Robust Bayesian Analysis of Early-Stage Parkinson's Disease Progression Using DaTscan Images

Yuan Zhou, *Member, IEEE,* Sule Tinaz, and Hemant D. Tagare, *Senior Member, IEEE*

***Abstract*—This paper proposes a mixture of linear dynamical systems model for quantifying the heterogeneous progress of Parkinson's disease from DaTscan Images. The model is fitted to longitudinal DaTscans from the Parkinson's Progression Marker Initiative. Fitting is accomplished using robust Bayesian inference with collapsed Gibbs sampling. Bayesian inference reveals three image-based progression subtypes which differ in progression speeds as well as progression trajectories. The model reveals characteristic spatial progression patterns in the brain, each pattern associated with a time constant. These patterns can serve as disease progression markers. The subtypes also have different progression rates of clinical symptoms measured by MDS-UPDRS Part III scores.**

***Index Terms*—Parkinson's disease, disease progression model, DaTscans, linear dynamical system, centrosymmetric matrix, *t*-distribution**

I. Introduction

Parkinson's disease (PD) is a neurodegenerative disease characterized by the loss of dopaminergic neurons in the substantia nigra. Different individuals with PD progress along different disease trajectories. This variability is called *progression heterogeneity*, or simply, *heterogeneity*. Heterogeneity is understood in terms of *progression subtypes*, each subtype being a prototypical progression trajectory.

Another characteristic of PD progression is that it exhibits specific spatial patterns in the brain. These patterns, called *Braak stages* [1], have mostly been analyzed by histology of deceased PD patient brains. Spatial progression patterns in living PD patients have not yet been reported using DaTscans.

The goal of this paper is to propose a mathematical model and a Bayesian analysis method to (i) quantify PD heterogeneity by identifying progression subtypes, and (ii) to identify spatial progression patterns and their time constants in living PD patients using longitudinal analysis of *DaTscan images*, or DaTscans. DaTscans are the commercial name for SPECT imaging with 123I-FP-CIT. DaTscans measure the local presynaptic dopamine transporter (DAT) density. DAT density decreases as dopaminergic neurons are lost in PD, and this manifests as signal loss in DaTscan images [2]. In early-stage PD, signal loss is most significant in the striatum [1], hence we study the dynamics of PD progression using the striatal binding ratio (SBR) [3] in the caudates and putamina. The SBR at voxel v is defined as $SBR_v = (I_v - \mu)/\mu$ where I_v is the intensity in voxel v and μ is the mean (or median) intensity in a reference region, such as the occipital lobe, that does not have specific ligand binding. The SBR normalizes for radioligand dose as well as compensates for the amount of nonspecific radioligand binding.

The model we propose is a *mixture of linear dynamical systems* (MLDS). In this model, PD subjects are assigned to different progression subtypes, where each subtype is defined by a multivariate linear dynamical system (LDS). The eigenvectors of the transition matrix of the dynamical system give spatial progression patterns of DAT loss in the brain. The corresponding eigenvalues give time constants of disease progression along these patterns. The data used to fit the model comes from the Parkinson's Progression Marker Initiative (PPMI) (https://www.ppmi-info.org/).

This paper introduces several novel techniques for PD DaTscan image analysis, and we briefly summarize them here: First, we model coupled progression of the disease in several regions of interest. Second, our model is specifically designed to capture progression heterogeneity. This is in contrast to most previous PD SPECT or PET image analyses, which only model a single region-of-interest (ROI) at a time (e.g. [4]) and do not model heterogeneity. Third, we identify a new constraint called *population mirror symmetry*. A justification for the constraint, based on DaTscan data, is presented in Section II-E. Finally, we use Bayesian analysis with a robust *t*-distribution to model the residues. Using the *t*-distribution makes the parameter estimates robust to outliers [5], [6].

The paper is organized as follows: We begin in Section II by a brief review of disease progression literature and of PD progression. The MLDS model is explained in Section III. Bayesian inference for the model is in Section IV. The results of fitting the model to the data are reported in Section V. Section VI contains a discussion, while Section VII concludes the paper. Preliminary work using a Gaussian distribution to model the noise was reported in [7].

Yuan Zhou is with the Dept. of Radiology and Biomedical Imaging, Yale University, New Haven, CT, USA. Sule Tinaz is with the Dept. of Neurology, Yale University, New Haven, CT, USA. Hemant D. Tagare is with the Dept. of Radiology and Biomedical Imaging, Dept. of Biomedical Engineering, Dept. of Statistics and Data Science, Yale University, New Haven, CT, USA. E-mail: zhouyuanzxcv@gmail.com, sule.tinaz@yale.edu, hemant.tagare@yale.edu. Supplementary downloadable material is available at http://ieeexplore.ieee.org, provided by the authors. The data and code of this research are available online under the BSD 3-Clause License at https://github.com/tagarelab/Disease-Progression-Model. Please contact zhouyuanzxcv@gmail.com for further questions about this work. The work is partially supported by NIH grant R01NS107328.

II. Progression Models, PD, and the PPMI Dataset

A. Disease Progression Models

Most disease progression models (DPM) reported in the literature are for Alzheimer's disease. These DPMs model the temporal progress of biomarkers such as brain MRI regional volumes, cerebrospinal fluid measures, and clinical scores. DPMs can be categorized as either event-based, explicit function of time-based, or differential equation-based.

Event-based models are discrete in time; they define various disease stages and model transitions from one disease state to another. An example is [8], where transitions from normal to severe atrophy in different brain regions are defined as events. The model finds a consistent ordering of these events in a group of subjects. Enhanced versions of this basic model, with more events and applied to the ADNI dataset are in [9], and with different orderings for different groups of subjects and subject specific orderings in [10], [11].

In contrast to event-based models, explicit function models characterize the continuous longitudinal progress of biomarkers by a parametric or a non-parametric function of time and other variates. An example of parametric modeling is [12] which regresses covariates such as time, baseline age, brain regional volume with cognitive scores [12]. A similar scheme with subject specific time shift is used with the PPMI data in [13]. Non-linear models with logistic [14] or sigmoidal [15] functions are also used. For high-dimensional data, clustering is used to reduce the number of parameters [16]. An example of a non-parametric model is [17] where disease trajectories are modeled with a group-wise monotonic Gaussian process trajectory plus an individual trajectory. In the above models, time explicitly enters the regression. In contrast are models where image features at different time points are regressed to clinical scores [18], [19].

Differential equation models use a differential equation to model the longitudinal trajectories of biomarkers, e.g. [20], [21]. Neurodegenerative diseases progress by toxic protein transmission along neuronal pathways [22]. This suggests that modeling neuronal pathways as edges in a graph can lead to using diffusion on the graph as a model for disease progression [23], [24]. An extension adds regional sporadic stimulus [25] to the model. Recently, a graph-based differential equation has been applied to MRI images of PD relating atrophy patterns to diffusion seeded at the substantia nigra [26].

Bayesian analysis has been used with neurodegenerative DPMs before [8], [10], [12], [13]. However, to the best of our knowledge, Bayesian modeling of a mixture of linear dynamical systems has not been reported with PD DaTscans.

As mentioned above, most of the above methods are designed for Alzheimer's disease, and they predominantly use MRI images. In contrast, our goal is to model Parkinson's disease progression using DaTscans. DaTscans do not provide any connectivity information.

B. Early Stage Parkinson's Disease

PD progression in DaTscans is quantified by ROI analysis which shows that the mean SBR in the putamen and caudate decreases exponentially with time [27], [28]. Exponential decrease is also observed with PET (non-DaTscan) imaging tracers [4]. The rates of SBR decrease vary widely, from 5% to 13% per annum, indicating strong heterogeneity [27]. Because the putamen is affected before the caudate in the early stages [1], the difference between the mean SBR in the putamen and the caudate is taken as an indicator of disease progression [29].

Early-stage PD is also asymmetric; one brain hemisphere is affected more than the other [30]. Asymmetry is caused by a complex interplay of hereditary and environmental factors [31]. Initially, either brain hemisphere may be affected with almost equal probability, but the disease becomes more symmetric as it progresses.

C. Parkinson's Disease Subtypes

In the PD literature, subtypes are usually derived from clinical examination, i.e. from the Movement Disorders Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) scores, resulting in subtypes such as akinetic/rigid-dominant or tremor-dominant [32]. Typically, clinical progression subtypes are found by clustering the baseline clinical scores and comparing the progression rates of these clusters [33], [34]. A review of these methods is available in [32]. Recently, a complex combination of neural networks, dynamic time warping, t-SNE embedding and k-means was used to cluster the PPMI data [35] into subtypes. To the best of our knowledge progression subtypes have not been found so far using DaTscans.

D. The PPMI Dataset

The PPMI DaTscan dataset has 449 early-stage PD subjects. Their demographics are as follows: 65% of the subjects are male, 35% are female. Their ages at the time of entry into the study are $34-85$ years, with a median age of 63 years. The subjects are scanned at baseline, and then approximately at 1, 2, 4, and 5 years from baseline (the imaging protocol for the PPMI DaTscans is documented in http://www.ppmi-info.org/wp-content/uploads/2013/02/PPMI-Protocol-AM5-Final-27Nov2012v6-2.pdf). Not all subjects have a scan for all of these time points, and the scan times for different subjects are not exactly at $1, 2, 4, 5$ years.

The PPMI dataset also has longitudinal MDS-UPDRS scores for the subjects. We relate the image subtypes to Part III of the scores.

E. Population-level Mirror Symmetry

There is a population-level symmetry in the PD DaTscans. For every PD patient whose brain is asymmetrically affected in one direction, there is another patient whose brain is asymmetrically affected in the reverse direction. Moreover these subjects progress in a mirror image fashion. Fig. 1 illustrate this. The subfigures plot the time series of the mean SBR in the bilateral caudates and putamina for the PPMI subjects. The time series for every subject is plotted as a sequence of vectors, each vector pointing from a time point to the subsequent time point. The vectors are rendered with continuously changing colors that denote time from the baseline DaTscan. The relation between color and time is in

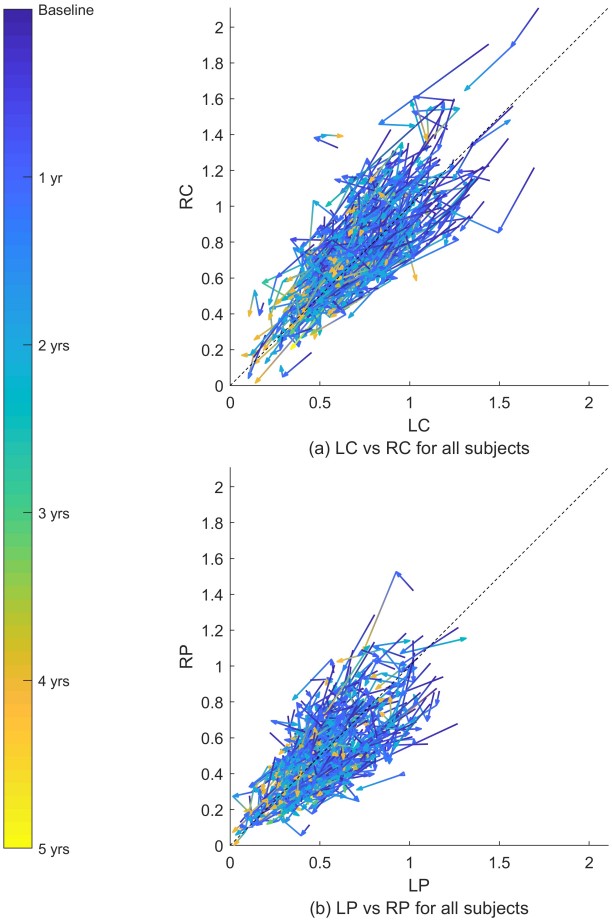


Fig. 1. Time series of the mean SBR in the left caudate (LC), right caudate (RC), left putamen (LP), and right putamen (RP) for 365 subjects from the PPMI dataset. The arrows are rendered with continuously changing colors that correspond to the 5 year period indicated by the colorbar on the left side. The asymmetry of the disease is observable at baseline; in approximately 45% of the subjects, the right hemisphere is more severely affected at baseline. The left hemisphere is more severely affected in the rest.

the colorbar on the left. A 45 degree line is also shown. Any departure from this line represents asymmetry. Note that the spread of the data exhibits mirror symmetry around the 45 degree line. That is, given a time series for a subject, its mirror image across the 45 degree line is also a valid time series. This implies that if we were to use a single model to describe all of the trajectories in a population, then the model should remain invariant if we swapped the right and left hemispheres for all subjects. We call this property *population mirror symmetry*.

III. The MLDS Model

Leaving aside the issue of subtypes for now, Fig. 1 suggests a single multivariate linear dynamical system (LDS) as a model for all disease trajectories. Suppose that the mean SBR in the left caudate (LC), left putamen (LP), right putamen (RP), right caudate (RC) are arranged in a vector $\mathbf{x} = [\text{LC, LP, RP, RC}]^T$, then the time evolution of $\mathbf{x}$ can be modeled as the LDS $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a $D \times D$ *transition matrix* ($D = 4$). This model is coupled as long as $\mathbf{A}$ is not a diagonal matrix. The solution of the LDS is $\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0)$, where $e^{\mathbf{A}t}$ is the matrix exponential and $\mathbf{x}(0)$ is the initial condition. The solution has interesting, and well-known, properties:

1) *The semi-group property*: Suppose $\mathbf{x}(t)$ is a time series that satisfies $d\mathbf{x}(t)/dt = \mathbf{A}\mathbf{x}(t)$. Further suppose that we observe this time series starting from some later point in time, i.e. suppose $\mathbf{y}(t) = \mathbf{x}(t+T)$, $T > 0$. Then $\mathbf{y}(t)$ continues to follow the same differential equation without time shift, i.e. $d\mathbf{y}(t)/dt = \mathbf{A}\mathbf{y}(t)$, an equation that is independent of T.
2) For a given $\mathbf{A}$, the initial condition $\mathbf{x}(0)$ determines the entire trajectory. Since different PD patients have different initial conditions, the trajectory determined by the differential equation is automatically subject-specific.

These properties suggest that as far as fitting a matrix $\mathbf{A}$ to a time series goes, it is not essential to know, or to model, the exact starting time for every patient.

Different progression subtypes can be described by different LDSs (with different transition matrices $\mathbf{A}$). This leads to the mixture of linear dynamical systems (MLDS) model. We make two comments about this model before we give mathematical details:

First, a subtype in this model does not correspond to a single speed of progress, neither does the model cluster the time series by progression speeds. The differential equation $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$ expresses a relation between the values of the SBRs (at any point in time) to the rates of change of SBRs at that point in time. The values of SBR and its rate (speed) can be arbitrary; all that the equation requires is that the relation between the two be similar for subjects to be modeled by the same equation.

Second, different subtypes modeled in this way cannot be seen as early or late stages of a single trajectory. Suppose $d\mathbf{x}_1(t) = \mathbf{A}_1\mathbf{x}_1(t)$ and $d\mathbf{x}_2(t)/dt = \mathbf{A}_2\mathbf{x}_2(t)$ with $\mathbf{A}_1 \neq \mathbf{A}_2$. Then, excluding trivial initial conditions similar to $\mathbf{x}_1(0) = \mathbf{x}_2(0) = 0$, there is no time shift $T \neq 0$ (positive or negative) such that $\mathbf{x}_1(t) = \mathbf{x}_2(t+T)$ for all t.

We now give mathematical details of this model beginning with the constraint on $\mathbf{A}$ due to population mirror symmetry.

A. Dynamical System with Population Mirror Symmetry

Given the arrangement of $\mathbf{x}$ that we use (i.e. LC, LP, RP, RC), population mirror symmetry is mathematically equivalent to saying that the differential equation $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$ remains invariant under a permutation that swaps the right SBRs with the left SBRs.

Definition 1. *$\pi : \mathbb{R}^D \to \mathbb{R}^D$ is a* symmetric permutation *if $(\pi(\mathbf{x}))_i = (\mathbf{x})_{D-i+1}$ for $i = 1, \cdots, D$, for any $\mathbf{x} \in \mathbb{R}^D$, where $(\mathbf{x})_k$ refers to the kth component of the vector $\mathbf{x}$.*

Because of the way the SBRs are arranged in the vector $\mathbf{x}$, population mirror symmetry corresponds to applying a symmetric permutation to $\mathbf{x}$. Applying this permutation to both sides of $d\mathbf{x}/dt = \mathbf{A}\mathbf{x}$ gives: $\pi(d\mathbf{x}/dt) = d\pi(\mathbf{x})/dt = \pi(\mathbf{A}\mathbf{x})$. Population mirror symmetry requires $d\pi(\mathbf{x})/dt = \mathbf{A}\pi(\mathbf{x})$, i.e. $\pi(\mathbf{A}\mathbf{x}) = \mathbf{A}\pi(\mathbf{x})$. It is easy to check that this constraint is mathematically equivalent to the pre-processing procedure in the clinical literature that relabels the two hemispheres of

all the subjects to dominant/non-dominant for analysis (hence ignores the left/right difference and assumes a mirror-like progression pattern) [4].

Population mirror symmetry is equivalent to $\mathbf{A}$ being a centrosymmetric matrix:

Definition 2. *A $D \times D$ matrix $\mathbf{A}$ is* centrosymmetric *if $\pi(\mathbf{Ax}) = \mathbf{A}\pi(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^D$, where π is a symmetric permutation [36].*

In terms of elements of $\mathbf{A}$, centrosymmetry means that $(\mathbf{A})_{i,j} = (\mathbf{A})_{D-i+1,D-j+1}$. Loosely speaking, centrosymmetry means that elements which are located on the same line through the center of the matrix, but which are on opposite sides of the center, are equal.

Definition 3. *The dynamical system $d\mathbf{x}/dt = \mathbf{Ax}$ has* population mirror symmetry *if $\mathbf{A}$ is a centrosymmetric matrix.*

From now on we assume that our dynamical system has mirror symmetry. The following properties of centrosymmetric matrices are important for developing and interpreting our model:

1) The set of all $D \times D$ centrosymmetric matrices is a subspace of the vector space of $D \times D$ matrices. This subspace has dimension $\lceil D^2/2 \rceil$.
2) If the eigenvalues of a centrosymmetric matrix are distinct, then the corresponding eigenvectors are either symmetric or anti-symmetric [36].

The first property indicates that the number of parameters used in fitting a centrosymmetric matrix is reduced by half for even numbered D. The second property has implications for interpreting the differential equation $d\mathbf{x}/dt = \mathbf{Ax}$.

To see the significance of interpreting a differential equation with a centrosymmetric transition matrix, note that the solution of the differential equation can be written as

$$\mathbf{x}(t) = \sum_i c_i e^{\lambda_i t} \mathbf{v}_i \tag{1}$$

where $\{\mathbf{v}_i : i = 1, \ldots, D\}$ are the eigenvectors of $\mathbf{A}$ and $\{\lambda_i\}$ are the corresponding eigenvalues, and $[c_1, \ldots, c_D]^T = \mathbf{V}^{-1}\mathbf{x}(0)$ where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_D]$. The eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_D$ are linearly independent but are not guaranteed to be orthonormal. Hence the coefficients of $\mathbf{x}$ along the eigenvectors are found by taking the inner product of $\mathbf{x}$ with the dual basis of $\mathbf{v}_1, \ldots, \mathbf{v}_D$. Suppose $\{\mathbf{u}_1, \ldots, \mathbf{u}_D\}$ is such a dual basis, i.e. $\mathbf{U}^T\mathbf{V} = \mathbf{I}$ where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_D]$. We have an exponential function for projected SBR values along each $\mathbf{u}_i$, $\mathbf{u}_i^T\mathbf{x}(t) = e^{\lambda_i t} c_i$.

The dual basis of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T$ with the same eigenvalues as that of $\mathbf{A}$. Since $\mathbf{A}^T$ is also centrosymmetric, the dual basis vectors are also either symmetric or anti-symmetric. The symmetry/anti-symmetry of the dual basis of the eigenvectors of the transition matrix has an interesting interpretation. If $\mathbf{u}_i$ is symmetric, i.e. $\mathbf{u}_i = [\alpha, \beta, \beta, \alpha]^T$, then the projection $\mathbf{u}_i^T\mathbf{x}(t)$ is the linear combination $\alpha \times$ LC $+ \beta \times$ LP $+ \beta \times$ RP $+ \alpha \times$ RC, i.e. the projection is a symmetric measurement across the brain hemispheres. If $\mathbf{u}_i$ is anti-symmetric, then the projection is an asymmetric measurement. Thus projecting on the dual basis tells us how symmetric and asymmetric parts of the SBR vector evolve – they evolve with the corresponding λ_i as time constants. Note however, that because the dual basis is not orthonormal, the orthogonal projection $\mathbf{u}_i^T\mathbf{x}(t)$ is not the component of $\mathbf{x}(t)$ along $\mathbf{u}_i$. Rather it is the component of $\mathbf{x}(t)$ along $\mathbf{v}_i$.

B. Discretization and Probabilistic Formulation

Suppose that SBRs are available for N subjects and the i^{th} subject has SBRs $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i,T_i}$ at time points $\tau_{i1}, \ldots, \tau_{i,T_i}$, where T_i is the total number of time points and the time points are not assumed to be evenly spaced. Then, the time series for the i^{th} subject can be modeled by a discrete version of the linear differential equation $d\mathbf{x}/dt = \mathbf{Ax}$ as:

$$\frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{ij}}{\Delta t_{ij}} = \mathbf{A}\mathbf{x}_{ij} + \boldsymbol{\epsilon}_{ij}, \tag{2}$$

where $\Delta t_{ij} = \tau_{i,j+1} - \tau_{ij}$, and $\boldsymbol{\epsilon}_{ij}$ is the model residue. The residue is assumed to follow a Student's *t*-distribution, i.e. $\boldsymbol{\epsilon}_{ij} \sim \mathcal{T}\left(\mathbf{0}, \sigma^2\mathbf{I}_D, \nu\right)$ where $\sigma^2\mathbf{I}_D$ is the scale matrix and ν is the degree of freedom.

Letting $\mathbf{x}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i,T_i}\}$ denote the entire time series for subject i,

$$p\left(\mathbf{x}_i|\mathbf{A}, \sigma^2, \nu\right) = p\left(\mathbf{x}_{i1}\right) \prod_{j=1}^{T_i-1} p\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij}, \mathbf{A}, \sigma^2, \nu\right), \tag{3}$$

where we assume that the probability distribution of the first element of the time series is $p(\mathbf{x}_{i1}) = \mathcal{N}(\mathbf{x}_{i1}|\mathbf{0}, \Sigma)$, and the conditional probability distribution is

$$\begin{aligned} &p\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij}, \mathbf{A}, \sigma^2, \nu\right) \\ =&\mathcal{T}\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij} + \Delta t_{ij}\mathbf{A}\mathbf{x}_{ij}, \Delta t_{ij}^2\sigma^2\mathbf{I}_D, \nu\right). \end{aligned} \tag{4}$$

The form of the conditional distribution follows from (2) and the *t*-distribution. The distribution $p(\mathbf{x}_{i1}) = \mathcal{N}(\mathbf{x}_{i1}|\mathbf{0}, \mathbf{\Sigma})$ is the same for every subject. It models the "spread" of initial data $\mathbf{x}_{i1}$, which we take to be independent of $\mathbf{A}, \sigma^2, \nu$.

Directly expressing $p\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij}, \mathbf{A}, \sigma^2, \nu\right)$ as the *t*-distribution causes technical problems in Bayesian inference – there are no conjugate priors for $\mathbf{A}, \sigma^2, \nu$. However, a standard modification makes it possible to create conjugate priors for $\mathbf{A}, \sigma^2, \nu$ [37], [38]. The modification follows from the observation that a *t*-distributed random variable can be generated by first sampling a scalar random variable from a Gamma distribution, and then sampling from a Gaussian distribution with a covariance matrix scaled by the Gamma distribution sample, i.e.

$$\mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}, \nu) = \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}/w)\, \mathrm{Ga}\left(w|\frac{\nu}{2}, \frac{\nu}{2}\right) dw, \tag{5}$$

where w is the scale parameter. To use this formulation, we introduce latent scale variables $\{w_{ij} : i = 1, \ldots, N, j = 1, \ldots, T_i - 1\}$ with $p(w_{ij}|\nu) = \mathrm{Ga}\left(w_{ij}|\frac{\nu}{2}, \frac{\nu}{2}\right)$, and write (4) with a normal distribution on the right hand side:

$$\begin{aligned} &p\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij}, w_{ij}, \mathbf{A}, \sigma^2\right) \\ =&\mathcal{N}\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij} + \Delta t_{ij}\mathbf{A}\mathbf{x}_{ij}, \Delta t_{ij}^2\sigma^2\mathbf{I}_D/w_{ij}\right). \end{aligned} \tag{6}$$

With this modification, we can rewrite (3) as

$$p\left(\mathbf{x}_i|\mathbf{w}_i,\mathbf{A},\sigma^2\right)=p(\mathbf{x}_{i1})\prod_{j=1}^{T_i-1}p\left(\mathbf{x}_{i,j+1}|\mathbf{x}_{ij},w_{ij},\mathbf{A},\sigma^2\right), \quad (7)$$

where $\mathbf{w}_i=(w_{i1},\ldots,w_{i,T_i-1})$ and $p\left(\mathbf{w}_i|\nu\right)=\prod_j p\left(w_{ij}|\nu\right)$. Note that according to (5), combining $p\left(\mathbf{x}_i|\mathbf{w}_i,\mathbf{A},\sigma^2\right)$ and $p\left(\mathbf{w}_i|\nu\right)$ with $\mathbf{w}_i$ integrated out gives the time series distribution in (3). This is the discretized probabilistic version of $d\mathbf{x}/dt=\mathbf{A}\mathbf{x}$ with *t*-distributed model residues.

C. The Mixture Model

All subjects that have the same transition matrix belong to the same subtype. To extend the model to K distinct progression subtypes, we allow each subtype to have its own transition matrix $\mathbf{A}_k$ and model residue σ_k. Let z_i be a latent random variable taking values in $\{1,2,\ldots,K\}$ and indicating the subtype of the i^{th} subject. Given z_i, the probability density of the time series $\mathbf{x}_i$ of the i^{th} subject is the density of (7) with $\mathbf{A}_{z_i}$ and σ_{z_i}:

$$p\left(\mathbf{x}_i|z_i,\mathbf{w}_i,\left\{\mathbf{A}_k,\sigma_k^2\right\}\right)=p\left(\mathbf{x}_i|\mathbf{w}_i,\mathbf{A}_{z_i},\sigma_{z_i}^2\right). \quad (8)$$

The latent variable z_i has a categorical distribution:

$$p\left(z_i|\boldsymbol{\pi}\right)=\text{Cat}\left(z_i|\boldsymbol{\pi}\right)=\prod_{k=1}^{K}\pi_k^{\mathbb{I}(z_i=k)}=\pi_{z_i}, \quad (9)$$

where $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_K)$, such that $\pi_l\geq 0$ for all l and $\sum_l\pi_l=1$. Also $\mathbb{I}\left(\cdot\right)=1$ if the argument of $\mathbb{I}$ is true and zero otherwise.

Finally, let $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, $\mathbf{W}=\{\mathbf{w}_1,\ldots,\mathbf{w}_N\}$, and $\mathbf{z}=(z_1,\ldots,z_N)$ denote the time series, the latent variables for the *t*-distribution, and the latent variables for the class labels. Setting $\boldsymbol{\theta}=\left\{\boldsymbol{\pi},\mathbf{A}_l,\sigma_l^2:l=1,\ldots,K\right\}$ gives

$$p\left(\mathbf{X}|\mathbf{z},\mathbf{W},\boldsymbol{\theta}\right)=\prod_{i=1}^{N}p\left(\mathbf{x}_i|z_i,\mathbf{w}_i,\left\{\mathbf{A}_l,\sigma_l^2\right\}\right), \quad (10)$$

$$p\left(\mathbf{W}|\nu\right)=\prod_{i=1}^{N}p\left(\mathbf{w}_i|\nu\right),\ p\left(\mathbf{z}|\boldsymbol{\theta}\right)=\prod_{i=1}^{N}p\left(z_i|\boldsymbol{\pi}\right) \quad (11)$$

as the complete model. Note that integrating out the latent variables $\mathbf{W}$ and $\mathbf{z}$ gives the mixture model

$$p\left(\mathbf{X}|\boldsymbol{\theta},\nu\right)=\prod_{i=1}^{N}p\left(\mathbf{x}_i|\boldsymbol{\theta},\nu\right)=\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_k p\left(\mathbf{x}_i|\mathbf{A}_k,\sigma_k^2,\nu\right).$$

where $p\left(\mathbf{x}_i|\mathbf{A}_k,\sigma_k^2,\nu\right)$ is defined in (3). We want to infer the parameters $\boldsymbol{\theta}$ and ν from the observed data $\mathbf{X}$.

IV. Bayesian Inference

Our Bayesian inference methodology is to use Gibbs sampling, and to keep the sampling scheme tractable we use priors that are conjugate to the conditional densities in (10) and (11):

$$p\left(\boldsymbol{\pi}|\boldsymbol{\alpha}\right)=\text{Dir}\left(\boldsymbol{\pi}|\boldsymbol{\alpha}\right), \quad (12)$$

$$p\left(\nu|\boldsymbol{\gamma}\right)\propto\left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{\xi_0}e^{\tau_0\nu},\ \nu>0, \quad (13)$$

$$\begin{aligned}p\left(\mathbf{a}_k,\sigma_k^2|\boldsymbol{\beta}\right)&=\text{NIG}\left(\mathbf{a}_k,\sigma_k^2|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0,\nu_0,\kappa_0\right)\\&=\mathcal{N}\left(\mathbf{a}_k|\boldsymbol{\mu}_0,\sigma_k^2\boldsymbol{\Lambda}_0^{-1}\right)\text{IG}\left(\sigma_k^2|\nu_0,\kappa_0\right), \quad (14)\end{aligned}$$

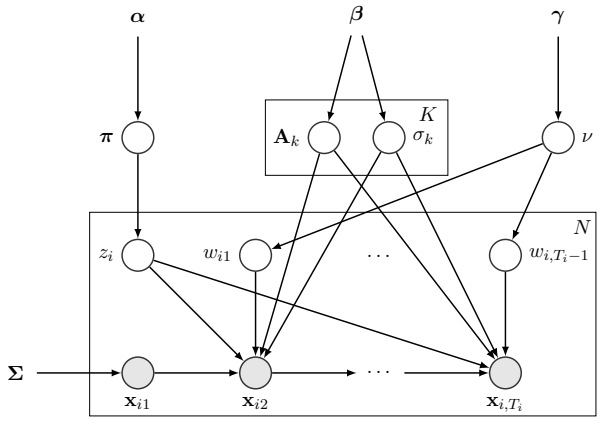


Fig. 2. Probabilistic graphical model of the MLDS with *t*-distributed residues. It features a standard structure of four layers: hyperparameters ($\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$), parameters to infer ($\boldsymbol{\theta}=\left\{\boldsymbol{\pi},\mathbf{A}_k,\sigma_k^2\right\},\nu$), latent variables ($\mathbf{z}=(z_1,\ldots,z_N)$, $\mathbf{W}=\{\mathbf{w}_1,\ldots,\mathbf{w}_N\}$), and observed data ($\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$).

where $\text{Dir}\left(\cdot\right)$ is the Dirichlet distribution, $\text{NIG}\left(\cdot\right)$ is the normal-inverse-gamma distribution, $\text{IG}\left(x|a,b\right)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$, and $\boldsymbol{\alpha}=(\alpha/K,\ldots,\alpha/K)$, $\boldsymbol{\gamma}=\{\xi_0,\tau_0\}$, $\boldsymbol{\beta}=\{\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0,\nu_0,\kappa_0\}$ are hyperparameters. We define the prior on $\mathbf{A}_k$ using its coordinates on a basis for centrosymmetric matrices, i.e. $\text{vec}\left(\mathbf{A}_k\right)=\mathbf{E}\mathbf{a}_k$ where the j^{th} column of $\mathbf{E}$ has the form $[\ldots,1,\ldots,1,\ldots]^T$ where 1 only appears at the j^{th} position and the $(D^2-j+1)^{\text{th}}$ position and the others are zero. The rationale for choosing (14) for $p\left(\mathbf{X}|\mathbf{z},\mathbf{W},\boldsymbol{\theta}\right)$ and (13) for $p\left(\mathbf{W}|\nu\right)$ is provided in Supplementary Section II.

With these priors, the probabilistic graphical model is shown in Fig. 2. We can infer the parameters by drawing samples from the posterior $p\left(\mathbf{z},\boldsymbol{\theta},\mathbf{W},\nu|\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}\right)$.

A. The Gibbs Sampler

A detailed derivation of the Gibbs sampler is available in Supplementary Section III. Here we briefly describe the salient points of the sampler. The sampler works by sampling $\mathbf{z},\theta$ and $\mathbf{W},\nu$ in sequence conditioned on the remaining random variables. The sampling proceeds as below:

1. Sample $p(\mathbf{z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$ as follows:
   - 1.1. Sample $\mathbf{z}$ from $p(\mathbf{z}|\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$ with $\boldsymbol{\theta}$ integrated out. This is known as *collapsed Gibbs sampling* [39]. Sampling $\mathbf{z}$ corresponds to sequentially sampling each z_i given the rest $\{z_j:j\neq i\}$.
   - 1.2. Sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{z},\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$ by sampling $\boldsymbol{\pi}$ from a Dirichlet distribution and $\mathbf{A}_k,\sigma_k^2$ from a NIG distribution.
2. Sample $p(\mathbf{W},\nu|\mathbf{X},\boldsymbol{\gamma},\mathbf{z},\boldsymbol{\theta})$ as follows:
   - 2.1. Sample $\mathbf{W}$ from $p(w_{ij}|\mathbf{X},\boldsymbol{\gamma},\mathbf{z},\boldsymbol{\theta},\nu)$ by independently sampling each w_{ij} from a Gamma distribution.
   - 2.2. Sample ν from $p(\nu|\mathbf{W},\boldsymbol{\gamma})$ using adaptive rejection sampling (ARS) [40].

A detailed version of this algorithm is in Supplementary Section III-B. The above steps are iterated till the chain converges and provides sufficient samples for parameter estimation.

We set the hyperparameters to $\alpha = K$, $\boldsymbol{\beta} = \left(\mathbf{0}, 10^{-8}\mathbf{I}_{\lceil D^2/2 \rceil}, 10^{-3}, 10^{-3}\right)$, $\boldsymbol{\gamma} = \left(10^{-3}, 10^{-3}\right)$, which corresponds to having weak priors. The initialization of each sampling chain is done by assigning each $z_i \in \{1, \ldots, K\}$ randomly, $w_{ij} = 1, \forall i, j$, $\nu = 30$. Following Section 11.4 of [41], we run 5 chains (1500 iterations each) with random initialization, discard the first half of each chain as burn-in samples, and split the remaining samples to calculate a ratio of between-sequence variance and within-sequence variance of $\log p\left(\mathbf{X}|\boldsymbol{\theta}, \nu\right)$ to check convergence. For any combination of chains, if this ratio is close to one, we conclude that this combination converges to the same distribution. We pick samples from one chain of a converged combination for analysis.

Following standard Bayesian methodology, we take the parameter estimates to be the averages of the post-burn-in samples of the converged chain. The estimates of the transition matrices $\mathbf{A}_k$ are of particular importance, since they define the progression subtypes. We estimate $\mathbf{A}_k$ by the mixture estimator [42]:

$$\hat{\mathbf{A}}_k = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\left(\mathbf{A}_k | \mathbf{z} = \mathbf{z}_l, \mathbf{W} = \mathbf{W}_{l-1}, \nu = \nu_{l-1}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right)$$

where L is the number of post-burn-in samples. The means on the right hand side are directly available when we sample $\mathbf{A}_k, \sigma_k^2$. The above estimator has a lower variance than the empirical estimator (i.e. averaging the samples) according to the Rao-Blackwell theorem (see Section 2.4.4 in [43]).

B. Model Selection

The Gibbs sampler described above estimates model parameters, once the number of subtypes (the number of components of the model) are known. To find the number of subtypes, we use cross validation and Bayesian model selection [44], [45]. For cross validation, we divide the dataset into 10 subsets (10-fold cross-validation). Using each subset as test set, we use the remaining data as training set to infer the parameters $\boldsymbol{\theta}, \nu$. Then, the log-likelihood of each test set is evaluated and the sum of these log-likelihood values is considered for each $K \in \{1, \ldots, K_{\text{max}}\}$, where K_{max} is the maximum number of components considered.

For Bayesian model selection, we denote $\boldsymbol{\eta}_K = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ for the hyperparameters with K components, and let $\mathcal{H} = \left\{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K_{\text{max}}}\right\}$. Assuming $p\left(K\right) \propto$ constant on $K = 1, \ldots, K_{\text{max}}$, we have

$$p\left(K|\mathbf{X}, \mathcal{H}\right) \propto p\left(K\right) p\left(\mathbf{X}|\mathcal{H}, K\right) \propto p\left(\mathbf{X}|\mathcal{H}, K\right) = p\left(\mathbf{X}|\boldsymbol{\eta}_K\right).$$

Finding the optimal $\hat{K} = \arg\max_K p\left(K|\mathbf{X}, \mathcal{H}\right)$ is equivalent to finding the maximum of $p\left(\mathbf{X}|\boldsymbol{\eta}_K\right)$ which can be evaluated by the integral

$$p\left(\mathbf{X}|\boldsymbol{\eta}_K\right) = \int p\left(\mathbf{X}|\boldsymbol{\theta}, \nu\right) p\left(\boldsymbol{\theta}, \nu|\boldsymbol{\eta}_K\right) d\boldsymbol{\theta} d\nu. \qquad (15)$$

Since the Gibbs sampler has already generated samples from $p\left(\mathbf{z}, \boldsymbol{\theta}, \mathbf{W}, \nu|\mathbf{X}, \boldsymbol{\eta}_K\right)$, we use importance sampling with the proposal distribution $p\left(\boldsymbol{\theta}, \nu|\mathbf{X}, \boldsymbol{\eta}_K\right)$ to calculate the integral. The details are in Supplementary Section IV.

V. Results

A. Data Preparation

The PPMI DaTscan dataset was described earlier in Section II-D. The DICOM headers for PPMI DaTscan images reveal that the images have a size of $109 \times 91 \times 91$ voxels, with 2 mm^3 isotropic voxels. The images are distributed by PPMI and already registered in standard Montreal Neurological Institute (MNI) space. However, we did find some misregistered images in the data. These were eliminated in the preprocessing step described below. After elimination, the mean SBRs in the caudates and the putamina were obtained using the MNI atlas, which is also explained below.

Image Pre-processing: We pre-processed the image data in two steps. First, we eliminated all subjects that had only one scan, since time series information cannot be gleaned from a single scan. This led to 382 remaining subjects. Next, we eliminated all subjects that had misregistered images. Misregistered images were found by taking the image sequence for every subject and calculating the correlation coefficient of all voxels outside the striatum between every pair of images in the series. The smallest correlation coefficient in this set was taken as the indicator of misregistration. If this indicator was less than the median minus three times the mean absolute deviation of correlation coefficients of all subjects, then the entire sequence for the subject was removed. This step eliminated 17 such subjects, leaving 365 subjects, which entered the analysis (id's of the eliminated subjects are available in the supplementary material). Of these subjects, 45 had 2 scans, 190 had 3 scans, 127 had 4 scans and 3 had 5 scans.

Next the SBR feature vectors $\mathbf{x}_{ij}$ were extracted from the images by using a set of 3D masks for the two caudates, two putamina, and the occipital lobe. Fig. 3 shows the masks overlaid on a subset of the axial slices of the mean baseline image of all subjects. The masks for the caudates and putamina were taken from the MNI atlas, dilated by 1 voxel and smoothed by a Gaussian filter with $\sigma = 0.5$ pixels to capture the partial volume effects. The occipital lobe mask was created manually, and is similar to the mask in [29]. Then, the median of the occipital lobe was used as the denominator to calculate the SBR at each voxel, and mean SBRs in the caudates and putamina were organized as $\mathbf{x}_{ij}$ as described in Section III-A. PPMI also provides imaging dates for all subjects, and these dates were used to calculate the time intervals Δt_{ij}.

B. Simulation on Synthetic Dataset

Because Bayesian analysis is new to PD DaTscan image analysis, we first evaluated its accuracy – especially clustering accuracy – by creating a synthetic dataset with known class labels. To create a synthetic dataset that is close to real DaTscan data, we used our algorithm on the real PPMI dataset with 3 clusters (see Section V-C) to obtain estimates of the model parameters ($\{\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}_k, \hat{\sigma}_k^2\}$, $\hat{\nu}$). Using these estimated parameters, we created a low, medium and large noise dataset (the noise σ_k's were set to $0.1^{\lambda}\hat{\sigma}_k$ with $\lambda = 2, 1, 0$) by keeping existing $\{\mathbf{x}_{i1}\}$ and $\{\Delta t_{ij}\}$ and generating the remaining data according to (4). We also set $\pi_k = 1/K$ to ensure that the data are evenly distributed across different classes.

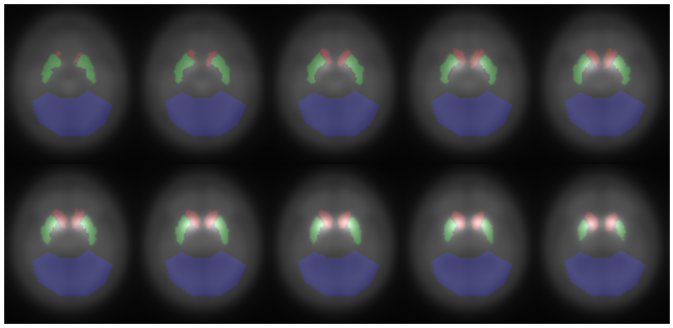


Fig. 3. Masks for the left and right caudates (red), left and right putamina (green), and the occipital lobe (blue). The background shows the 33rd - 42nd slices of the mean baseline image.

TABLE I
CLUSTERING AND PREDICTION ACCURACIES OF GIBBS SAMPLING VS. THE EM-ALGORITHM USING A SYNTHETIC DATASET.

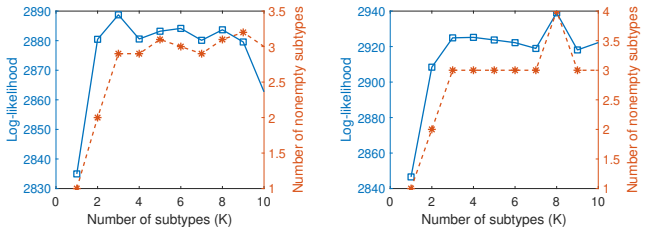
Noise Level	$0.01\hat{\sigma}_k$			$0.1\hat{\sigma}_k$			$\hat{\sigma}_k$		
	Purity	Rd Id	Pr Er	Purity	Rd Id	Pr Er	Purity	Rd Id	Pr Er
Gibbs-*t* [b]	**1.000**[a]	**1.000**	**0.004**	**0.997**	**0.996**	**0.037**	**0.853**	**0.644**	**0.393**
Gibbs	**1.000**	**1.000**	**0.004**	**0.997**	**0.996**	**0.037**	0.781	0.642	0.401
EM [c]	0.988	0.938	0.022	0.988	0.957	0.045	0.758	0.629	0.408

[a] The best entry in each category is boldfaced.
[b] Gibbs-*t* / Gibbs is the proposed Gibbs sampling with *t*-distributed / normal-distributed model residues.
[c] EM is the version that maximizes the likelihood without the centrosymmetric constraint.

We divided this synthetic dataset randomly into 10 subsets where 9 subsets were retained for training and 1 subset for testing (rotated over all subsets). This procedure was repeated 10 times and Gibbs sampling was run on the 10 by 10 training sets. For comparison, we also ran two other algorithms: 1. An EM algorithm that maximizes the log-likelihood of $p(\mathbf{X}|\boldsymbol{\theta}, \nu)$ of Section III-C, but with Gaussian noise. 2. A simplified Gibbs sampling with Gaussian noise and the centrosymmetric constraint [7].

We used three measures to compare clustering accuracies of the algorithms. The first measure is *purity*, which measures the percentage of overlap of estimated and true class labels. The second measure is *Rand index*, which measures the proportion of data point pairs that are in agreement with the true labels in terms of falling in the same class or different classes [46]. Purity and Rand index have range 0 - 1, where 1 represents perfect clustering. The third measure is *prediction error*. We use the MLDS model to predict the SBR values for time points ≥ 3 and take the prediction error to be the difference (L1 norm) between the prediction and true values (details in Supplementary Section V).

Table I shows the mean purity and mean Rand index over the 10 by 10 training sets, and mean prediction error over the 10 by 10 test sets, for the three methods. We see that the Gibbs sampling algorithms outperform the EM algorithm for all performance measures. The two Gibbs samplers perform very similarly except for the large noise case, where the *t*-distribution version has a higher purity and Rand index and lower prediction error. This analysis of synthetic data justifies the use of Bayesian analysis with *t*-distributions over maximum-likelihood methods.



(a) Cross Validation (b) Bayesian Model Selection

Fig. 4. Model selection using cross validation (a) and Bayesian (b). The y-axis has two scales corresponding to log-likelihood value (blue solid square) and final number of nonempty subtypes (orange dashed star) respectively. The number of nonempty subtypes is averaged over 10 folds for cross validation.

C. Fitting MLDS to the PPMI Data

Having established the superiority of Gibbs sampling with synthetic data, we turn to analyzing real PPMI data. We first determined the number of subtypes using cross-validation and Bayesian model selection as described in Section IV-B, and then used Bayesian analysis to explore the posterior distribution of the parameters.

1) Determining the number of subtypes: The results of using cross-validation and Bayesian model selection are shown in Fig. 4. The number of subtypes explored was between 1 and 10 ($K_{\max} = 10$). The blue solid curve in Fig. 4(a) shows the log-likelihood of the test sets as a function of the number of subtypes. As the number of subtypes increases, several subtypes turn out to be empty, i.e. no subjects are assigned to that subtype. The orange dashed curve in Fig. 4(a) shows the mean number of nonempty subtypes. Fig. 4(b) shows the same two quantities for Bayesian model selection.

Fig. 4 clearly shows that the log-likelihood values for cross-validation and Bayesian model selection behave similarly. The log-likelihood increases monotonically from 1 to 3 subtypes and then appears to saturate. The number of nonempty subtypes found by both methods is similar as well. In the final model, we chose 3 subtypes ($K = 3$) for further analysis.

2) Parameter estimation and interpretation: The MLDS model with three subtypes ($K = 3$) was fit to the PPMI dataset using Bayesian analysis. The clustering results are shown in Fig. 5. The subtype label is created by calculating $p(z_i|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ from the samples, and assigning $\arg\max_k p(z_i = k|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ to subject i. This gives us 46, 257, and 62 subjects in subtypes 1, 2, and 3 respectively. As the estimated trajectories show, different subtypes progress with different speeds with subtype 1 being the fastest and subtype 3 being the slowest. The mean and standard deviation of the posterior distribution of the parameters are shown in the top half of Table II for each subtype. The bottom half of Table II (rows indicated by $\lambda, \mathbf{v}, \mathbf{u}$) shows the eigenvalues, eigenvectors and dual basis of the eigenvectors of the mean transition matrix for each subtype.

The main characteristics of Table II are: Row π_k (mixing coefficient) in Table II indicates that subtype 2 has the highest occupancy; a little over half of the subjects are contained in this subtype. Subtype 3 and 1 have sequentially smaller occupancy. All subtypes have similar values for σ_k, suggesting

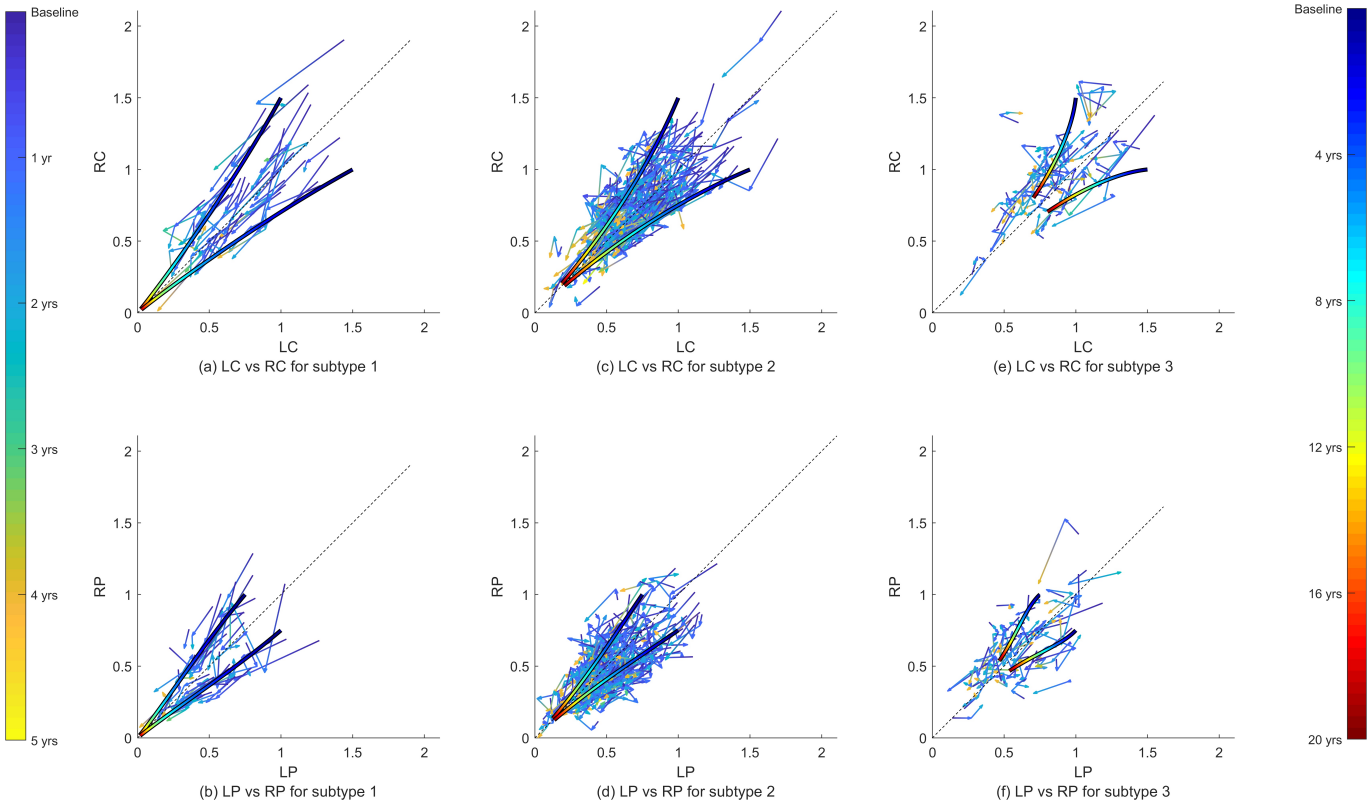


Fig. 5. Clustered time series of the mean SBR for all the PPMI subjects. The columns show the subtypes discovered by the proposed approach. The arrows are rendered with continuously changing colors that correspond to the 5 year period indicated by the colorbar on the left side. We can see that different subtypes exhibit different progression rates. We also show our estimated trajectory starting from a fixed point ($\mathbf{x} = [1, 0.75, 1, 1.5]^T$) and its reflection (duration indicated by the colorbar on the right side).

that all subtypes have similar model residues.

Different subtypes have different progression rates and progression trajectories, i.e. the MLDS model has successfully captured PD heterogeneity. The variability in progression rates is apparent in the eigenvalues of the transition matrices in Table II (row λ). All eigenvalues are real, distinct, and negative. Checking the magnitude of the eigenvalues, we see that subtype 1 is the fastest progressing subtype, followed by subtype 2 and then by subtype 3. Further evidence for the relative speeds of the subtypes can be directly found in Fig. 6, which shows histograms of starting speeds (i.e. the magnitude of $\frac{\mathbf{x}_{i2}-\mathbf{x}_{i1}}{\Delta t_{i1}}$) of all subjects in each subtype (initial changes are the largest and therefore present the clearest evidence in presence of noise).

The subtypes differ not only in speed but also in the shape of the SBR trajectories as well. This is apparent in Fig. 5 which shows the SBR trajectories of subjects in each subtype. The figure also shows model trajectories (smooth curves overlaid on raw trajectories) for two initial points. These trajectories clearly have different speed, extent, and shape.

The spatial patterns of progression as evident in the dual basis of the eigenvectors of the transition matrices are especially interesting. Recall from Section III-A that a symmetric or anti-symmetric dual basis vector can be interpreted as representing the symmetry or asymmetry of the disease across the two brain hemispheres. Since all eigenvalues are real and negative, symmetric/anti-symmetric dual basis vectors capture how the

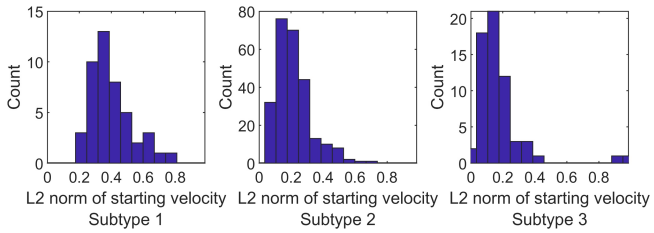


Fig. 6. Histograms of the initial speeds for subtypes 1, 2, and 3. The median velocities are 0.36, 0.19, and 0.13 SBR/year respectively.

symmetry/asymmetry of dopamine transporter concentration (i.e. the mean/difference of $\alpha \times$Caudate$+\beta \times$Putamen between both hemispheres) changes as the disease progresses.

The leading dual basis vector in every subtype in Table II (row $\mathbf{v}$) is anti-symmetric, with α, β having opposite signs. This implies that the loss of asymmetry in the disease is the fastest spatial progression pattern among all linear combinations of SBRs. The last dual basis vector in subtype 1 and 2 is symmetric with α, β having the same sign. This dynamical mode clearly represents the "mean" of all four regions. Since this mode has the smallest eigenvalue, the mean SBR is the slowest index of disease progression in early-stage PD. Thus Table II suggests that the rate of asymmetry change is several times faster than the change in the mean SBR.

3) Relation to Demographics: Fig. 7 shows violin plots of the age and sex distribution of PD subjects in the three

TABLE II
PARAMETERS AND EIGENSTRUCTURE ESTIMATED BY BAYESIAN INFERENCE.

Estimated Parameters (mean (std))[a]												
Subtype	$k=1$				$k=2$				$k=3$			
ν	2.85 (0.21)											
π_k	0.146 (0.036)				0.612 (0.081)				0.242 (0.076)			
σ_k	0.064 (0.007)				0.059 (0.003)				0.067 (0.005)			
$\mathbf{A}_k$	−0.33 (0.06)	0.16 (0.07)	−0.10 (0.06)	0.09 (0.05)	−0.21 (0.02)	0.12 (0.03)	−0.01 (0.03)	0.04 (0.02)	−0.09 (0.04)	0.05 (0.05)	−0.08 (0.05)	0.08 (0.04)
	0.15 (0.06)	−0.44 (0.08)	0.09 (0.06)	−0.05 (0.05)	0.06 (0.02)	−0.22 (0.03)	0.07 (0.02)	−0.02 (0.02)	0.13 (0.03)	−0.24 (0.05)	0.03 (0.05)	−0.01 (0.04)
	−0.05 (0.05)	0.09 (0.06)	−0.44 (0.08)	0.15 (0.06)	−0.02 (0.02)	0.07 (0.02)	−0.22 (0.03)	0.06 (0.02)	−0.01 (0.04)	0.03 (0.05)	−0.24 (0.05)	0.13 (0.03)
	0.09 (0.05)	−0.10 (0.06)	0.16 (0.07)	−0.33 (0.06)	0.04 (0.02)	−0.01 (0.03)	0.12 (0.03)	−0.21 (0.02)	0.08 (0.04)	−0.08 (0.05)	0.05 (0.05)	−0.09 (0.04)
Estimated Eigenstructure[b]												
λ	−0.71	−0.39	−0.23	−0.20	−0.37	−0.22	−0.16	−0.09	−0.37	−0.18	−0.08	−0.02
$\mathbf{v}$ LC	0.46	0.26	0.58	0.59	0.50	0.64	0.60	0.59	0.40	0.13	0.58	0.59
$\mathbf{v}$ LP	−0.54	−0.66	0.40	0.39	−0.50	−0.30	0.38	0.39	−0.58	0.70	0.41	0.39
$\mathbf{v}$ RP	0.54	−0.66	−0.40	0.39	0.50	−0.30	−0.38	0.39	0.58	0.70	−0.41	0.39
$\mathbf{v}$ RC	−0.46	0.26	−0.58	0.59	−0.50	0.64	−0.60	0.59	−0.40	0.13	−0.58	0.59
$\mathbf{u}$ LC	0.41	0.40	0.54	0.67	0.39	0.46	0.51	0.35	0.41	−0.54	0.58	0.97
$\mathbf{u}$ LP	−0.58	−0.60	0.46	0.27	−0.61	−0.70	0.51	0.75	−0.58	0.82	0.40	−0.18
$\mathbf{u}$ RP	0.58	−0.60	−0.46	0.27	0.61	−0.70	−0.51	0.75	0.58	0.82	−0.40	−0.18
$\mathbf{u}$ RC	−0.41	0.40	−0.54	0.67	−0.39	0.46	−0.51	0.35	−0.41	−0.54	−0.58	0.97

[a] π is the fraction of the PD subjects that are contained in each subtype. $\mathbf{A}_k$ and σ_k^2 are the transition matrix and the unscaled variance for each subtype.
[b] The bottom half shows the eigenvalues (λ), the eigenvectors ($\mathbf{v}$) and the dual basis of the eigenvectors ($\mathbf{u}$) of the mean transition matrices.

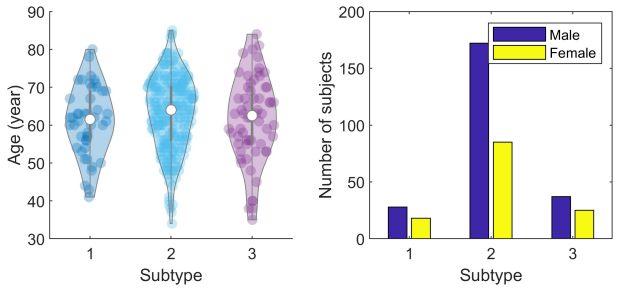


Fig. 7. Age and sex distribution in image-based subtypes.

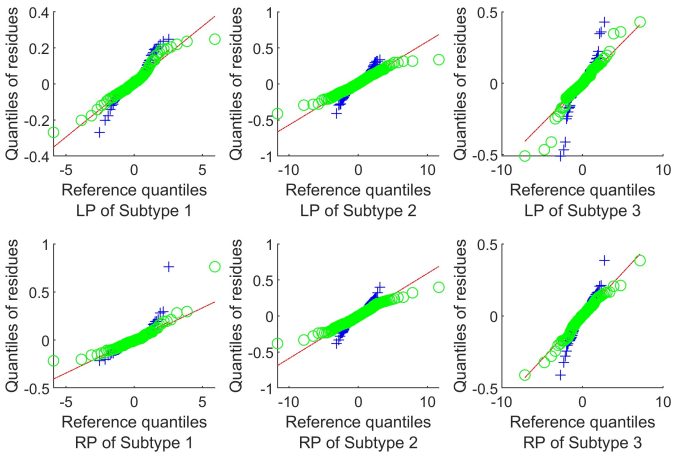


Fig. 8. Q-Q plots of the model residue vs normal (blue cross) and vs t-distribution (green circle) for the putamina.

subtypes. The 95-percentile age range in the three subtypes is 61.0 ± 18.1, 63.2 ± 19.3, 61.5 ± 22.2 years respectively. A t-test with a null hypothesis of equality of means of the ages of subtype 1 vs 2, 2 vs 3, 1 vs 3 gives p-values of 0.14, 0.26, 0.82. Thus the null hypothesis cannot be rejected, suggesting that the mean ages in the subtypes are equal. The male population is distributed in the three subtypes as 11.8%, 72.6%, 15.6%. The female population is distributed as 14.1%, 66.4%, 19.5%. A chi-square test, evaluating the null hypothesis that these distributions are equal, gives a p-value of 0.46, suggesting that the null hypothesis cannot be rejected again. In spite of that, the female population is slightly more dispersed with subtypes 1 and 3 having larger fractions of the population.

D. Model Validation and Results Sensitivity

Finally, we turn to evaluating other aspects of the model: the use of t-distributions for model residues, the train-test consistency of the model residue, and the sensitivity of the result to caudate and putamen templates.

1) Validating the residue distribution: We validate the model residue distribution by using a Q-Q plot, i.e. by plotting the quantiles of the model residues against that of a normal distribution and a t-distribution for each region and each subtype. Fig. 8 shows Q-Q plots of the putamen residues (the caudate residues show a similar trend and are omitted to save space) and fitted lines representing a perfect fit to a residue model. A Q-Q plot crossing the line at a steeper slope implies that the data have heavier tails than the assumed distribution. It is clear from Fig. 8 that the residues in all three subtypes have significantly heavier tails than the normal distribution. And the t-distribution assumption appears to be a significant improvement over the normal distribution assumption in every subtype.

2) Train-test consistency of the model residue: To test whether our algorithm overfits the data, we also evaluated the train-test consistency of the model residue with 10-fold cross validation. Specifically, nine folds were taken as the training set with $K = 3$ to estimate parameters $\hat{\boldsymbol{\theta}}, \hat{\nu}$. These parameters were applied to the subjects in the remaining fold (test set) to predict their subtypes (via $\arg\max_k p\left(z = k|\mathbf{x}, \hat{\boldsymbol{\theta}}, \hat{\nu}\right)$ using Eq. (S20) in Supplementary Section V) and the norms of the model residues $\|\boldsymbol{\epsilon}_{ij}\|$ were calculated from the subtype $\hat{\mathbf{A}}_k$

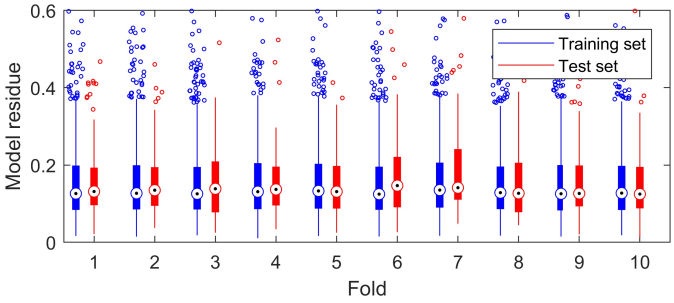
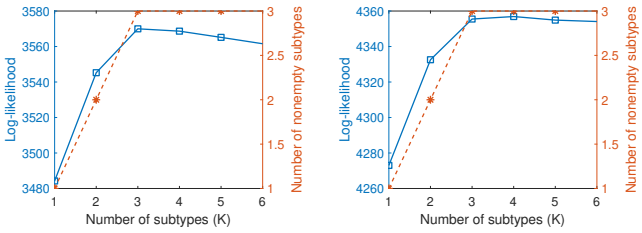


Fig. 9. Box plots of model residues from training sets and test sets in the 10-fold cross validation. The whiskers correspond to $\pm 2.7\sigma$ and 99.3% coverage if the residues are normally distributed. Residues beyond the limit of y-axis are not shown.



(a) Masks dilated by 1 voxel (b) Masks dilated by 2 voxels

Fig. 10. Bayesian model selection for dilated masks. The y-axis has two scales corresponding to log-likelihood value (blue solid square) and final number of nonempty subtypes (orange dashed star) respectively.

according to Eq. (2).

Fig. 9 shows box plots of training and test set residues. For clarity, only the scatter of the residues in the extreme quantiles is shown. Fig. 9 shows that our algorithm does not overfit the data (overfitting would lead to substantially larger test errors).

3) Sensitivity to caudate and putamen masks: Partial voluming and small variations in subject-specific anatomy can potentially affect the estimated model parameters. To test the sensitivity to these, we further dilated the caudate and putamen masks by 1 and 2 voxels and then filtered them with a Gaussian filter having $\sigma = 0.5$ pixels. We re-estimated the parameters using Bayesian analysis with these dilated masks. Dilating by 1 voxel increases the volume of the mask by 55% for caudate and 41% for putamen (133% and 101% when dilating 2 voxels). Even with such large changes to the masks, the number of subtypes (see Fig. 10) as well as the parameter estimates were similar to the original estimates. The number of subtypes remained at 3, and the relative changes in the transition matrices of the three subtypes, calculated as $\|\hat{\mathbf{A}}_k - \hat{\mathbf{A}}_k^{\text{dilated}}\|_F / \|\hat{\mathbf{A}}_k\|_F$, were 0.04, 0.08, 0.11 for masks dilated by 1 voxel and 0.47, 0.19, 1.67 for masks dilated by 2 voxels. We also calculated the Rand index between the subtype labels in Section V-C and the labels estimated using the dilated masks. For masks dilated by 1 voxel, the Rand index was 0.92, while for masks dilated by 2 voxels, the Rand index was 0.73. This implies that the progression pattern and subtyping are not sensitive to partial volume effects or anatomical variations.

E. Correlation with MDS-UPDRS scores

Finally, we sought correlation between DaTscan-based progression subtypes and clinical movement scores as present in the Part III of the MDS-UPDRS exam. In PPMI, longitudinal scores of each patient are sampled more frequently than DaTscan images: at 3 months intervals for the first year, at 6 months intervals for the next 4 years, and at 1 year intervals for the following 3 years. We retained only those scores that corresponded to the imaging times. Part III scores can be influenced by medication, but PPMI provides scores for subjects in the off-medication state. We only used the off-medication scores.

MDS-UPDRS Part III has 36 scores of which we retained the first 33 for every subject for every imaging time. The last three ratings ("were dyskinesias present?", "did these movements interfere with your ratings?", "Hoehn and Yahr stage") were discarded either because they were non-informative (all subjects scored the same score) or because they could be considered a summary of other ratings (e.g. "Hoehn and Yahr stage").

In the PD literature, Part III scores are added to create a single number which summarizes the state of movement disorder for the patient. This summed score is called *Total Movement Score* (TMS) [47]. Larger values of TMS reflect worse PD symptoms. Fig. 11 shows the scatter plots of TMS for all subjects in the three progression subtypes. The plots in Fig. 11 also show the best-fit linear time regression to the scores. A slope, intercept and a p-value are calculated from the best-fit line. The p-value corresponds to testing the null hypothesis that the slope of the best-fit line is 0. A p-value less than 0.05 indicates that the slope is not zero, at a significance level of 0.05.

Fig. 11 shows that regressing TMS linearly with time for all subtypes gives a positive slope with a p-value significantly less than 0.05. Moreover, the slope for subtype 1 is bigger than the slopes for subtype 2 and subtype 3, which is consistent with image-based progression – subtype 1 has the fastest progression. The slope for subtype 2 is only slightly bigger than that for subtype 3, implying that the difference in terms of clinical symptom progression rate of TMS between subtype 2 and subtype 3 is smaller. Thus, TMS progression is consistent with DaTscan progression subtypes.

VI. Discussion

A. The MLDS Model

MLDS models the mean SBR from the caudates and the putamina. This is consistent with almost all of the PD DaTscan analysis literature, e.g. [48], [47]. In spite of its popularity, we readily acknowledge that there are limitations to this paradigm: the posterior-to-anterior anatomical gradient of disease progression in the striatum cannot be captured by mean SBRs. Moreover, extra-striatal structures such as the globus pallidus and the thalamus are also affected by PD [49], [50], but they are not included in the model. Clearly, what is needed is a finer grained model which also takes striatal subregions and extra-striatal regions into account. We plan to address this in the future.

The MLDS model can be generalized to other longitudinal datasets as long as the underlying progression satisfies a linear differential equation. For example, it can be naturally

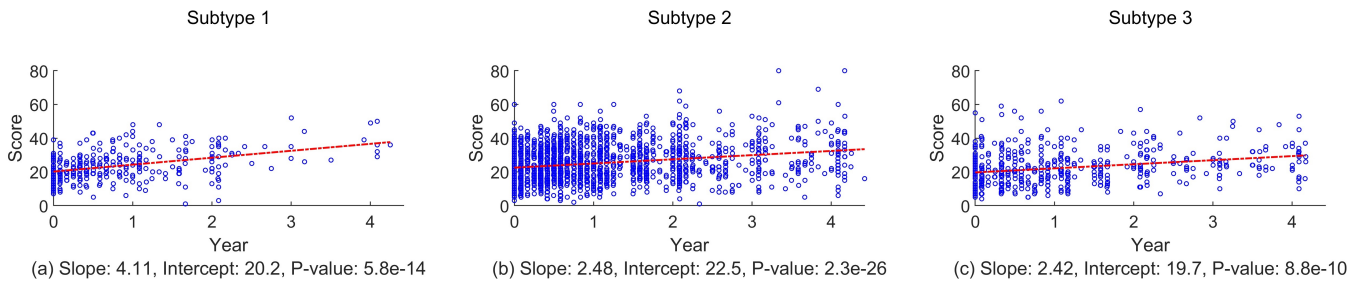


Fig. 11. Linear regression to the MDS-UPDRS Part III total scores for each subtype. The p-value is for the null hypothesis that the slope is 0.

extended to high dimensional features (e.g. voxelwise SBR) from DaTscan images by imposing a low-rank constraint on the transition matrices. Another example is the graph diffusion equation for modeling the progression of misfolded protein in the brain's connectivity network [23], where the connectivity information can be encoded in the prior to constrain the transition matrices.

B. Inferred Subtypes and Progression Patterns

The subtypes found by Bayesian inference clearly capture the progression heterogeneity as it manifests in DaTscans. Evidence, presented in Supplementary Section VI-B, shows that the subtypes do not represent time delayed versions of a single progression prototype.

The subtypes have different progression speeds with subtype 1 having the fastest progression, subtype 2 having a more moderate progression, and subtype 3 having the slowest progression. The values for π_k in Table II suggest that slightly more than half of the PD subjects belong to subtype 2, the remaining divided between subtype 1 and 3. Thus, one interpretation of the subtypes is that subtype 2 represents typical progression, while subtypes 1 and 3 represent the extremes of progression.

The presence of three progression subtypes in the PPMI dataset is also supported by other machine learning approaches. For example, Zhang et al. combine image and non-image features in the PPMI dataset (SBR, clinical scores, biospecimen exams) in a deep learning framework to find moderate, mild, and rapid progression subtypes [35]. However, that analysis does not reveal any eigenvectors or time constants. And our analysis only uses DaTscans.

It is remarkable that the eigenvector with the fastest time constant in all three subtypes corresponds to conversion from asymmetry to symmetry. The decrease of asymmetry in (non-DaTscan) PET images has been noted in the previous research [4], [51]. What MLDS reveals is that the decrease in asymmetry has the fastest possible time constant amongst all linear combinations of LC, LP, RP, RC.

The correlation between image-based subtypes and MDS-UPDRS Part III scores shows that at a group level, the progression rates measured by DaTscans reflect the progression rates of clinical symptoms. Note that in the PD literature, the reported correlations between DaTscans and UPDRS scores are usually quite small, ranging in magnitude from 0.1 to 0.3. The PD literature also suggests that correlations between *changes* in DaTscans and MDS-UPDRS are not significant [47]. However, these studies do not take subtypes into account. Our results show that TMS changes in the subtypes are similar to the subtype progression rates.

VII. Conclusions

This paper introduced a new longitudinal model and a Bayesian inference methodology for identifying progression subtypes and for finding disease progression patterns and their time constants for PD. The model is a mixture of linear dynamical systems, and is based on identifying key properties of PD progression. The model introduces several new ideas to PD modeling: coupled progression of multiple regions with population mirror symmetry, t-distributed model residues, mixtures of dynamical systems for heterogeneity, and a proper Bayesian analysis using Gibbs sampling.

Three image-based progression subtypes are found, differing in progression speeds. Each subtype displays characteristic spatial progression patterns with associated time constants. The fastest progression pattern in all subtypes is the loss of hemispheric asymmetry, while the slowest progression pattern is the change in the mean SBR. This finding has implications for clinical trials that assess the effectiveness of disease modifying therapies. The DaTscan-based subtypes also have different TMS progression rates.

Acknowledgements

This research was supported by the NIH grant R01NS107328. The data used in the preparation of this article was obtained from the Parkinson's Progression Markers Initiative (PPMI) database (up-to-date information is available at www.ppmi-info.org). PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and multiple funding partners. The full list of PPMI funding partners can be found at ppmi-info.org/fundingpartners.

References

[1] H. Braak, K. D. Tredici, U. Rub, R. A. de Vos, E. N. J. Steur, and E. Braak, "Staging of brain pathology related to sporadic Parkinson's disease," *Neurobiology of Aging*, vol. 24, no. 2, pp. 197 – 211, 2003.

[2] J. Booij *et al.*, "Practical benefit of [123I]FP-CIT SPET in the demonstration of the dopaminergic deficit in Parkinson's disease," *European Journal of Nuclear Medicine*, vol. 24, no. 1, pp. 68–71, Jan 1997.

[3] R. B. Innis *et al.*, "Consensus nomenclature for in vivo imaging of reversibly binding radioligands," *Journal of Cerebral Blood Flow & Metabolism*, vol. 27, no. 9, pp. 1533–1539, 2007.

[4] R. Nandhagopal *et al.*, "Longitudinal progression of sporadic Parkinson's disease: a multi-tracer positron emission tomography study," *Brain*, vol. 132, no. 11, pp. 2970–2979, 08 2009.

[5] F. R. Hampel, E. Ronchetti, P. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, 01 1986, vol. 29.

[6] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[7] Y. Zhou and H. D. Tagare, "Bayesian longitudinal modeling of early stage Parkinson's disease using DaTscan images," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 405–416.

[8] H. M. Fonteijn *et al.*, "An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease," *NeuroImage*, vol. 60, no. 3, pp. 1880–1889, 2012.

[9] A. L. Young *et al.*, "A data-driven model of biomarker changes in sporadic alzheimer's disease," *Brain*, vol. 137, no. 9, pp. 2564–2577, 2014.

[10] ——, "Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference," *Nature communications*, vol. 9, no. 1, pp. 1–16, 2018.

[11] V. Venkatraghavan, E. E. Bron, W. J. Niessen, S. Klein, A. D. N. Initiative *et al.*, "Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling," *NeuroImage*, vol. 186, pp. 518–532, 2019.

[12] K. Li and S. Luo, "Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer's disease," *Statistical methods in medical research*, vol. 28, no. 2, pp. 327–342, 2019.

[13] S. Iddi *et al.*, "Estimating the evolution of disease in the Parkinson's Progression Markers Initiative," *Neurodegenerative Diseases*, vol. 18, pp. 173–190, 2018.

[14] M. C. Donohue *et al.*, "Estimating long-term multivariate progression from short-term data," *Alzheimer's & Dementia*, vol. 10, pp. S400–S410, 2014.

[15] B. M. Jedynak *et al.*, "A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort," *Neuroimage*, vol. 63, no. 3, pp. 1478–1486, 2012.

[16] R. V. Marinescu *et al.*, "DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders," *NeuroImage*, vol. 192, pp. 166–177, 2019.

[17] M. Lorenzi *et al.*, "Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease," *NeuroImage*, 2017.

[18] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1095–1103.

[19] H. Wang *et al.*, "High-order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction," in *Advances in Neural Information Processing Systems*, 2012, pp. 1277–1285.

[20] N. P. Oxtoby *et al.*, "Learning imaging biomarker trajectories from noisy Alzheimer's disease data using a Bayesian multilevel model," in *Bayesian and grAphical Models for Biomedical Imaging*, M. J. Cardoso, I. Simpson, T. Arbel, D. Precup, and A. Ribbens, Eds. Cham: Springer International Publishing, 2014, pp. 85–94.

[21] ——, "Data-driven models of dominantly-inherited Alzheimer's disease progression," *Brain*, vol. 141, no. 5, pp. 1529–1544, 03 2018.

[22] W. W. Seeley, R. K. Crawford, J. Zhou, B. L. Miller, and M. D. Greicius, "Neurodegenerative diseases target large-scale human brain networks," *Neuron*, vol. 62, no. 1, pp. 42–52, 2009.

[23] A. Raj, A. Kuceyeski, and M. Weiner, "A network diffusion model of disease progression in dementia," *Neuron*, vol. 73, no. 6, pp. 1204–1215, 2012.

[24] A. Raj and F. Powell, "Models of network spread and network degeneration in brain disorders," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2018.

[25] C. Hu, X. Hua, J. Ying, P. M. Thompson, G. E. Fakhri, and Q. Li, "Localizing sources of brain disease progression with network diffusion model," *IEEE journal of selected topics in signal processing*, vol. 10, no. 7, pp. 1214–1225, 2016.

[26] S. Pandya *et al.*, "Predictive model of spread of Parkinson's pathology using network diffusion," *NeuroImage*, vol. 192, pp. 178–194, 2019.

[27] W. L. Au, J. R. Adams, A. Troiano, and A. J. Stoessel, "Parkinson's disease: in vivo assessment of disease progression using positron emission tomography," *Mol. Brain Research*, vol. 134, pp. 24–33, 2005.

[28] R. Hilker *et al.*, "Nonlinear progression of Parkinson disease as determined by serial positron emission tomographic imaging of striatal fluorodopa F 18 activity," *Archives of Neurology*, vol. 62, no. 3, pp. 378–382, 2005.

[29] I. G. Zubal, M. Early, O. Yuan, D. Jennings, K. Marek, and J. P. Seibyl, "Optimized, automated striatal uptake analysis applied to SPECT brain scans of Parkinson's disease patients," *The Journal of Nuclear Medicine*, vol. 48, no. 6, pp. 857–864, 2007.

[30] M. M. Hoehn and M. D. Yahr, "Parkinsonism: onset, progression and mortality," *Neurology*, vol. 17, pp. 427–442, 1967.

[31] P. Riederer, K. Jellinger, P. Kolber, G. Hipp, J. Sian-Hülsmann, and R. Krüger, "Lateralisation in Parkinson disease," *Cell and tissue research*, vol. 373, no. 1, pp. 297–312, 2018.

[32] M. A. Thenganatt and J. Jankovic, "Parkinson disease subtypes," *JAMA neurology*, vol. 71, no. 4, pp. 499–504, 2014.

[33] S.-M. Fereshtehnejad, S. R. Romenets, J. B. Anang, V. Latreille, J.-F. Gagnon, and R. B. Postuma, "New clinical subtypes of parkinson disease and their longitudinal progression: a prospective cohort comparison with other phenotypes," *JAMA neurology*, vol. 72, no. 8, pp. 863–873, 2015.

[34] M. Lawton *et al.*, "Developing and validating Parkinson's disease subtypes and their motor and cognitive progression," *J Neurol Neurosurg Psychiatry*, vol. 89, no. 12, pp. 1279–1287, 2018.

[35] X. Zhang *et al.*, "Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study," *Scientific reports*, vol. 9, no. 1, p. 797, 2019.

[36] J. R. Weaver, "Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors," *The American Mathematical Monthly*, vol. 92, no. 10, pp. 711–717, 1985.

[37] S. Fröhwirth-Schnatter and S. Kaufmann, "Model-based clustering of multiple time series," *Journal of Business & Economic Statistics*, vol. 26, no. 1, pp. 78–89, 2008.

[38] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution," *Statistics and computing*, vol. 10, no. 4, pp. 339–348, 2000.

[39] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[40] W. R. Gilks and P. Wild, "Adaptive rejection sampling for gibbs sampling," *Applied Statistics*, pp. 337–348, 1992.

[41] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[42] J. S. Liu, W. H. Wong, and A. Kong, "Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, vol. 81, no. 1, pp. 27–40, 1994.

[43] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[44] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[45] G. J. McLachlan and S. Rathnayake, "On the number of components in a Gaussian mixture model," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 5, pp. 341–355, 2014.

[46] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[47] T. Simuni *et al.*, "Longitudinal change of clinical and biological measures in early Parkinson's disease: Parkinson's progression markers initiative cohort," *Movement Disorders*, vol. 33, no. 5, pp. 771–782, 2018.

[48] R. Prashanth, S. D. Roy, P. K. Mandal, and S. Ghosh, "Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3333–3342, 2014.

[49] A. Rajput, H. Sitte, A. Rajput, M. Fenton, C. Pifl, and O. Hornykiewicz, "Globus pallidus dopamine and Parkinson motor subtypes: clinical and brain biochemical correlation," *Neurology*, vol. 70, no. 16 Part 2, pp. 1403–1410, 2008.

[50] P. Remy, M. Doder, A. Lees, N. Turjanski, and D. Brooks, "Depression in Parkinson's disease: loss of dopamine and noradrenaline innervation in the limbic system," *Brain*, vol. 128, no. 6, pp. 1314–1322, 2005.

[51] J. F. Fu *et al.*, "Joint pattern analysis applied to PET DAT and VMAT2 imaging reveals new insights into Parkinson's disease induced presynaptic alterations," *NeuroImage: Clinical*, vol. 23, p. 101856, 2019.

Supplementary Material for "Robust Bayesian Analysis of Early-Stage Parkinson's Disease Progression Using DaTscan Images"

Yuan Zhou, Sule Tinaz, and Hemant D. Tagare

Abstract

This supplementary material gives details of the Gibbs sampling algorithm mentioned in Section IV of the manuscript. We start with introducing the background of Gibbs sampling. Then we present the posterior predictive distributions related to the conjugate priors used in the paper. These predictive distributions are used in deriving the collapsed Gibbs sampler. Implementation issues such as convergence and label switching are discussed afterwards. Besides the sampling algorithm, this document also gives details on calculating the marginal likelihood for model selection, and on predicting future time points. Finally, additional results for the pre-processing procedure and the clustered subtypes are included.

I. Introduction

We begin by recalling that we refer to the time series of the i^{th} subject as $\mathbf{x}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i}\}$ with time intervals $\{\Delta t_{i1}, \ldots, \Delta t_{i,T_i-1}\}$ and the set of all time series as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Our goal is to infer the parameters $\boldsymbol{\theta}$ and ν from $\mathbf{X}$. To achieve this, we use Gibbs sampling to generate samples of the parameters as well as the latent variables, i.e. we generate samples from $p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{W}, \nu | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. The Gibbs sampling algorithm we use for this purpose is iterative; each iteration first samples from $\mathbf{z}, \boldsymbol{\theta}$ conditioned on remaining variables, and then samples from $\mathbf{W}, \nu$ conditioned on remaining variables:

$$\mathbf{z}, \boldsymbol{\theta} \sim p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{W}, \nu),$$
$$\mathbf{W}, \nu \sim p(\mathbf{W}, \nu | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\theta}).$$

Because the above random variables are related via a graphical model, the conditioning simplifies to conditioning on the *Markov blanket* of the variables that are being sampled. The Markov blanket $\text{MB}(u)$ of any random variable u includes its parents, children, and children's parents. The Markov blankets can be easily read off from Fig. S1(a). According to Fig. S1(a), we only need to sample using the following simplified conditioning:

$$p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{W}, \nu) = p(\mathbf{z}, \boldsymbol{\theta} | \text{MB}(\mathbf{z}, \boldsymbol{\theta})) = p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}),$$
$$p(\mathbf{W}, \nu | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{W}, \nu | \text{MB}(\mathbf{W}, \nu)) = p(\mathbf{W}, \nu | \mathbf{X}, \boldsymbol{\gamma}, \mathbf{z}, \boldsymbol{\theta}).$$

Our Gibbs sampler is standard, except when sampling $\mathbf{z}$. Sampling $\mathbf{z}$ is via a special scheme called *collapsed Gibbs sampling*. The reason for using collapsed Gibbs sampling is that it is more efficient than regular Gibbs sampling [1], [2]. In collapsed sampling, we sample from $p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$ with $\boldsymbol{\theta}$ integrated out, i.e. we sample from $p(\mathbf{z} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$. It is worth noting that after integrating out $\boldsymbol{\theta}$, the original graphical model in Fig. S1(a) becomes the graphical model in Fig. S1(b), where the conditional independence of $\{z_i\}$ given $\boldsymbol{\pi}$ no longer holds and all the z_is in $p(\mathbf{z} | \boldsymbol{\alpha})$ are dependent on each other. The same happens after integrating out $\mathbf{A}_k, \sigma_k^2$, i.e. $\{\mathbf{x}_i\}$ becomes dependent given $\boldsymbol{\beta}$. Collapsed sampling of $p(\mathbf{z} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$ proceeds by sampling each z_i sequentially from

$$p(z_i | \text{MB}(z_i)) = p(z_i = k | \mathbf{z}_{-i}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}). \tag{S1}$$

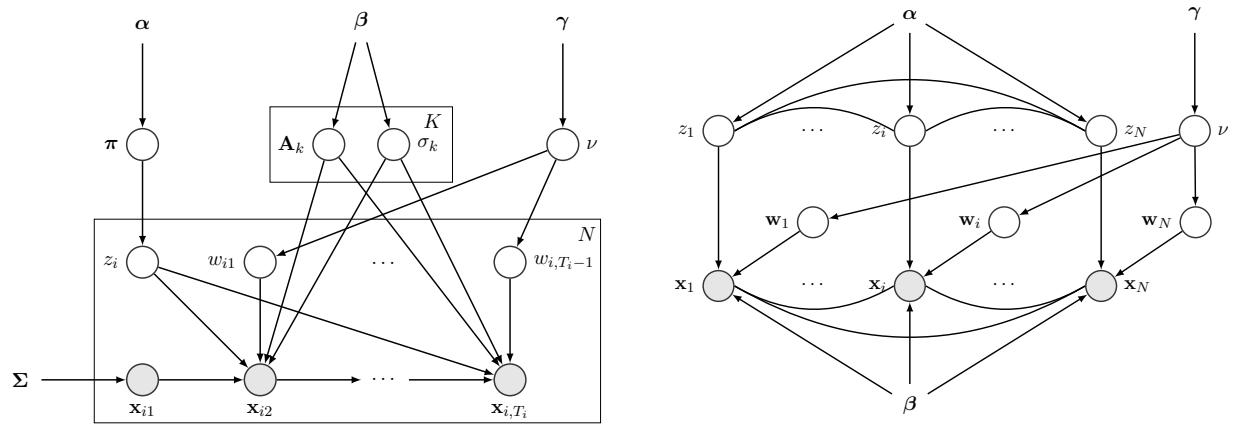
The details of how this sampling is carried out are given in Section III-A.

Once $\mathbf{z}$ is sampled from $p(\mathbf{z} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$, we sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta} | \mathbf{z}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$ since

$$p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) = p(\mathbf{z} | \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) p(\boldsymbol{\theta} | \mathbf{z}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}).$$

Finally, sampling $\mathbf{W}, \nu$ follows the standard Gibbs sampler by sampling one variable conditioned on its Markov blanket. Following Fig. S1(a):

$$p(\boldsymbol{\pi} | \text{MB}(\boldsymbol{\pi})) = p(\boldsymbol{\pi} | \mathbf{z}, \boldsymbol{\alpha}), \tag{S2}$$



(a) Original graphical model

(b) Graphical model with $\boldsymbol{\theta}$ integrated out

Fig. S1. Probabilistic graphical model of the MLDS with t-distributed residues. (a) shows the original model that features a standard structure of four layers: hyperparameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$), parameters to infer ($\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}_k, \sigma_k^2\}, \nu$), latent variables ($\mathbf{z} = (z_1, \ldots, z_N)$, $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$), and observed data ($\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$). (b) shows the model used in the collapsed Gibbs sampling where $\boldsymbol{\theta}$ is integrated out.

$$p\left(\mathbf{A}_k, \sigma_k^2 | \text{MB}\left(\mathbf{A}_k, \sigma_k^2\right)\right) = p\left(\mathbf{A}_k, \sigma_k^2 | \boldsymbol{\beta}, \mathbf{X}, \mathbf{z}, \mathbf{W}\right) = p\left(\mathbf{A}_k, \sigma_k^2 | \boldsymbol{\beta}, \mathbf{X}_k, \mathbf{W}_k\right), \tag{S3}$$

$$p\left(w_{ij} | \text{MB}\left(w_{ij}\right)\right) = p\left(w_{ij} | \mathbf{x}_{i,j+1}, z_i = k, \mathbf{A}_k, \sigma_k^2, \nu, \mathbf{x}_{ij}\right), \tag{S4}$$

$$p\left(\nu | \text{MB}\left(\nu\right)\right) = p\left(\nu | \mathbf{W}, \boldsymbol{\gamma}\right). \tag{S5}$$

where $\mathbf{X}_k = \{\mathbf{x}_j : z_j = k, j = 1, \ldots, N\}$ (the same for $\mathbf{W}_k$).

The formulae for Eqs. (S2) and (S4) are given in Section III-A. To obtain formulae for Eqs. (S3) and (S5), we need conjugate priors on $\mathbf{A}_k, \sigma_k^2$ and ν. These priors are provided in Section II. Combining the conjugate priors and the derived closed forms of these distributions, the algorithm samples from Eqs. (S1) - (S5) iteratively. The whole algorithm is presented in detail in Section III-B. Preliminary results on convergence are provided in Section III-C. Finally, besides the sampling scheme, additional information on model selection and prediction using the model is given in Section IV and Section V respectively.

II. Conjugate priors, Posteriors, and Posterior Predictive densities

A. Conjugate prior for $\mathbf{A}_k, \sigma_k^2$

Suppose $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{x}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i}\}$. Combining

$$p\left(\mathbf{x}_{i,j+1} | \mathbf{x}_{ij}, w_{ij}, \mathbf{A}, \sigma^2\right) = \mathcal{N}\left(\mathbf{x}_{i,j+1} | \mathbf{x}_{ij} + \Delta t_{ij} \mathbf{A} \mathbf{x}_{ij}, \Delta t_{ij}^2 \sigma^2 \mathbf{I}_D / w_{ij}\right). \tag{S6}$$

for all j, the distribution of $\mathbf{x}_i$ is

$$\begin{aligned} p\left(\mathbf{x}_i | \mathbf{A}, \sigma^2, \mathbf{w}_i\right) &= p\left(\mathbf{x}_{i1}\right) \prod_{j=1}^{T_i - 1} p\left(\mathbf{x}_{i,j+1} | \mathbf{x}_{ij}, \mathbf{A}, \sigma^2, w_{ij}\right) \\ &= \frac{p\left(\mathbf{x}_{i1}\right) \prod_j w_{ij}^{D/2}}{\left(2\pi\sigma^2\right)^{(T_i - 1)D/2} \prod_j \Delta t_{ij}^D} e^{-\frac{1}{2\sigma^2} \sum_j w_{ij} \| \mathbf{v}_{ij} - \mathbf{A}\mathbf{x}_{ij} \|^2}, \end{aligned} \tag{S7}$$

where $\mathbf{v}_{ij} = \frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{ij}}{\Delta t_{ij}}$. Representing $\mathbf{A}$ using a basis $\mathbf{E}$ of the subspace of centrosymmetric matrices, we have $\text{vec}\left(\mathbf{A}\right) = \mathbf{E}\mathbf{a}$. Rewriting Eq. (S7) by expanding the sum in its exponential term as a quadratic term of $\mathbf{a}$, we have

$$p\left(\mathbf{x}_i | \mathbf{a}, \sigma^2, \mathbf{w}_i\right) = \frac{h(\mathbf{x}_i)}{\sigma^{d_i}} e^{-\frac{1}{2\sigma^2}\left[\mathbf{a}^T \boldsymbol{\Lambda}(\mathbf{x}_i)\mathbf{a} - 2\boldsymbol{\mu}(\mathbf{x}_i)^T \mathbf{a} + \epsilon(\mathbf{x}_i)\right]} \tag{S8}$$

where $d_i = (T_i - 1)\, D$,

$$h(\mathbf{x}_i) = \frac{p\left(\mathbf{x}_{i1}\right) \prod_j w_{ij}^{D/2}}{(2\pi)^{(T_i-1)D/2} \prod_j \Delta t_{ij}^{D}}, \qquad \boldsymbol{\mu}(\mathbf{x}_i) = \mathbf{E}^T \sum_{j=1}^{T_i-1} w_{ij} \text{vec}\left(\mathbf{v}_{ij}\mathbf{x}_{ij}^T\right), \tag{S9}$$

$$\boldsymbol{\Lambda}(\mathbf{x}_i) = \mathbf{E}^T \left(\sum_{j=1}^{T_i-1} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^T \otimes \mathbf{I}_D \right) \mathbf{E}, \qquad \epsilon(\mathbf{x}_i) = \sum_{j=1}^{T_i-1} w_{ij}\mathbf{v}_{ij}^T\mathbf{v}_{ij}. \tag{S10}$$

Eqs. (S9) and (S10) define the notations $h, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \epsilon$ via the formulae on the right hand sides. Note that $\{\Delta t_{ij} : j = 1, \ldots, T_i - 1\}$, T_i, and $\mathbf{w}_i$ are implicitly used and change per index i. Hence, $h, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \epsilon$ are not functions of $\mathbf{x}_i$ (i.e. $\mathbf{x}_i = \mathbf{x}_j$ does not imply $h(\mathbf{x}_i) = h(\mathbf{x}_j)$) but rather functions of the subject index.

The exponent in equation (S8) is a quadratic semi-definite form in $\mathbf{a}$, while σ^2 plays the role similar to the variance of a normal density. From standard conjugate prior theory, we know that that the conjugate prior to $\mathbf{a}$ and σ^2 is normal-inverse-Gamma [3]. The exact form is given in the theorem below.

Theorem 1. *Suppose $\{\mathbf{x}_i \in \mathbb{R}^{D\times T_i} : i = 1, ..., N\}$ are N conditionally independent random variables with probability densities $p\left(\mathbf{x}_i|\mathbf{a}, \sigma^2\right)$ given by Eq.* (S8)*, then the joint density of $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ has a conjugate prior in the form of normal-inverse-gamma (NIG) with hyperparameters $\boldsymbol{\beta} = \{\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\}$,*

$$\begin{aligned} p\left(\mathbf{a}, \sigma^2|\boldsymbol{\beta}\right) &= \mathit{NIG}\left(\mathbf{a}, \sigma^2|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\right) \\ &= \mathcal{N}\left(\mathbf{a}|\boldsymbol{\mu}_0, \sigma^2\boldsymbol{\Lambda}_0^{-1}\right) \mathit{IG}\left(\sigma^2|\nu_0, \kappa_0\right) \end{aligned}$$

where $\boldsymbol{\Lambda}_0$ is positive definite and $\mathit{IG}\left(x|a, b\right) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-\frac{b}{x}}$. Using the conjugate prior, the posterior density of $\mathbf{a}, \sigma^2$ is also NIG:

$$p\left(\mathbf{a}, \sigma^2|\mathbf{X}, \boldsymbol{\beta}\right) = \mathit{NIG}\left(\mathbf{a}, \sigma^2|\boldsymbol{\mu}_p, \boldsymbol{\Lambda}_p, \nu_p, \kappa_p\right),$$

where

$$\begin{aligned} \nu_p &= \nu_0 + \frac{1}{2}\sum_i d_i, \\ \boldsymbol{\Lambda}_p &= \boldsymbol{\Lambda}_0 + \sum_i \boldsymbol{\Lambda}(\mathbf{x}_i), \\ \boldsymbol{\mu}_p &= \boldsymbol{\Lambda}_p^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \sum_i \boldsymbol{\mu}(\mathbf{x}_i)\right), \ \textit{and,} \\ \kappa_p &= \kappa_0 + \frac{1}{2}\left(\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \sum_i \epsilon(\mathbf{x}_i) - \boldsymbol{\mu}_p^T\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p\right). \end{aligned} \tag{S11}$$

Proof: To simplify the notation in this proof, we use $h_i, \epsilon_i, \boldsymbol{\mu}_i$ and $\boldsymbol{\Lambda}_i$, to represent $h(\mathbf{x}_i), \epsilon(\mathbf{x}_i), \boldsymbol{\mu}(\mathbf{x}_i)$ and $\boldsymbol{\Lambda}(\mathbf{x}_i)$ respectively.

The probability density of $\mathbf{X}$ conditioned on $\mathbf{a}, \sigma^2$ is

$$\begin{aligned} p\left(\mathbf{X}|\mathbf{a}, \sigma^2\right) = \prod_i p\left(\mathbf{x}_i|\mathbf{a}, \sigma^2\right) &= \frac{\prod_i h_i}{\sigma^{\sum_i d_i}} e^{-\frac{1}{2\sigma^2}\sum_i\left(\mathbf{a}^T\boldsymbol{\Lambda}_i\mathbf{a} - 2\boldsymbol{\mu}_i^T\mathbf{a} + \epsilon_i\right)} \\ &= \frac{\prod_i h_i}{\sigma^{\sum_i d_i}} e^{-\frac{1}{2\sigma^2}\left[\mathbf{a}^T\left(\sum_i \boldsymbol{\Lambda}_i\right)\mathbf{a} - 2\left(\sum_i \boldsymbol{\mu}_i\right)^T\mathbf{a} + \sum_i \epsilon_i\right]}. \end{aligned}$$

Using the prior, the posterior density of $\mathbf{a}, \sigma^2$ is

$$\begin{aligned} &p\left(\mathbf{a}, \sigma^2|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\right) \propto p\left(\mathbf{a}, \sigma^2|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\right) p\left(\mathbf{X}|\mathbf{a}, \sigma^2\right) \\ &= |\boldsymbol{\Lambda}_0|^{\frac{1}{2}} (2\pi)^{-\frac{D}{2}} \frac{\kappa_0^{\nu_0}}{\Gamma(\nu_0)} \left(\frac{1}{\sigma^2}\right)^{\frac{D}{2}+\nu_0+1} e^{-\frac{1}{2\sigma^2}\left[2\kappa_0 + (\mathbf{a}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}_0(\mathbf{a}-\boldsymbol{\mu}_0)\right]} \frac{\prod_i h_i}{\sigma^{\sum_i d_i}} e^{-\frac{1}{2\sigma^2}\left[\mathbf{a}^T\left(\sum_i \boldsymbol{\Lambda}_i\right)\mathbf{a} - 2\left(\sum_i \boldsymbol{\mu}_i\right)^T\mathbf{a} + \sum_i \epsilon_i\right]} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{D}{2}+\nu_0+1+\frac{\sum_i d_i}{2}} e^{-\frac{1}{2\sigma^2}\left[2\kappa_0 + (\mathbf{a}-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}_0(\mathbf{a}-\boldsymbol{\mu}_0) + \mathbf{a}^T\left(\sum_i \boldsymbol{\Lambda}_i\right)\mathbf{a} - 2\left(\sum_i \boldsymbol{\mu}_i\right)^T\mathbf{a} + \sum_i \epsilon_i\right]}. \end{aligned}$$

Straightforward algebraic manipulation simplifies the exponential term to

$$\begin{aligned} & 2\kappa_0 + (\mathbf{a}-\boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0 (\mathbf{a}-\boldsymbol{\mu}_0) + \mathbf{a}^T \left(\sum_i \boldsymbol{\Lambda}_i\right) \mathbf{a} - 2\left(\sum_i \boldsymbol{\mu}_i\right)^T \mathbf{a} + \sum_i \epsilon_i \\ = \quad & \left(\mathbf{a}-\boldsymbol{\mu}_p\right)^T \boldsymbol{\Lambda}_p \left(\mathbf{a}-\boldsymbol{\mu}_p\right) + 2\kappa_p, \end{aligned}$$

where

$$\boldsymbol{\Lambda}_p = \boldsymbol{\Lambda}_0 + \sum_i \boldsymbol{\Lambda}_i, \quad \boldsymbol{\mu}_p = \boldsymbol{\Lambda}_p^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \sum_i \boldsymbol{\mu}_i\right), \quad \kappa_p = \kappa_0 + \frac{1}{2}\left(\boldsymbol{\mu}_0^T\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \sum_i \epsilon_i - \boldsymbol{\mu}_p^T\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p\right).$$

Hence, the posterior is NIG

$$p\left(\mathbf{a},\sigma^2|\mathbf{X},\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0,\nu_0,\kappa_0\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{D}{2}+\nu_p+1} e^{-\frac{1}{2\sigma^2}\left[\left(\mathbf{a}-\boldsymbol{\mu}_p\right)^T\boldsymbol{\Lambda}_p\left(\mathbf{a}-\boldsymbol{\mu}_p\right)+2\kappa_p\right]},$$

with $\nu_p = \nu_0 + \frac{\sum_i d_i}{2}$. ■

One consequence of Theorem 1 is that the posterior predictive density can be explicitly stated:

Corollary 2. *Following the definitions and notations of Theorem 1, supposing that we have a new time series* $\mathbf{x} \in \mathbb{R}^{D\times T}$ *with the same form in Eq.* (S8) *and functions* $h, \epsilon, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ *of* $\mathbf{x}$*, the posterior predictive density of* $\mathbf{x}$ *given old data* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *and hyperparameter* $\boldsymbol{\beta}$ *is*

$$\begin{aligned} p\left(\mathbf{x}|\mathbf{X},\boldsymbol{\beta}\right) &= \int\int p\left(\mathbf{x}|\mathbf{a},\sigma^2\right) p\left(\mathbf{a},\sigma^2|\mathbf{X},\boldsymbol{\beta}\right) d\mathbf{a}d\sigma^2 \\ &= \int\int p\left(\mathbf{x}|\mathbf{a},\sigma^2\right) NIG\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p(\mathbf{X}),\boldsymbol{\Lambda}_p(\mathbf{X}),\nu_p(\mathbf{X}),\kappa_p(\mathbf{X})\right) d\mathbf{a}d\sigma^2 \\ &= h(\mathbf{x}) \frac{\left|\boldsymbol{\Lambda}_p\right|^{1/2}}{\left|\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\left(\mathbf{x}\right)\right|^{1/2}} \frac{\Gamma\left(\nu_p+\frac{d}{2}\right)}{\Gamma\left(\nu_p\right)} \kappa_p^{-\frac{d}{2}} \left[1+\frac{Q\left(\mathbf{x}\right)}{2\kappa_p}\right]^{-\left(\nu_p+\frac{d}{2}\right)}, \end{aligned}$$

where

$$Q(\mathbf{x}) = \boldsymbol{\mu}_p^T\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p + \epsilon(\mathbf{x}) - \left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p + \boldsymbol{\mu}(\mathbf{x})\right)^T \left(\boldsymbol{\Lambda}_p + \boldsymbol{\Lambda}(\mathbf{x})\right)^{-1} \left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p + \boldsymbol{\mu}(\mathbf{x})\right).$$

and $d = (T-1)D$.

Proof: As we did for the proof of Theorem 1, in this proof too we simplify notations. We use $\boldsymbol{\Lambda}, \boldsymbol{\mu}$ and Q to represent $\boldsymbol{\Lambda}(\mathbf{x}), \boldsymbol{\mu}(\mathbf{x})$ and $Q(\mathbf{x})$. Rearranging

$$p\left(\mathbf{a},\sigma^2|\mathbf{x},\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right) = \frac{p\left(\mathbf{x}|\mathbf{a},\sigma^2\right) \mathrm{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)}{\int\int p\left(\mathbf{x}|\mathbf{a},\sigma^2\right) \mathrm{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right) d\mathbf{a}d\sigma^2}$$

gives

$$\int\int p\left(\mathbf{x}|\mathbf{a},\sigma^2\right) \mathrm{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right) d\mathbf{a}d\sigma^2 = \frac{p\left(\mathbf{x}|\mathbf{a},\sigma^2\right) \mathrm{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)}{p\left(\mathbf{a},\sigma^2|\mathbf{x},\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)}.$$

Thus the integral can be calculated by evaluating the numerator and denominator on the right hand side.

Using Theorem 1, we have

$$p\left(\mathbf{a},\sigma^2|\mathbf{x},\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right) = \mathrm{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p',\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda},\nu_p+\frac{d}{2},\kappa_p'\right),$$

where

$$\begin{aligned} \boldsymbol{\mu}_p' &= \left(\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p+\boldsymbol{\mu}\right) \\ \kappa_p' &= \kappa_p + \frac{1}{2}\left(\boldsymbol{\mu}_p^T\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p + \epsilon - \left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p+\boldsymbol{\mu}\right)^T\left(\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p+\boldsymbol{\mu}\right)\right). \end{aligned}$$

Thus

$$\begin{aligned}
&\int\int p\left(\mathbf{x}|\mathbf{a},\sigma^2\right)\text{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)d\mathbf{a}d\sigma^2\\
&=\frac{p\left(\mathbf{x}|\mathbf{a},\sigma^2\right)\text{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)}{p\left(\mathbf{a},\sigma^2|\mathbf{x},\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)}\\
&=\frac{\frac{h}{\sigma^d}e^{-\frac{1}{2\sigma^2}\left(\mathbf{a}^T\boldsymbol{\Lambda}\mathbf{a}-2\boldsymbol{\mu}^T\mathbf{a}+\epsilon\right)}\left|\boldsymbol{\Lambda}_p\right|^{\frac{1}{2}}(2\pi)^{-\frac{D}{2}}\frac{\kappa_p^{\nu_p}}{\Gamma(\nu_p)}\left(\frac{1}{\sigma^2}\right)^{\frac{D}{2}+\nu_p+1}e^{-\frac{1}{2\sigma^2}\left[2\kappa_p+\left(\mathbf{a}-\boldsymbol{\mu}_p\right)^T\boldsymbol{\Lambda}_p\left(\mathbf{a}-\boldsymbol{\mu}_p\right)\right]}}{\left|\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right|^{\frac{1}{2}}(2\pi)^{-\frac{D}{2}}\frac{\left(\kappa_p'\right)^{\nu_p+\frac{d}{2}}}{\Gamma\left(\nu_p+\frac{d}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{D}{2}+\nu_p+\frac{d}{2}+1}e^{-\frac{1}{2\sigma^2}\left[2\kappa_p'+\left(\mathbf{a}-\boldsymbol{\mu}_p'\right)^T\left(\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right)\left(\mathbf{a}-\boldsymbol{\mu}_p'\right)\right]}}.
\end{aligned}$$

The three exponential terms cancel out, giving

$$\begin{aligned}
&\int\int p\left(\mathbf{x}|\mathbf{a},\sigma^2\right)\text{NIG}\left(\mathbf{a},\sigma^2|\boldsymbol{\mu}_p,\boldsymbol{\Lambda}_p,\nu_p,\kappa_p\right)d\mathbf{a}d\sigma^2\\
&=\frac{h}{\sigma^d}\frac{\left|\boldsymbol{\Lambda}_p\right|^{\frac{1}{2}}}{\left|\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right|^{\frac{1}{2}}}\frac{\Gamma\left(\nu_p+\frac{d}{2}\right)}{\Gamma\left(\nu_p\right)}\left(\frac{1}{\sigma^2}\right)^{-\frac{d}{2}}\frac{\kappa_p^{\nu_p}}{\left(\kappa_p'\right)^{\nu_p+\frac{d}{2}}}\\
&=h\frac{\left|\boldsymbol{\Lambda}_p\right|^{\frac{1}{2}}}{\left|\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right|^{\frac{1}{2}}}\frac{\Gamma\left(\nu_p+\frac{d}{2}\right)}{\Gamma\left(\nu_p\right)}\kappa_p^{-\frac{d}{2}}\left[1+\frac{Q}{2\kappa_p}\right]^{-\left(\nu_p+\frac{d}{2}\right)},
\end{aligned}$$

where

$$Q=\boldsymbol{\mu}_p^T\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p+\epsilon-\left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p+\boldsymbol{\mu}\right)^T\left(\boldsymbol{\Lambda}_p+\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_p\boldsymbol{\mu}_p+\boldsymbol{\mu}\right).$$

■

B. Conjugate prior for ν

The conditional density $p\left(\mathbf{W}|\nu\right)$ belongs to the exponential family:

$$\begin{aligned}
p\left(\mathbf{W}|\nu\right)&=\prod_{i=1}^{N}\prod_{j=1}^{T_i-1}\text{Ga}\left(w_{ij}|\frac{\nu}{2},\frac{\nu}{2}\right)\\
&=\left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{\sum_i(T_i-1)}\left(\prod_i\prod_j w_{ij}^{-1}\right)e^{\nu\cdot\frac{1}{2}\sum_{i,j}(\log w_{ij}-w_{ij})}.
\end{aligned}$$

and hence has the conjugate prior

$$p\left(\nu|\boldsymbol{\gamma}\right)\propto\left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{\xi_0}e^{\tau_0\nu},\ \nu>0$$

where $\boldsymbol{\gamma}=(\xi_0,\tau_0)$. The posterior has the same form as the prior

$$p\left(\nu|\mathbf{W},\boldsymbol{\gamma}\right)\propto p\left(\nu|\boldsymbol{\gamma}\right)p\left(\mathbf{W}|\nu\right)=\left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{\xi_0+\sum_i(T_i-1)}e^{\nu\left[\tau_0+\frac{1}{2}\sum_{i,j}(\log w_{ij}-w_{ij})\right]}.\tag{S12}$$

III. Gibbs Sampling for the MLDS

The Gibbs sampler uses the conjugate priors and predictive densities established in the previous section. The reader is warned that the notation for describing a Gibbs sampler in mixture models gets pretty complicated. The complication arises because we have to repeat calculations that are similar to those in Eqs. (S11) for different subsets of $\mathbf{X}$. To simplify the notation, we will first define a notation for the sets that we use, and then define the how the calculations in Eqs. (S11) are carried out for these sets:

Notation for sets:

1) We use subscript k to denote random variables associated with all subjects whose z takes the value k, e.g. $\mathbf{X}_k = \{\mathbf{x}_j | z_j = k, j = 1, \ldots, N\}$, $\mathbf{W}_k = \{\mathbf{w}_j | z_j = k, j = 1, \ldots, N\}$.
2) We use subscript $-i$ to denote random variables associated with subjects whose indices are not i, e.g., $\mathbf{z}_{-i} = \{z_j : j \neq i, j = 1, \ldots, N\}$, $\mathbf{X}_{-i} = \{\mathbf{x}_j : j \neq i, j = 1, \ldots, N\}$.
3) We also define the sets $\mathbf{X}_{-i,k} = \{\mathbf{x}_j : j \neq i, z_j = k, j = 1, \ldots, N\}$, and similarly, $\mathbf{W}_{-i,k} = \{\mathbf{w}_j : j \neq i, z_j = k, j = 1, \ldots, N\}$.

Notation for calculations: To represent calculations that are similar to those in Eqs. (S11) but carried out with the above sets, we adopt the following convention: Suppose $\mathbf{U}$ is any subset of $\mathbf{X}$, e.g. $\mathbf{U} = \mathbf{X}_k$, or $\mathbf{U} = \mathbf{X}_{-i,k}$. Then we use the notation $\boldsymbol{\Lambda}_p(\mathbf{U})$ to denote the calculation of $\boldsymbol{\Lambda}_p$ in Eqs. (S11) where the sum on the right hand side is carried out using information from all subjects whose time series is in $\mathbf{U}$. Similarly $\nu_p(\mathbf{U}), \boldsymbol{\mu}_p(\mathbf{U})$, and $\kappa_p(\mathbf{U})$ denote the calculations in Eq. (S11) where the sums in the right hand side of the equations are carried out using information from all subjects whose time series are in $\mathbf{U}$. Thus, terms such as $\boldsymbol{\Lambda}_p(\mathbf{X}_{-i})$ and $\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})$ are well defined. Finally we remind the reader that subscript p in $\boldsymbol{\Lambda}_p$ etc. refers to the posterior probability density. Without the subscript, the terms $h, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \epsilon$ are defined only for a single time series as given in Eqs. (S9) and (S10). Similar to $h, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \epsilon$, the notations $\nu_p, \boldsymbol{\Lambda}_p, \boldsymbol{\mu}_p, \kappa_p$ should not be seen as functions of $\mathbf{U}$, but rather they are functions of the subject indices.

A. Derivation of the Gibbs sampler

1) Sample $p(\mathbf{z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$ by collapsed Gibbs sampling: This is achieved by sampling each z_i sequentially conditioned on the other z_i's:

$$\begin{aligned} & p(z_i = k|\mathbf{z}_{-i}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) \\ \propto & p(z_i = k|\mathbf{X}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})\, p(\mathbf{x}_i|z_i = k, \mathbf{X}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) \\ = & p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha})\, p(\mathbf{x}_i|z_i = k, \mathbf{X}_{-i,k}, \boldsymbol{\beta}, \mathbf{W}_k)\,. \end{aligned} \tag{S13}$$

An explanation of the last step in equation (S13) is as follows: Consider the term $p(z_i = k|\mathbf{X}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$. According to Fig. S1(b), the Markov blanket of z_i is $\boldsymbol{\alpha}, \mathbf{z}_{-i}, \mathbf{x}_i, \mathbf{w}_i, \boldsymbol{\beta}$. Since $\mathbf{x}_i$ is not involved in the conditioning, the correlated $\mathbf{X}_{-i}$ and its parents ($\boldsymbol{\beta}$ and $\mathbf{W}$) are independent of z_i. Thus $p(z_i = k|\mathbf{X}_{-i}, \mathbf{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$ can be reduced to $p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha})$. The second term in Eq. (S13) simplifies because each $\mathbf{X}_k = \{\mathbf{x}_j : z_j = k, j = 1, \ldots, N\}$ is independently generated by $\boldsymbol{\beta}$ after integrating out $\mathbf{A}_k$, σ_k^2. If the parents of $\mathbf{x}_i$ are known and $z_i = k$, all the other $\{\mathbf{X}_l : l \neq k\}$ become independent of $\mathbf{x}_i$.

The first term in Eq. (S13) can be calculated by the following strategy: Assuming $\boldsymbol{\alpha} = (\alpha/K, \ldots, \alpha/K)$, and integrating out $\boldsymbol{\pi}$ gives

$$p(\mathbf{z}|\boldsymbol{\alpha}) = \int p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_i p(z_i|\boldsymbol{\pi})\, d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}, \tag{S14}$$

where $N_k = \sum_{i=1}^{N} \mathbb{I}(z_i = k)$ and $\mathbb{I}(\cdot) = 1$ if the the argument of $\mathbb{I}$ is true and zero otherwise. Hence the first term in Eq. (S13) is

$$p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha}) = \frac{p(z_i = k, \mathbf{z}_{-i}|\boldsymbol{\alpha})}{p(\mathbf{z}_{-i}|\boldsymbol{\alpha})} = \frac{N_{-i,k} + \alpha/K}{N + \alpha - 1}$$

where $N_{-i,k} = \sum_{j \neq i} \mathbb{I}(z_j = k)$.

The second term in Eq. (S13) is

$$\begin{aligned} & p(\mathbf{x}_i|z_i = k, \mathbf{X}_{-i,k}, \boldsymbol{\beta}, \mathbf{W}_k) \\ = & \int\int p\left(\mathbf{x}_i|\mathbf{a}_k, \sigma_k^2, \mathbf{w}_i\right) p\left(\mathbf{a}_k, \sigma_k^2|\mathbf{X}_{-i,k}, \boldsymbol{\beta}, \mathbf{W}_{-i,k}\right) d\mathbf{a}_k d\sigma_k^2 \end{aligned}$$

which is calculated according to Corollary 2 in Section II-A.

2) Sample $p(\boldsymbol{\theta}|\mathbf{z},\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$: Given $\mathbf{z}$, we can partition $\mathbf{X}$, the set of all time series into subsets $\{\mathbf{X}_k : k = 1,\ldots,K\}$ where the k^{th} subset contains all time series for which $z = k$. Then, sampling $\mathbf{A}_k$ and σ_k^2 follows Theorem 1:

$$p\left(\mathbf{a}_k,\sigma_k^2|\boldsymbol{\beta},\mathbf{X}_k,\mathbf{W}_k\right) = \text{NIG}\left(\mathbf{a}_k,\sigma_k^2|\boldsymbol{\mu}_p(\mathbf{X}_k),\boldsymbol{\Lambda}_p(\mathbf{X}_k),\nu_p(\mathbf{X}_k),\kappa_p(\mathbf{X}_k)\right).$$

Sampling $\boldsymbol{\pi}$ follows the standard posterior density

$$p\left(\boldsymbol{\pi}|\mathbf{z},\boldsymbol{\alpha}\right) = \text{Dir}\left(\boldsymbol{\pi}|\alpha/K+N_1,\ldots,\alpha/K+N_K\right). \tag{S15}$$

3) Sample $p(\mathbf{W},\nu|\mathbf{X},\boldsymbol{\gamma},\mathbf{z},\boldsymbol{\theta})$: Next, we sample $\mathbf{W}$ from $p(w_{ij}|\mathbf{x}_{i,j+1},z_i=k,\mathbf{A}_k,\sigma_k^2,\nu,\mathbf{x}_{ij})$, and sample ν from $p(\nu|\mathbf{W},\boldsymbol{\gamma})$. Using $p\left(w_{ij}|\nu\right) = \text{Ga}\left(w_{ij}|\frac{\nu}{2},\frac{\nu}{2}\right)$ and $p(\mathbf{x}_{i,j+1}|w_{ij},z_i=k,\mathbf{A}_k,\sigma_k^2,\mathbf{x}_{ij})$ in (S6), the posterior of w_{ij} is calculated as

$$\begin{aligned}
&p\left(w_{ij}|\mathbf{x}_{i,j+1},z_i=k,\mathbf{A}_k,\sigma_k^2,\nu,\mathbf{x}_{ij}\right)\\
\propto&p\left(w_{ij}|\nu\right)p\left(\mathbf{x}_{i,j+1}|w_{ij},z_i=k,\mathbf{A}_k,\sigma_k^2,\mathbf{x}_{ij}\right)\\
=&\text{Ga}\left(w_{ij}|\frac{D}{2}+\frac{\nu}{2},\frac{\nu}{2}+\frac{1}{2\sigma_k^2}\|\mathbf{v}_{ij}-\mathbf{A}_k\mathbf{x}_{ij}\|^2\right).
\end{aligned} \tag{S16}$$

Using a conjugate prior, $p\left(\nu|\mathbf{W},\boldsymbol{\gamma}\right)$ has a form in (S12) (see Section II-B). Since $p\left(\nu|\boldsymbol{\gamma}\right)$ is a log-concave function ($\frac{d^2\log p(\nu|\boldsymbol{\gamma})}{d\nu^2} < 0$ for all $\nu > 0$), adaptive rejection sampling (ARS) is used to sample ν [4].

B. The Gibbs sampling algorithm

We can now state the entire Gibbs sampling algorithm in detail.

1. First sample $p(\mathbf{z},\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$ as follows:
   - 1.1. Sample $\mathbf{z}$ from $p(\mathbf{z}|\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$ by sequentially sampling z_i from

$$p(z_i=k|\mathbf{z}_{-i},\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W}) \propto p(z_i=k|\mathbf{z}_{-i},\boldsymbol{\alpha})p(\mathbf{x}_i|z_i=k,\mathbf{X}_{-i,k},\boldsymbol{\beta},\mathbf{W}_k). \tag{S17}$$

     - 1.1.1. Calculate $p(z_i=k|\mathbf{z}_{-i},\boldsymbol{\alpha}) = \frac{N_{-i,k}+\alpha/K}{N+\alpha-1}$.
     - 1.1.2. Calculate

$$\begin{aligned}
&p(\mathbf{x}_i|z_i=k,\mathbf{X}_{-i,k},\boldsymbol{\beta},\mathbf{W}_k)\\
=&\frac{h(\mathbf{x}_i)}{\kappa_p(\mathbf{X}_{-i,k})^{\frac{d_i}{2}}}\frac{|\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})+\boldsymbol{\Lambda}(\mathbf{x}_i)|^{\frac{1}{2}}}\frac{\Gamma\left(\nu_p(\mathbf{X}_{-i,k})+\frac{d_i}{2}\right)}{\Gamma\left(\nu_p(\mathbf{X}_{-i,k})\right)}\left[1+\frac{Q}{2\kappa_p(\mathbf{X}_{-i,k})}\right]^{-\left(\nu_p(\mathbf{X}_{-i,k})+\frac{d_i}{2}\right)}
\end{aligned} \tag{S18}$$

     where

$$\begin{aligned}
Q &= -\left(\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})\boldsymbol{\mu}_p(\mathbf{X}_{-i,k})+\boldsymbol{\mu}(\mathbf{x}_i)\right)^T\left(\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})+\boldsymbol{\Lambda}(\mathbf{x}_i)\right)^{-1}\times\\
&\quad\left(\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})\boldsymbol{\mu}_p(\mathbf{X}_{-i,k})+\boldsymbol{\mu}(\mathbf{x}_i)\right)+\boldsymbol{\mu}_p(\mathbf{X}_{-i,k})^T\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k})\boldsymbol{\mu}_p(\mathbf{X}_{-i,k})+\epsilon(\mathbf{x}_i)
\end{aligned}$$

     and $d_i = (T_i-1)\,D$, where T_i is the number of time points.
   - 1.2. Sample $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{z},\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{W})$.
     - 1.2.1. Sample $\boldsymbol{\pi}$ from $\text{Dir}(\boldsymbol{\pi}|\frac{\alpha}{K}+N_1,\ldots,\frac{\alpha}{K}+N_K)$.
     - 1.2.2. Sample $\mathbf{A}_k,\sigma_k^2$ from $\text{NIG}\left(\mathbf{a}_k,\sigma_k^2|\boldsymbol{\mu}_p(\mathbf{X}_k),\boldsymbol{\Lambda}_p(\mathbf{X}_k),\nu_p(\mathbf{X}_k),\kappa_p(\mathbf{X}_k)\right)$.
2. Next, sample $p(\mathbf{W},\nu|\mathbf{X},\boldsymbol{\gamma},\mathbf{z},\boldsymbol{\theta})$:
   - 2.1. Sample w_{ij} from $\text{Ga}(w_{ij}|\frac{D}{2}+\frac{\nu}{2},\frac{\nu}{2}+\frac{1}{2\sigma_k^2}\|\mathbf{v}_{ij}-\mathbf{A}_k\mathbf{x}_{ij}\|^2)$ where $k=z_i$.
   - 2.2. Sample ν from $p(\nu|\mathbf{W},\boldsymbol{\gamma}) \propto \left[\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{\xi_0+\sum_i(T_i-1)} \times e^{\nu\left[\tau_0+\frac{1}{2}\sum_{i,j}(\log w_{ij}-w_{ij})\right]}$ using adaptive rejection sampling (ARS) [4].

The above algorithm has an intuitive interpretation. Each class label z_i is sampled based on the cluster size $N_{-i,k}$ and similarity of its data point $\mathbf{x}_i$ to that cluster. The robustness comes from the weight variables $\{w_{ij}\}$. Since a Gamma distribution $\text{Ga}\left(x|a,b\right)$ has mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$, when we have a large $\|\mathbf{v}_{ij}-\mathbf{A}_k\mathbf{x}_{ij}\|$ (e.g. outlier $\mathbf{x}_{i,j+1}$), we are likely to sample a small w_{ij}, which in turn makes this data negligible in calculating the sufficient statistics in (S9) and (S10).

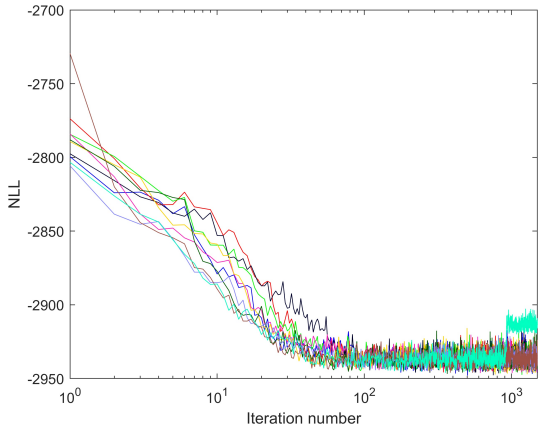


Fig. S2. NLL of 10 runs on the PPMI DaTscan dataset.

We want to make some remarks about the implementation. First, $h(\mathbf{x}_i)$ can be removed in calculating (S18) because it appears in the numerator and the denominator. The term $h(\mathbf{x}_i)$ appears in the denominator because it appears in (S17) which involves a partition function that sums over all k. Because $h(\mathbf{x}_i)$ involves an evaluation of $p\left(\mathbf{x}_{i1}\right)$, and $h(\mathbf{x}_i)$ cancels, the density of the initial data can be ignored.

Second, most of the complexity of the algorithm comes from sampling z_i sequentially. Examining the algorithm carefully shows that the posterior predictive density relies on a sum of the sufficient statistics $\boldsymbol{\mu}(\mathbf{x}_i), \boldsymbol{\Lambda}(\mathbf{x}_i), \epsilon(\mathbf{x}_i)$ for all the subjects in a cluster (see Eq. (S11)). Storing these sums in each cluster, and when assigning a new class label to a subject, subtracting the sufficient statistics of the subject from the sums in the old cluster and adding the same amount to the new cluster is an efficient way to update the sufficient statistics. With this in mind, the computational cost of sampling z_i can be analyzed as follows.

Let R be the number of basis in $\mathbf{E}$, calculating $\boldsymbol{\mu}(\mathbf{x}_i), \boldsymbol{\Lambda}(\mathbf{x}_i), \epsilon(\mathbf{x}_i)$ takes $O(T_iD^2 + RD^2)$, $O(T_iD^2 + RD^4 + R^2D^2)$, $O(T_iD)$ respectively. The complexity of adding and subtracting these terms from the summed sufficient statistics is negligible. The calculation of the posterior parameters for $\mathbf{X}_{-i,k}$ using Eq. (S11), is dominated by the inversion of $\boldsymbol{\Lambda}_p$ which has complexity $O(R^3)$. Then, using the calculated posterior parameters, calculating Q in Eq. (S18) is dominated by the inversion of $\boldsymbol{\Lambda}_p(\mathbf{X}_{-i,k}) + \boldsymbol{\Lambda}(\mathbf{x}_i)$, whose complexity is $O(R^3)$. Since the determinant in (S18) also has computational complexity $O(R^3)$, the entire calculation of Eq. (S18) has complexity $O(R^3)$. The above computation needs to be repeated for K clusters in Eq. (S17), N subjects, and L iterations. In our case, $R = \frac{1}{2}D^2$, $T_i < R$, hence the entire complexity of sampling $\mathbf{z}$ is $O(LNKD^6)$. The complexity of sampling $\mathbf{z}$ dominates the entire algorithm. Though $O(LNKD^6)$ looks expensive, for our data, $D = 4$, $K = 3$, $N = 365$, $L = 1500$, which makes the algorithm fairly efficient on a modern computer. For example, on a computer with a Intel Xeon W-2155 CPU, 64 GB memory, running 5 chains in parallel, takes 450 seconds.

C. Convergence and other issues

Next we discuss the convergence of the sampler and its robustness to label switching.

1) Convergence: Using the PPMI data, we ran 10 Markov chains of the Gibbs sampler for 1500 iterations using random initialization. Fig. S2 shows the negative log-likelihoods (NLL) in each chain as a function of iteration number. Clearly all chains have converged within 100 iterations and entered their stationary distributions (this fast convergence is likely due to the collapsed Gibbs sampler). To be conservative, when processing PPMI data, we use 1500 iterations and discard the first half as a burn-in period. Also notice that in Fig. S2, 9 of the 10 chains have converged to the same distribution. Empirically, we found that running 5 chains and picking the commonly converged chains was sufficient to give a stable result most of the time. We discarded the remaining chains, and used one converged chain for analysis.

2) Label switching: In Gibbs sampling for mixture models, there is a label switching problem because the posterior $p\left(\mathbf{z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}\right)$ has $K!$ symmetric modes [5]. When a Markov chain is run for extended time, it tends to visit all modes uniformly. The resulting $p\left(\mathbf{z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}\right)$ has its labels switched (each mode corresponds to a permutation of the class labels) and the mean of the samples generated by these chains is not very meaningful. Fortunately, label switching seldom happens for relatively short chains [6].

To check if label switching happens in our dataset, we employ the following strategy: First, from amongst all z's generated by the sampler, we choose the one that has the highest log-likelihood, and call it a template. Then for every z generated by the sampler, we find the permutation of labels that makes z most consistent with the template. If this permutation is the identity permutation, then no label switching has occurred. Thus, we count the percent of z's in the entire sampling chain in which no label switching has occurred. In the synthetic experiments ($3 \times 10 \times 10$ runs) (Section V-B in the paper), for the overwhelming majority of the runs this percentage was 100% (one run scored at 99.9%, and one run scored at 99.5% for the synthetic data with large noise). No label switching was detected in the PPMI dataset.

IV. Model Selection

We use importance sampling to approximate the integral $\int p(\mathbf{X}|\boldsymbol{\theta},\nu)\, p(\boldsymbol{\theta},\nu|\boldsymbol{\eta}_K)\, d\boldsymbol{\theta} d\nu$ since samples from $p(\boldsymbol{\theta},\nu|\boldsymbol{\eta}_K)$ mostly contribute zeros to $p(\mathbf{X}|\boldsymbol{\theta},\nu)$. Denoting $f(\boldsymbol{\theta},\nu) = p(\mathbf{X}|\boldsymbol{\theta},\nu)$ and $\tilde{p}(\boldsymbol{\theta},\nu) = p(\boldsymbol{\theta},\nu|\boldsymbol{\eta}_K)$, our goal is to calculate $\int f(\boldsymbol{\theta},\nu)\, \tilde{p}(\boldsymbol{\theta},\nu)\, d\boldsymbol{\theta} d\nu$.

We use the proposal density $q(\boldsymbol{\theta},\nu) = p(\boldsymbol{\theta},\nu|\mathbf{X},\boldsymbol{\eta}_K)$ since the Gibbs sampler has already generated samples $\{\boldsymbol{\theta}_l, \nu_l : l = 1,\ldots,L\}$ from this density, and it is also the optimal importance distribution in terms of minimizing the variance of the estimate (see Section 23.4 in [1]). Since

$$q(\boldsymbol{\theta},\nu) = p(\boldsymbol{\theta},\nu|\mathbf{X},\boldsymbol{\eta}_K) \propto p(\boldsymbol{\theta},\nu|\boldsymbol{\eta}_K)\, p(\mathbf{X}|\boldsymbol{\theta},\nu,\boldsymbol{\eta}_K),$$

we use the unnormalized version $\tilde{q}(\boldsymbol{\theta},\nu) = p(\boldsymbol{\theta},\nu|\boldsymbol{\eta}_K)\, p(\mathbf{X}|\boldsymbol{\theta},\nu)$ for importance sampling. It has the integral

$$\int f(\boldsymbol{\theta},\nu)\, \tilde{p}(\boldsymbol{\theta},\nu)\, d\boldsymbol{\theta} d\nu \approx \sum_{l=1}^{L} w_l f(\boldsymbol{\theta}_l,\nu_l),$$

where $w_l = \frac{\tilde{w}_l}{\sum_i \tilde{w}_i}$ and

$$\tilde{w}_l = \frac{\tilde{p}(\boldsymbol{\theta}_l,\nu_l)}{\tilde{q}(\boldsymbol{\theta}_l,\nu_l)} = \frac{1}{p(\mathbf{X}|\boldsymbol{\theta}=\boldsymbol{\theta}_l,\nu=\nu_l)}.$$

Hence,

$$\begin{aligned}
\int f(\boldsymbol{\theta},\nu)\, \tilde{p}(\boldsymbol{\theta},\nu)\, d\boldsymbol{\theta} d\nu &\approx \sum_{l=1}^{L} \frac{\tilde{w}_l}{\sum_i \tilde{w}_i} f(\boldsymbol{\theta}_l,\nu_l) \\
&= \sum_{l=1}^{L} \frac{\frac{1}{p(\mathbf{X}|\boldsymbol{\theta}=\boldsymbol{\theta}_l,\nu=\nu_l)}}{\sum_i \frac{1}{p(\mathbf{X}|\boldsymbol{\theta}=\boldsymbol{\theta}_i,\nu=\nu_i)}} p(\mathbf{X}|\boldsymbol{\theta}=\boldsymbol{\theta}_l,\nu=\nu_l) \\
&= \frac{L}{\sum_l p(\mathbf{X}|\boldsymbol{\theta}=\boldsymbol{\theta}_l,\nu=\nu_l)^{-1}}.
\end{aligned}$$

V. Prediction

Given the samples from the posterior distribution $p(\mathbf{z},\boldsymbol{\theta},\mathbf{W},\nu|\mathbf{X},\boldsymbol{\eta}_K)$, we can predict values at a future time point of a new test series. With slight abuse of notation, suppose the test series is $\mathbf{x} = \{\mathbf{x}_1,\ldots,\mathbf{x}_T\}$ with Δt_j being the time interval between $\mathbf{x}_j$ and $\mathbf{x}_{j+1}$. Given parameters $\boldsymbol{\theta},\nu$ and the first j points of a series, the probability density function of the $(j+1)$th time point is

$$\begin{aligned}
p(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu) &= \sum_k p(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},z=k,\boldsymbol{\theta},\nu)\, p(z=k|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu) \\
&= \sum_k p\left(\mathbf{x}_{j+1}|\mathbf{x}_j,\mathbf{A}_k,\sigma_k^2,\nu\right) p(z=k|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu),
\end{aligned}$$

where $\mathbf{x}_{1:j} = \{\mathbf{x}_1,\ldots,\mathbf{x}_j\}$, $p\left(\mathbf{x}_{j+1}|\mathbf{x}_j,\mathbf{A}_k,\sigma_k^2,\nu\right)$ is given by

$$p\left(\mathbf{x}_{j+1}|\mathbf{x}_j,\mathbf{A}_k,\sigma_k^2,\nu\right) = \mathcal{T}\left(\mathbf{x}_{j+1}|\mathbf{x}_j + \Delta t_j \mathbf{A}_k \mathbf{x}_j, \Delta t_j^2 \sigma_k^2 \mathbf{I}_D, \nu\right) \tag{S19}$$

and

$$p\left(z=k|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu\right)=\frac{\pi_k p\left(\mathbf{x}_{1:j}|\mathbf{A}_k,\sigma_k^2,\nu\right)}{\sum_l \pi_l p\left(\mathbf{x}_{1:j}|\mathbf{A}_l,\sigma_l^2,\nu\right)}. \tag{S20}$$

Then the expected value of $\mathbf{x}_{j+1}$ given the first j time points and $\boldsymbol{\theta},\nu$ is easily obtained as

$$\begin{aligned}\mathbb{E}\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu\right)&=\int \mathbf{x}_{j+1}p\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu\right)d\mathbf{x}_{j+1}\\&=\sum_k p\left(z=k|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu\right)\left(\mathbf{x}_j+\Delta t_j\mathbf{A}_k\mathbf{x}_j\right).\end{aligned} \tag{S21}$$

Using Eq. (S21), we can predict $\mathbf{x}_{j+1}$ given a point estimate of $\boldsymbol{\theta},\nu$ or the full posterior.

If we have a point estimate $\hat{\boldsymbol{\theta}},\hat{\nu}$ of the model parameters, e.g. $\hat{\boldsymbol{\theta}}=\mathbb{E}\left(\boldsymbol{\theta}|\mathbf{X},\boldsymbol{\eta}_K\right)$, $\hat{\nu}=\mathbb{E}\left(\nu|\mathbf{X},\boldsymbol{\eta}_K\right)$ from the inference algorithm, we can directly plug it into (S21) and use the expected value as prediction. If we have a full posterior $p\left(\boldsymbol{\theta},\nu|\mathbf{X},\boldsymbol{\eta}_K\right)$, we can do it by integrating $\boldsymbol{\theta},\nu$ out. Considering

$$p\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\mathbf{X},\boldsymbol{\eta}_K\right)=\int p\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu\right)p\left(\boldsymbol{\theta},\nu|\mathbf{X},\boldsymbol{\eta}_K\right)d\boldsymbol{\theta}d\nu,$$

the expected value given the training data is

$$\begin{aligned}\mathbb{E}\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\mathbf{X},\boldsymbol{\eta}_K\right)&=\int \mathbf{x}_{j+1}p\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\mathbf{X},\boldsymbol{\eta}_K\right)d\mathbf{x}_{j+1}\\&=\int p\left(\boldsymbol{\theta},\nu|\mathbf{X},\boldsymbol{\eta}_K\right)\int \mathbf{x}_{j+1}p\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\boldsymbol{\theta},\nu\right)d\mathbf{x}_{j+1}d\boldsymbol{\theta}d\nu\\&\approx\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left(\mathbf{x}_{j+1}|\mathbf{x}_{1:j},\boldsymbol{\theta}=\boldsymbol{\theta}_l,\nu=\nu_l\right),\end{aligned}$$

which averages the expected values from the generated parameter samples by reusing (S21).

VI. Results

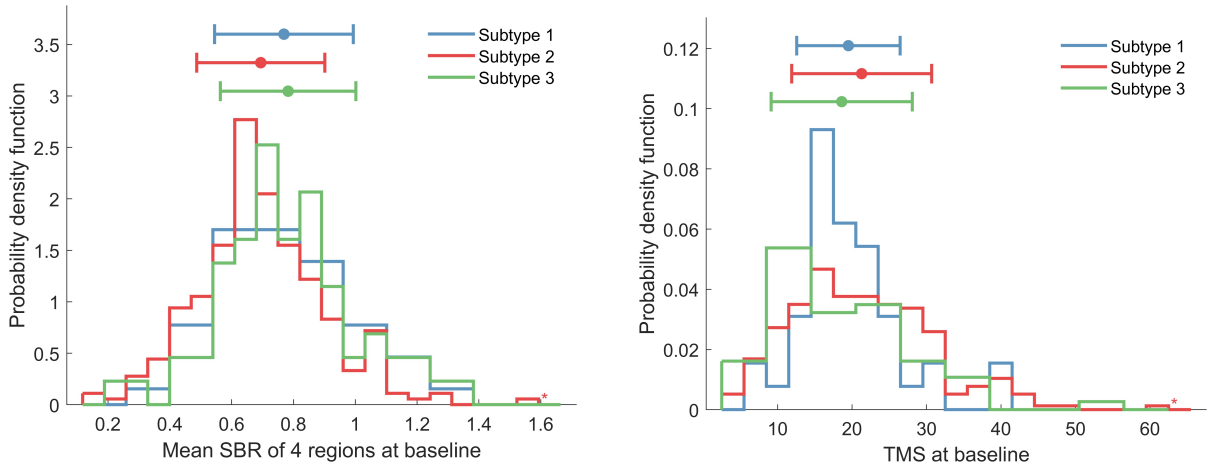
A. Pre-processing

Recall that the PPMI image data are pre-processed in two steps: first, subjects with only one scan are excluded; and second, subjects with misregistered images are excluded. The IDs of the excluded subjects are listed below.

- **Single scan** (67 in total): 3006, 3025, 3026, 3081, 3127, 3129, 3133, 3164, 3167, 3177, 3210, 3232, 3236, 3279, 3280, 3281, 3282, 3284, 3285, 3288, 3289, 3290, 3311, 3314, 3322, 3330, 3331, 3332, 3333, 3376, 3413, 3447, 3501, 3510, 3533, 3534, 3535, 3536, 3618, 3623, 3626, 3628, 3632, 3633, 3663, 3764, 3800, 3827, 3833, 3837, 3858, 3863, 3867, 3958, 3962, 3971, 4003, 4006, 4016, 4017, 4061, 4062, 4069, 4075, 4097, 4136, 4137.
- **Misregistration** (17 in total): 3360, 3407, 3419, 3420, 3421, 3422, 3434, 3451, 3455, 3557, 3605, 3711, 3791, 3953, 3972, 4121, 4135.

B. The three subtypes

We provide additional results to show that the subtypes do not correspond to different stages. If the subtypes correspond to different stages of a single set of disease trajectories, then the baseline SBR and TMS distributions in the subtypes should show a separation. These distributions are shown as histograms in Fig. S3 with the mean $\pm$ standard deviation bars on the top. The figure clearly shows that there is no systematic drift in the histograms to make such a separation. This is not surprising given the properties of our model. First, since the $\mathbf{A}$ matrices are very different (see Table II in the main paper), trajectories from different subtypes cannot be obtained by time transformations of each other. Second, assuming different subtypes being at different stages of a single set of trajectories would imply a single subtype model fitting the data given the semi-group property, which is not true as shown in Fig. 4 and Fig. 10 in the main paper.



(a) Histograms of mean SBR at baseline

(b) Histograms of TMS at baseline

Fig. S3. Histograms of mean SBR (a) and TMS (b) at baseline for the three subtypes. Mean SBR is calculated by averaging the SBRs of the 4 regions. Bars indicating mean $\pm$ std are shown above the histograms.

REFERENCES

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] E. B. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[3] J. M. Bernardo and A. F. Smith, *Bayesian theory*. John Wiley & Sons, 2009.

[4] W. R. Gilks and P. Wild, "Adaptive rejection sampling for gibbs sampling," *Applied Statistics*, pp. 337–348, 1992.

[5] A. Jasra, C. C. Holmes, and D. A. Stephens, "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, pp. 50–67, 2005.

[6] D. D. Walker and E. K. Ringger, "Model-based document clustering with a collapsed gibbs sampler," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 704–712.