$

For ease of notation we will drop the m and c indices and will give formulas for a single modality. In [26], authors provide an approximation of the KL for the maps $\boldsymbol{B}$:

$$
-\mathcal{D}[q_1(\boldsymbol{B})||p(\boldsymbol{B})] = \sum_{n,f} k_1 h(k_2 + k_3 \log(\alpha_{n,f})) - 0.5\log(1+\alpha_{n,f}^{-1}) - k_1,
$$

where h is the sigmoid function and $k_1 = 0.63576,\ k_2 = 1.87320,\ k_3 = 1.48695$.

In the case of $\boldsymbol{\Omega}$ and $\boldsymbol{W}$, we've seen that they have Gaussian priors and approximations which are detailed in Sections 3.3.1 and 3.5. As a result we can obtain closed-form formulas for their KL, leading to:

$$
\begin{aligned}
\mathcal{D}[q_2(\boldsymbol{\Omega})|p(\boldsymbol{\Omega})] &= \frac{1}{2}\sum_{n,j} \boldsymbol{Q}_{n,j}^2 l_n + \boldsymbol{R}_{n,j}^2 l_n - 1 - \log(\boldsymbol{Q}_{n,j}^2 l_n), \\
\mathcal{D}[q_3(\boldsymbol{W})|p(\boldsymbol{W})] &= \frac{1}{2}\sum_{n,j} \boldsymbol{V}_{n,j}^2 + \boldsymbol{T}_{n,j}^2 - 1 - \log(\boldsymbol{V}_{n,j}^2).
\end{aligned}
$$

By summation over the different modalities we finally obtain the total KL divergences.

Appendix C.

We provide in this Appendix details for the experiments on real data.

- The number of random features for the GP estimation was set to 10, as it was enough to recover the temporal sources in the synthetic experiments.
- The lower bound was optimized using the ADAM optimizer.
- We used an alternate optimization scheme between the spatio-temporal parameters and the time-shift of [2000, 1000] iterations repeated 20 times, followed by 30000 iterations in which we only optimized the spatio-temporal parameters.
- The expectation terms in the lower bound were approximated using only one Monte-Carlo sample as proposed in [18].
- The table below gives the learning rates (LR) of all the parameters of the model.

Table 1: Learning rates (LR) of the different parameters of the model.

	$\boldsymbol{\theta}$	$\boldsymbol{M}$	$\boldsymbol{P}$	$\boldsymbol{Z}$	$\sigma,\ \nu$	$\boldsymbol{\delta}$
LR	10^{-2}	10^{-3}	10^{-1}	10^{-1}	10^{-2}	10^{-4}