Bayesian Longitudinal Modeling of Early Stage Parkinson's Disease Using DaTscan Images

Yuan Zhou[1(✉)] and Hemant D. Tagare[1,2]

[1] Department of Radiology and Biomedical Imaging, Yale University,
New Haven, CT 06510, USA
zhouyuanzxcv@gmail.com
[2] Department of Biomedical Engineering, Yale University,
New Haven, CT 06520, USA

Abstract. This paper proposes a disease progression model for early stage Parkinson's Disease (PD) based on DaTscan images. The model has two novel aspects: first, the model is fully coupled across the two caudates and putamina. Second, the model uses a new constraint called *model mirror symmetry* (MMS). A full Bayesian analysis, with collapsed Gibbs sampling using conjugate priors, is used to obtain posterior samples of the model parameters. The model identifies PD progression subtypes and reveals novel fast modes of PD progression.

Keywords: Parkinson's Disease · Disease progression model · Bayesian analysis

1 Introduction

Parkinson's Disease (PD) is a common neurodegenerative disease characterized by loss of dopaminergic neurons, and accompanied by progressively worsening clinical motor and non-motor symptoms. PD is also a heterogeneous disease; it exhibits vastly different rates of progression in different subjects.

DaTscan imaging is the commercial name for SPECT imaging with 123 I-FP-CIT. DaTscans measure the local density of presynaptic dopamine transporters (DaT). Dopaminergic neural loss decreases DaT density and is visible as signal loss in DaTscan images. Our goal is to model the progression of PD as it manifests in DaTscans. We use the Parkinson's Progression Marker Initiative (PPMI) dataset, described in more details in Sect. 2.

For quantitative analysis, intensity values in every voxel of a DaTscan are converted to Striatal Binding Ratios (SBR). 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 the occipital lobe [8]. SBR is a measure of DaT density in the voxel. The sum or the mean of the SBR in a brain region is taken as a measure of DaT density in the region.

DaTscan and PD characteristics that are important in modeling early stage PD are listed below. Our model takes these characteristics into account:

1. *Disease Stages*: PD progresses along stages called Braak Stages [3]. Early stage PD affects the striatum, with the putamen affected more than the caudate.
2. *Coupled Progress*: Early stage PD is also asymmetric; one brain hemisphere is affected more than the other hemisphere [7]. As the disease progresses, the disease becomes symmetric, demonstrating a coupling between ROIs.
3. *Exponential Loss*: SBR loss in the striatum due to PD progression is approximately exponential [1,6].
4. *Heterogeneity*: PD is a heterogeneous disease, with patients progressing at different rates and exhibiting variable clinical symptoms. Different PD subtypes have been proposed using clinical symptoms (MDS-UPDRS ratings), e.g. rigidity-dominant vs. tremor-dominant [4], or early-onset vs. late-onset [13]. To the best of our knowledge, the existence of image-based progression subtypes has not yet been reported in the literature. One of our goals is to investigate such subtypes.

Besides these previously known properties, we identify a new property called *model mirror symmetry* (MMS) which is critical in reducing the dimension of the model:

1. *Model Mirror Symmetry*: Progression from the asymmetric state to the symmetric state does not depend on the hemisphere that the disease originally affected. This implies that the progression model should remain invariant if left hemisphere voxels (or regions) are swapped with right hemisphere voxels (or regions). We call this *model mirror symmetry.*

The above properties suggest that PD progression in DaTscan images can be modeled as a mixture of linear dynamical systems (LDS), where the transition matrices of the dynamical systems are constrained to be centrosymmetric. This is explained further in Sect. 3.1. We fit the model using a Bayesian methodology: collapsed Gibbs sampling (Sect. 4.1) and Bayesian model selection for the optimal number of mixture components (Sect. 4.2). Sampling avoids entrapment in spurious local maxima, a common problem in maximum-likelihood methods and the EM algorithm.

Modeling brain disease progression is a relatively new problem. Using prior knowledge (about toxic protein aggregation followed by transmission along neuronal pathways) graph theoretic approaches have been proposed for Alzheimer's disease progression, e.g. [14]. Other approaches use discrete event models or generalized linear time models for disease progression [5,9]. Regression of image features with longitudinal clinical scores are also used for disease progression modeling [17]. Most of these methods are applied to MRI images of Alzheimer's disease. DaTscans have a lower resolution than MRI images and contain very different kind of information (DaT density rather than structural information). MRI-based models are not directly applicable to DaTscans.

2   Data

The PPMI dataset has 449 early stage PD subjects. The subjects are imaged at baseline and approximately at 1, 2, 4, and 5 years out. Not all subjects have a complete time series. And the time series for different subjects has slightly different timings. The PPMI DaTscan images are registered to the MNI space; but, we did find some misregistered images in the data. We preprocessed PPMI data to eliminate subjects with only a single scan, and subjects with misregistered images (using a simple non-parametric correlation test). In the end, data from 365 subjects survived, and entered the analysis. Of these subjects, 320 subjects had 3 or more scans, 130 subjects had 4 or more scans, and 3 subjects had 5 scans.
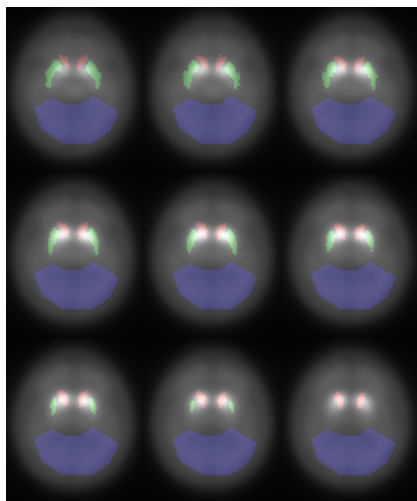

Fig. 1. Caudates (red), putamina (green), and occipital lobe (blue) ROIs superimposed on the mean baseline DaTscan image (36th - 44th slices shown). (Color figure online)

Since early stage PD mostly affects the caudate and putamen, we modeled the mean SBR in the two caudates and putamina. The MNI atlas was used to identify caudates and putamina, and manual ROI (similar to the one in [18]) was created for the occipital lobe (Fig. 1 shows these regions on 9 slices). Figure 2 shows mean SBR time series for the 365 surviving subjects. The time series for left and right caudates and the left and right putamina are shown in different plots. For each subject, the sequence of the mean SBRs are shown as blue, orange, yellow, and purple arrows which correspond to data from time 1 to 2, 2 to 3, 3 to 4, and 4 to 5 respectively. Note that, modulo noise, the data are symmetric around the $45°$ line. The figure shows that any model that fits well to this data should remain invariant if the "left" and "right" labels are swapped. This is model mirror symmetry.

3   Disease Progression Model

3.1   Time Series Model

The four mean SBRs are arranged in a 4 by 1 vector $\mathbf{x}$ in the order: left caudate (LC), left putamen (LP), right putamen (RP), and right caudate (RC). Note that $\mathbf{x} \in \mathbb{R}^D$ with $D = 4$. This vector is the feature we extract from every image, and our goal is to model the time series of this feature for all subjects.

Keeping in mind that SBR is known to decay exponentially in PD, a continuous time evolution model for $\mathbf{x}$ is the linear differential equation: $\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{D \times D}$. Since $\mathbf{A}$ is not required to be diagonal, this model captures coupled progression between the caudates and putamina.
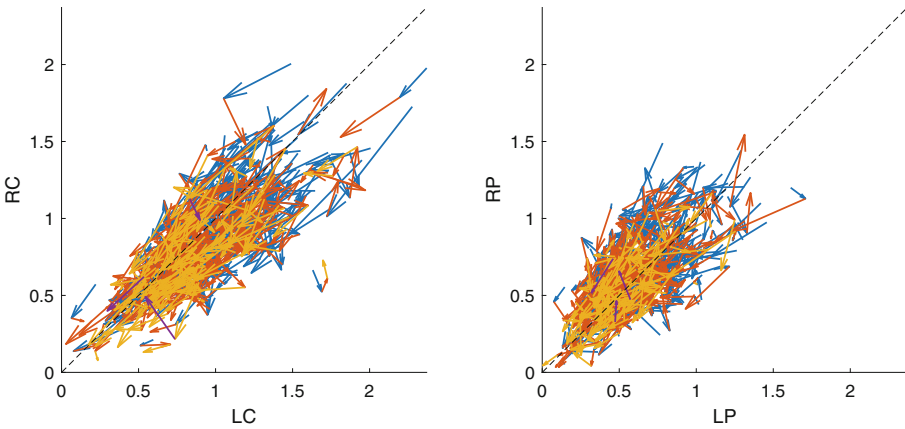

Fig. 2. Time series of the mean SBR in the caudates and putamina. Blue, orange, yellow, purple arrows represent the vectors from time 1 to 2, 2 to 3, 3 to 4, and 4 to 5 respectively. LC = left caudate, RC = right caudate, LP = left putamen, RP = right putamen. (Color figure online)

Next, we account for the fact that the longitudinal time series for each subject is discrete and not necessarily uniformly sampled in time. Letting $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i}$ denote the discrete time series for the ith subject, and letting Δt_{ij} denote the difference in time between the jth and the $(j+1)$th imaging times, the differential equation can be discretized as $\frac{\mathbf{x}_{i,j+1} - \mathbf{x}_{ij}}{\Delta t_{ij}} = \mathbf{A}\mathbf{x}_{ij} + \boldsymbol{\epsilon}_{ij}$, where we assume that the noise $\boldsymbol{\epsilon}_{ij}$ has a Gaussian distribution, $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_D\right)$. Denoting the entire time series for the subject as $\mathbf{x}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT_i}\}$, we have

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

(1)

where $\mathbf{v}_{ij} = \frac{\mathbf{x}_{i,j+1}-\mathbf{x}_{ij}}{\Delta t_{ij}}$ and $p(\mathbf{x}_{i1}) = \mathcal{N}(\mathbf{x}_{i1}|\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ estimated from all the $\mathbf{x}_{i1}$.

Next, we impose MMS:

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* $i = 1, \cdots, D$, *and* $(\mathbf{x})_k$ *refers to the kth component of vector* $\mathbf{x}$. *A* $D \times D$ *matrix* $\mathbf{A}$ *is centrosymmetric if* $\pi(\mathbf{Ax}) = \mathbf{A}\pi(\mathbf{x})$ *for all* $\mathbf{x}$, *where* π *is a symmetric permutation.*

Because of how LC, LP, RP, and RC are arranged in $\mathbf{x}$, swapping right and left hemisphere ROIs corresponds to applying a symmetric permutation to $\mathbf{x}$. Thus, MMS implies that $\mathbf{A}$ should be a centrosymmetric matrix. The set of all $D \times D$ centrosymmetric matrices forms a subspace of all $D \times D$ matrices. This subspace has dimension $\lceil D^2/2 \rceil$, hence the number of parameters for fitting the data is reduced by approximately half. Also, if a centrosymmetric matrix has distinct eigenvalues, its eigenvectors are either symmetric or skew-symmetric [16]. We will see the importance of this property when interpreting results.

Let $\mathbf{e}_j$ be a vector of the form $[\ldots, 1, \ldots, 1, \ldots]^T$ where 1 only appears at the jth position and the $(D^2 - j + 1)$th position and the other components are zero. Then, any centrosymmetric matrix $\mathbf{A}$ can be represented as $\text{vec}(\mathbf{A}) = \mathbf{Ea}$ where $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_{\lceil D^2/2 \rceil}]$ is the basis and $\mathbf{a}$ contains the coordinates for expressing $\text{vec}(\mathbf{A})$ in this basis. Hence, the density $p(\mathbf{x}_i|\mathbf{A}, \sigma^2)$ in (1) can be reorganized as $p(\mathbf{x}_i|\mathbf{a}, \sigma^2)$:

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

where

$$h_i = \frac{\mathcal{N}(\mathbf{x}_{i1}|\mathbf{0}, \boldsymbol{\Sigma})}{(2\pi)^{(T_i-1)D/2} \prod_{j=1}^{T_i-1} \Delta t_{ij}^D}, \quad \epsilon_i = \sum_{j=1}^{T_i-1} \mathbf{v}_{ij}^T \mathbf{v}_{ij}, \quad d_i = (T_i - 1)D,$$

$$\boldsymbol{\Lambda}_i = \mathbf{E}^T \left(\sum_{j=1}^{T_i-1} \mathbf{x}_{ij}\mathbf{x}_{ij}^T \otimes \mathbf{I}_D \right) \mathbf{E}, \quad \boldsymbol{\mu}_i = \mathbf{E}^T \sum_{j=1}^{T_i-1} \text{vec}\left(\mathbf{v}_{ij}\mathbf{x}_{ij}^T\right). \tag{3}$$

Note that $h_i, \boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \epsilon_i$ are functions of $\mathbf{x}_i$, but the dependence is not explicitly shown to simplify notation. For Bayesian analysis we will need conjugate priors for $\mathbf{a}, \sigma^2$ and a posterior predictive probability density for $\mathbf{x}$. The quadratic form for $\mathbf{a}$ in Eq. (2) is either strictly positive definite or positive semi-definite depending on the length of the time series. In either case, the density belongs to the exponential family and a conjugate prior is available [2]. For completeness, we state the conjugate prior and the posterior for the specific form in (2):

Theorem 1. *Suppose* $\mathbf{x}_i$, $i = 1, \ldots, N$ *are conditionally independent random variables, with densities given by Eq. (2) where* $\boldsymbol{\Lambda}_i$ *is symmetric. Then,* $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ *has a conjugate prior in the form of normal-inverse-gamma (NIG)*

$$p(\mathbf{a}, \sigma^2|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0) = NIG(\mathbf{a}, \sigma^2|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0)$$
$$= \mathcal{N}(\mathbf{a}|\boldsymbol{\mu}_0, \sigma^2\boldsymbol{\Lambda}_0^{-1}) IG(\sigma^2|\nu_0, \kappa_0) \tag{4}$$

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

$$p\left(\mathbf{a}, \sigma^2 | \mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\right) = NIG\left(\mathbf{a}, \sigma^2 | \boldsymbol{\mu}_p, \boldsymbol{\Lambda}_p, \nu_p, \kappa_p\right) \tag{5}$$

with

$$\nu_p = \nu_0 + \frac{\sum_i d_i}{2}, \ \boldsymbol{\Lambda}_p = \boldsymbol{\Lambda}_0 + \sum_i \boldsymbol{\Lambda}_i, \ \boldsymbol{\mu}_p = \boldsymbol{\Lambda}_p^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \sum_i \boldsymbol{\mu}_i\right),$$

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

In Eq. (4), $\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0$ are parameters of the prior (hyperparameters). We jointly refer to them as $\boldsymbol{\beta} = \{\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\}$. A direct consequence of Theorem 1 is that the posterior predictive distribution has a closed form:

Corollary 1. *Suppose $\mathbf{x}_{N+1}$ has density given by Eq. (2) where $\boldsymbol{\Lambda}_{N+1}$ is symmetric. Then, the posterior predictive density of $\mathbf{x}_{N+1}$ given $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is*

$$p\left(\mathbf{x}_{N+1}|\mathbf{X}, \boldsymbol{\beta}\right) = \int\int p\left(\mathbf{x}_{N+1}|\mathbf{a}, \sigma^2\right) p\left(\mathbf{a}, \sigma^2|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0, \kappa_0\right) d\mathbf{a}d\sigma^2$$

$$= \frac{|\boldsymbol{\Lambda}_p|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}_p + \boldsymbol{\Lambda}_{N+1}|^{\frac{1}{2}}} \frac{\Gamma\left(\nu_p + \frac{d_{N+1}}{2}\right)}{\Gamma\left(\nu_p\right)} \frac{h_{N+1}}{\kappa_p^{\frac{d_{N+1}}{2}}}\left[1 + \frac{Q}{2\kappa_p}\right]^{-\left(\nu_p + \frac{d_{N+1}}{2}\right)} \tag{6}$$

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

3.2 Heterogeneity and Mixture Models

The above model holds for all the subjects satisfying a single differential equation. It does not account for heterogeneity. Heterogeneity implies that different subjects may be modeled by different transition matrices $\mathbf{A}$ (as represented by their coordinates $\mathbf{a}$) and corresponding noise variances σ^2. Assuming that there are K distinct $\mathbf{a}_k$ and σ_k^2, $k = 1, \ldots, K$, then the time series for each subject can be modeled as generated by first picking a latent variable $z_i \in \{1, 2, \ldots, K\}$, $i = 1, \ldots, N$, and then sampling from the distribution $p\left(\mathbf{x}_i|\mathbf{a}_{z_i}, \sigma_{z_i}^2\right)$. The density of $\mathbf{x}_i$ given all $\mathbf{a}_k$ and σ_k^2 is, of course, a mixture model

$$p\left(\mathbf{x}_i|\boldsymbol{\pi}, \{\mathbf{a}_l, \sigma_l^2\}\right) = \sum_{k=1}^{K} p\left(z_i = k|\boldsymbol{\pi}\right) p\left(\mathbf{x}_i|z_i = k, \{\mathbf{a}_l, \sigma_l^2\}\right) = \sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_i|\mathbf{a}_k, \sigma_k^2\right),$$

where $z_i|\boldsymbol{\pi} \sim \text{Cat}\,(\boldsymbol{\pi})$ with $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$. We call this a *mixture of linear dynamical systems* model, or *mixture LDS* for short. We are interested in estimating $\boldsymbol{\theta} = \left\{\boldsymbol{\pi}, \mathbf{a}_k, \boldsymbol{\sigma}_k^2 : k = 1, \ldots, K\right\}$ for this model.

Since $\boldsymbol{\pi}$ has a categorical distribution, for Bayesian analysis we use its conjugate prior, a Dirichlet distribution $\boldsymbol{\pi}|\boldsymbol{\alpha} \sim \text{Dir}\,(\boldsymbol{\alpha})$. For $\mathbf{a}_k, \sigma_k^2$, we use NIG as its conjugate prior, i.e. $\mathbf{a}_k, \sigma_k^2 \sim \text{NIG}\,(\boldsymbol{\beta})$, from Theorem 1. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyperparameters. The mixture LDS model captures all of the characteristics of PD progression listed in Sect. 1.

4   Bayesian Analysis

Bayesian analysis of the above model consists of generating samples from the posterior $p\,(\mathbf{z}, \boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ using MCMC methods, where $\mathbf{z} = (z_1, \ldots, z_N)$ is the vector of latent variables. The overall strategy is to sample from $\mathbf{z}, \boldsymbol{\pi}, \mathbf{a}_k, \sigma_k^2$ sequentially (a.k.a. Gibbs sampling). We use a collapsed Gibbs sampler for sampling $\mathbf{z}$ from $p\,(\mathbf{z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, and then sample the rest from $p\,(\boldsymbol{\theta}|\mathbf{z}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

4.1   Collapsed Gibbs Sampling

To sample $\mathbf{z}$, we integrate out $\boldsymbol{\theta}$ and sample from $p\,(\mathbf{z}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. This is collapsed Gibbs sampling and it is well known that it leads to faster convergence [12,15]. The samples of $\mathbf{z}$ are generated one component at a time based on

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

where $\mathbf{z}_{-i} = \{z_j : j \neq i, j = 1, \ldots, N\}$ (the same for $\mathbf{X}_{-i}$), and $\mathbf{X}_{-i,k} = \{\mathbf{x}_j : j \neq i, z_j = k, j = 1, \ldots, N\}$. Assuming that $\boldsymbol{\alpha} = (\alpha/K, \ldots, \alpha/K)$, the first term in the product is easily shown to be $p\,(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\alpha}) = \frac{N_{-i,k} + \alpha/K}{N + \alpha - 1}$, where $N_{-i,k} = \sum_{j \neq i} \mathbb{I}\,(z_j = k)$ and $\mathbb{I}\,(\cdot) = 1$ if its argument is true and zero otherwise. The second term is calculated from (6) according to Corollary 1.

To sample $\boldsymbol{\theta}$ from $p\,(\boldsymbol{\theta}|\mathbf{z}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, we sample $\boldsymbol{\pi}$ and then $\mathbf{a}_k, \sigma_k^2$. Sampling $\boldsymbol{\pi}$ is straightforward since $p\,(\boldsymbol{\pi}|\mathbf{z}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p\,(\boldsymbol{\pi}|\mathbf{z}, \boldsymbol{\alpha})$ and $\boldsymbol{\pi}|\mathbf{z}, \boldsymbol{\alpha} \sim \text{Dir}(\alpha/K + N_1, \ldots, \alpha/K + N_K)$. Sampling $\mathbf{a}_k, \sigma_k^2$ is done by sampling from NIG in (5) following Theorem 1 since $p\,(\mathbf{a}_k, \sigma_k^2|\mathbf{z}, \mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p\,(\mathbf{a}_k, \sigma_k^2|\mathbf{X}_k, \boldsymbol{\beta})$ where $\mathbf{X}_k = \{\mathbf{x}_i : z_i = k, i = 1, \ldots, N\}$.

We use weak priors by setting the hyperparameters to $\alpha = 1$ and $\boldsymbol{\beta} = (\mathbf{0}, 10^{-8}\mathbf{I}_{\lceil D^2/2 \rceil}, 10^{-3}, 10^{-3})$. For the results reported below, we use 3000 MCMC iterations with initial 40% samples discarded.

4.2   Choosing the Number of Clusters

Our model requires K, the number of clusters to be chosen a priori. We choose the number of clusters using Bayesian model selection as well as cross validation [10,11]. For cross validation, we divide the dataset into 10 subsets (10-fold cross

validation). Taking each subset as test set, we use the remaining data as training set to infer the parameters $\boldsymbol{\theta}$ and then evaluate the log-likelihood of the test set.

For Bayesian model selection, we denote $\mathbf{M}_K$ for the model with K components, and $\mathcal{M} = \{\mathbf{M}_1, \ldots, \mathbf{M}_{K_{\max}}\}$. Assuming $p(K) \propto$ constant on $K = 1, \ldots, K_{\max}$, we have $p(K|\mathbf{X}, \mathcal{M}) \propto p(K)\, p(\mathbf{X}|\mathcal{M}, K) \propto p(\mathbf{X}|\mathcal{M}, K) = p(\mathbf{X}|\mathbf{M}_K)$. Then, finding the optimal $\hat{K} = \arg\max_K p(K|\mathbf{X}, \mathcal{M})$ is equivalent to finding the maximum of $p(\mathbf{X}|\mathbf{M}_K)$ with respect to K. The density $p(\mathbf{X}|\mathbf{M}_K)$ is evaluated by the integral

$$p(\mathbf{X}|\mathbf{M}_K) = \int p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{M}_K)\, d\boldsymbol{\theta}. \tag{7}$$

The Gibbs sampler discussed above generates samples from $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Using $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{M}_K)$ as a proposal distribution and the generated samples of $\boldsymbol{\theta}$, we evaluate the integral in (7) by importance sampling.

5   Results

Recall from Sect. 2 that time series data from 365 subjects survived preprocessing. This is the dataset that we analyze. First we determine the number of clusters, then we use the determined number of clusters to sample from the posterior of the model parameters. Finally, we interpret the mean of the posterior.

5.1   Determine the Number of Clusters

Figure 3 shows plots for Bayesian model selection (with $K_{\max} = 10$) as well as 10-fold cross validation as a function of number of clusters. In each plot, the blue curve shows the value of the log-likelihood vs. the number of clusters. The log-likelihood behavior is similar in both plots, consistent with the common understanding that empirical Bayesian and cross validation have similar results. We observed that as the number of clusters increased, empty clusters were created. In both plots, the orange curve shows the number of clusters that were not empty. Note that the log-likelihood increases with number of clusters and approximately plateaus from 4 clusters onwards. The one exception is when we initialize with 9 clusters, where Bayesian model selection gives 5 non-empty clusters with a slightly higher log-likelihood value. For cross validation, the average number of final number of clusters (over the 10 folds) remains around 4. Considering these results, we use 4 clusters for the dataset.

5.2   Interpreting the Model

Having chosen the number of clusters, Table 1 shows the mean of the posterior distribution for $\boldsymbol{\theta}$, which we take as the "fit" of the model to the data. The table does not directly show matrices $\mathbf{A}_k$, instead it shows the eigenvalues and eigenvectors of the matrices. These are easier to interpret, as we discuss below. Figure 4 shows the raw SBR time series for subjects for each cluster.
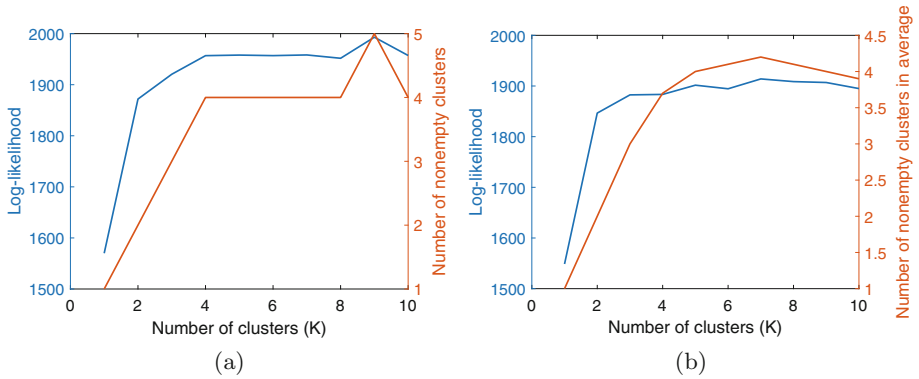
(a)   (b)

Fig. 3. Model selection using Bayesian (a) and cross validation (b). The y-axis has two scales corresponding to log-likelihood value (blue curve) and number of nonempty clusters (orange curve) separately. (Color figure online)

Table 1. $\mathbb{E}(\boldsymbol{\theta})$ from averaging the generated samples.

Class	$k = 1$				$k = 2$				$k = 3$				$k = 4$			
π_k	0.382				0.072				0.397				0.148			
σ_k	0.069				0.268				0.128				0.087			
λ^{a}	-0.3	-0.2	-0.2	-0.1	-0.7	-0.4	-0.4	-0.1	-0.4	-0.3	-0.1	-0.1	-0.6	-0.4	-0.2	-0.2
$\mathbf{u}$	0.4	-0.3	-0.6	-0.6	0.5	0.7	-0.7	-0.6	-0.4	0.5	-0.6	0.6	0.5	-0.2	-0.6	0.6
	-0.6	0.7	-0.3	-0.4	-0.5	-0.1	-0.2	-0.4	0.6	-0.5	-0.4	0.4	-0.5	0.7	-0.4	0.4
	0.6	0.7	0.3	-0.4	0.5	-0.1	0.2	-0.4	-0.6	-0.5	0.4	0.4	0.5	0.7	0.4	0.4
	-0.4	-0.3	0.6	-0.6	-0.5	0.7	0.7	-0.6	0.4	0.5	0.6	0.6	-0.5	-0.2	0.6	0.6

[a] **A** is represented by its eigenvalues λ and eigenvectors $\mathbf{u}$. Eigenvectors are shown to first significant digit to conserve horizontal space.
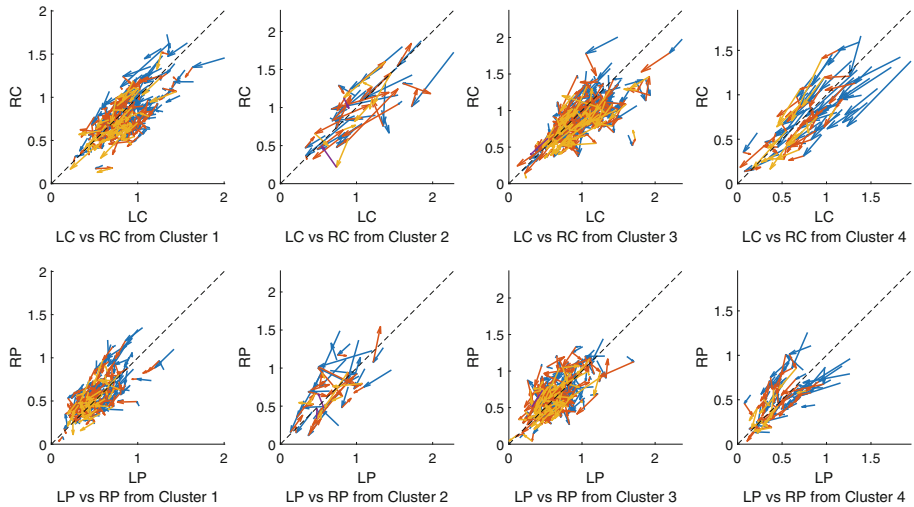

Fig. 4. Four clusters from Gibbs sampling on the whole data.

To interpret the model, first note that cluster 2 has the largest σ and the smallest π_k. Also, the SBR trajectories for this cluster (Fig. 4) are more disorganized than SBR trajectories for other clusters. Quite likely, this cluster represents additional outliers in the data. We focus on the remaining clusters. For these clusters, the eigenstructure of $\mathbf{A}_k$'s is particularly illuminating: All eigenvalues of $\mathbf{A}_k$'s are real and negative, showing that all linear combinations of SBRs decrease with time. Cluster 1 has the least negative eigenvalues, while cluster 4 has the most negative eigenvalues. Cluster 3 is intermediate. Thus cluster 4 captures the fastest evolving subjects, cluster 1 the slowest, and cluster 3 the intermediate. The eigenvalues for cluster 4 are almost twice the eigenvalues for cluster 1, suggesting that DaT loss proceeds approximately twice as fast in cluster 4. Further evidence for this comes directly from the data. Figure 5 shows the histograms of the magnitude of initial velocities ($\|\mathbf{v}_{i1}\|$) as calculated from the raw SBRs for each cluster. The medians of the histograms are 0.17, 0.61, 0.27, 0.36 respectively, which verifies our speed analysis.

The solution to a LDS $\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}$ is determined by the eigenvalues $\{\lambda_i\}$ and eigenvectors $\{\mathbf{u}_i\}$ of $\mathbf{A}$. The eigenvector determines the subspace in which the time series proceeds with the eigenvalue as the time constant. Recall from Sect. 3.1 that the eigenvectors of centrosymmetric matrices are either symmetric or skew-symmetric. This symmetry or antisymmetry represents symmetry or asymmetry of disease progression. To see this, suppose $\mathbf{u}$ is a skew-symmetric eigenvector, say $\mathbf{u} = (a, b, -b, -a)$. Because of how we have arranged the caudates and putamina in $\mathbf{x}$, this suggests that $a \times \mathrm{LC} + b \times \mathrm{LP} - (b \times \mathrm{RP} + a \times \mathrm{RC})$ goes to zero with speed λ. In other words, a skew-symmetric eigenvector and its eigenvalue capture how the asymmetry (of the linear combination $a \times \mathrm{Caudate} + b \times \mathrm{Putamen}$) between the two hemispheres decreases to zero. Similarly, a symmetric eigenvector and its eigenvalue capture how the symmetry (i.e. the weighted "mean" of the SBRs) decreases to zero. Finally, recall from Sect. 1 that the difference between SBR uptake in the caudate and putamen reflects the extent of PD in each brain hemisphere. Thus if a and b have opposite signs, then a skew-symmetric eigenvector represents asymmetry in the disease across the hemispheres, while a symmetric eigenvector represents the mean disease in both hemispheres.
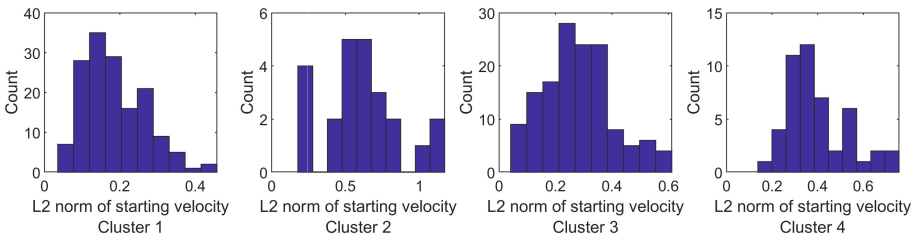

Fig. 5. Histogram of starting velocity magnitude for each cluster. The medians for the 4 clusters are 0.17, 0.61, 0.27, 0.36 respectively.

Applying this interpretation to each cluster of Table 1 gives the following: The most negative eigenvalue in each cluster has a skew-symmetric eigenvector. This implies that the loss of asymmetry of disease across hemispheres has the fastest speed of *all possible linear combinations of SBRs*. Also note that the first eigenvector in clusters 1, 3 and 4 is numerically quite similar. The next significant eigenvalue in each cluster has a symmetric eigenvector. The last eigenvector in all clusters represents the "mean" of all four regions and has the slowest eigenvalue, hence the mean (or net) SBR in the four regions is the slowest indicator of early stage PD.

6   Discussion and Conclusion

The results presented above show that PD progression as manifest in DaTscans is heterogeneous, with one subtype (cluster 4) progressing almost twice as fast as the slowest subtype (cluster 1). This is the first significant finding of this paper. The second, and potentially more interesting, finding is that within each subtype (cluster) the change in SBR asymmetry across hemispheres has a faster time constant than the change in mean SBR across hemispheres. Moreover, the change in asymmetry is the fastest change among all linear combinations of SBRs. Whether this finding can be utilized to create a sensitive disease progression index is an open question. We plan to investigate this in the future. Such a disease progression index is likely to have significant implications for clinical trials that use DaTscans.

Model mirror symmetry is also a novel idea with broader implications. Extending the MMS idea to high dimensional data (more structures, voxel level rather than ROI modeling) could be challenging as the dimension increases. One possible route to higher dimensions is to require **A** to be sparse or low rank. This will be the focus of forthcoming research.

In conclusion, this paper proposes a mixture LDS model for PD progression in DaTscans. The model reveals that progression in DaTsans is heterogeneous, with a significant range of progression time constants. The model also suggests that the change in asymmetry may be a more sensitive index of PD disease progression.

References

1. Au, W.L., Adams, J.R., Troiano, A., Stoessel, A.J.: Parkinson's disease: in vivo assessment of disease progression using positron emission tomography. Mol. Brain Res. **134**, 24–33 (2005)
2. Bernardo, J.M., Smith, A.F.: Bayesian Theory. Wiley, Hoboken (2009)
3. Braak, H., Tredici, K.D., Rub, U., de Vos, R.A., Steur, E.N.J., Braak, E.: Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol. Aging **24**(2), 197–211 (2003)
4. Eggers, C., Kahraman, D., Fink, G.R., Schmidt, M., Timmermann, L.: Akinetic-rigid and tremor-dominant Parkinson's disease patients show different patterns of FP-CIT single photon emission computer tomography. Mov. Disord. **26**(3), 416–423 (2011)
5. Fonteijn, H.M., et al.: An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage **60**(3), 1880–1889 (2012)
6. Hilker, R., et al.: Nonlinear progression of Parkinson disease as determined by serial positron emission tomographic imaging of striatal fluorodopa F 18 activity. Archiv. Neurol. **62**(3), 378–382 (2005)
7. Hoehn, M.M., Yahr, M.D.: Parkinsonism: onset, progression and mortality. Neurology **17**, 427–442 (1967)
8. Innis, R.B., et al.: Consensus nomenclature for in vivo imaging of reversibly binding radioligands. J. Cereb. Blood Flow Metab. **27**(9), 1533–1539 (2007)
9. Jedynak, B.M., et al.: A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort. Neuroimage **63**(3), 1478–1486 (2012)
10. Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 773–795 (1995)
11. McLachlan, G.J., Rathnayake, S.: On the number of components in a Gaussian mixture model. Wiley Interdisc. Rev.: Data Min. Knowl. Discov. **4**(5), 341–355 (2014)
12. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge (2012)
13. Quinn, N., Critchley, P., Marsden, C.D.: Young onset Parkinson's disease. Mov. Disord. **2**, 73–91 (1987)
14. Raj, A., Powell, F.: Models of network spread and network degeneration in brain disorders. Biol. Psychiatry: Cogn. Neurosci. Neuroimaging **3**, 788–797 (2018)
15. Sudderth, E.B.: Graphical models for visual object recognition and tracking. Ph.D. thesis, Massachusetts Institute of Technology (2006)
16. Weaver, J.R.: Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors. Am. Math. Mon. **92**(10), 711–717 (1985)
17. Zhou, J., Liu, J., Narayan, V.A., Ye, J.: Modeling disease progression via fused sparse group lasso. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1095–1103. ACM (2012)
18. Zubal, I.G., Early, M., Yuan, O., Jennings, D., Marek, K., Seibyl, J.P.: Optimized, automated striatal uptake analysis applied to SPECT brain scans of Parkinson's disease patients. J. Nucl. Med. **48**(6), 857–864 (2007)