Abstract

Disease modelling is an essential tool to describe disease progression and its heterogeneity across patients. Usual approaches use continuous data such as biomarkers to assess progression. Nevertheless, categorical or ordinal data such as item responses in questionnaires also provide insightful information about disease progression. In this work, we propose a disease progression model for ordinal and categorical data. We built it on the principles of disease course mapping, a technique that uniquely describes the variability in both the dynamics of progression and disease heterogeneity from multivariate longitudinal data. This extension can also be seen as an attempt to bridge the gap between longitudinal multivariate models and the field of Item Response Theory. Application to the Parkinson's Progression Markers Initiative cohort illustrates the benefits of our approach: a fine-grained description of disease progression at the item level, as compared to the aggregated total score, together with improved predictions of the patient's future visits. The analysis of the heterogeneity across individual trajectories highlights known disease trends such as tremor dominant (TD) or postural instability and gait difficulties (PIGD) subtypes of Parkinson's disease.

Multivariate disease progression modelling with longitudinal ordinal data

Pierre-Emmanuel Poulet, Stanley Durrleman[1]

[1]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière

1 Introduction

Neurodegenerative diseases become more and more prevalent with the global aging of the developed countries. Understanding these diseases is key for monitoring their progression and designing efficient interventions. Disease progression models are statistical models that describe and predict disease progression at the population or individual level. They provide insights into how the disease progresses on average, and the range of disease trajectories across patients. They can then be used to forecast progression with the aim to support medical decisions or stratify patients according to their progression profiles in clinical studies. These models are estimated from observational data where patients at different stages of the disease are followed over several years. Data in these longitudinal studies include clinical assessments of different functions such as cognition, behavior, mood or motor skills, biological measurements and image-derived biomarkers such as volumes of brain structures.

Neurodegenerative diseases and other age-related disorders raise a specific challenge: the age at onset is usually unknown since the disease progressively deviates from a trajectory of normal aging, and there is no temporal marker of progression whereas the disease starts at various ages and different pace for different patients. Disease modeling should therefore establish a matching between the age at which the patient is observed to a disease stage in a common normalized timeline.

We may describe disease progression with a single measurement, such as the Alzheimer's disease assessment scale [17] or the unified Parkinson's rating scale (MDS-UPDRS) [9] for instance. In this case, patients can differ only by the dynamics of the progression of the score. This variability

[0]**Abbreviations: Parkinson's Progression Markers Initiative (PPMI), Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS), Item Response Theory (IRT)**

may be modeled using a time-warp function matching the real age of the subject to a point in the common disease timeline [6, 7, 29].

These global scores aggregate responses to multiple items assessing an array of functional domains. Here we aim to enrich disease progression models by breaking down aggregate scores into sub-scores or items. In this multivariate analysis, the inter-subject variability is not only described by changes in the dynamics of progression but also by varying values for each item at a given disease stage. Disease course mapping is an approach that is designed to estimate and disentangle these two sources of variability in a longitudinal data set. However, this approach has been designed for continuous measurements. The goal of this paper is to extend this approach to binary or categorical variables. The study of such data constitutes the field of Item Response Theory (IRT), but most former work model the variability of the variables in a cross-sectional setting, therefore missing important information about the progression of the disease.

In IRT, a lot of the concerns of longitudinal models are not taken into account when dealing with non-continuous data. To our knowledge, there is no method based on binary or ordinal items which checks all the following points:

Disease timeline: the model should account for individual variability in ages. Indeed patients can have an early or late onset, and we can refer to them as slow or fast progressors. With an adequate time reparametrization, we can map all individuals to a shared disease timeline where it is easier to compare them. As mentioned before, a common disease timeline helps in understanding the standard disease duration and eventually framing disease stages.

Multivariate: the method should make use of the multiple observations in the patients' records and jointly model them. Features should not be considered independent of each other, learning the dependencies between them is a key part in understanding disease patterns.

Inter-patient variability: disease heterogeneity being a major focus, the model should provide a flexible framework for individual variability. Therefore there should be individual parameters, which are expected to be expressive enough (thus multidimensional) in order to disentangle different disease mechanisms.

In this work we provide a backbone for the longitudinal analysis of binary and ordinal data with the class of mixed-effect models [19], and more specifically a disease course mapping model [30]. We will first review the literature of disease progression models and the few attempts at modelling non-continuous data. Then we will introduce our model in order to account for the three points mentioned above. Finally we showcase the interest of item modelling rather than using aggregated scores in a simple synthetic experiment. The last part of our work is dedicated to applications of the model to a medical cohort of Parkinson's disease.

1.1 Related work: Disease progression modelling

Disease progression models can be divided in several classes. One such class includes time-to-event modelling, especially with SuStaIn [34] which

is a method mixing time-to-event to define disease stages and subtyping. SuStaIn has recently been extended to deal with ordinal data as well [35], however this method is taylored for cross-sectional data and does not account for repeated measurements of individuals. The other common framework for longitudinal modelling is the one of mixed-effects models [19]. Mixed-effect models are based on two kinds of parameters: population parameters called fixed effects, which reflect the average disease progression, and individual parameters called random effects, which account for the individual variability around the average disease progression. Linear mixed-effects models (LMM) are the simplest ones, yet their application to the medical domain is often inadequate, for instance because of the non-linearity of biological mechanisms over long periods of time. Indeed the progression of biomarkers in Alzheimer's disease for instance are hypothesised to be sigmoid-shaped [14]. Therefore many non-linear mixed-effects models (NLMM) have been proposed over the past years. A direct extension of LMM to non-linear is through the use of a link function, like the logit, as in the framework of generalized linear models [24]. Another method used an *ad hoc* statistical model [15], also including time-reparametrization, which is desirable, but the authors still used a univariate framework. This has been extended to a multivariate setting in subsequent work [2, 3]. Multivariate methods are very diverse, including very fine-grained models based on imaging [23], or dynamical models relying on differential equations [32]. New models for disease progression modelling seem to be increasingly infusing knowledge into the models, such as in [27] where the author proposed a non-linear mixed-effect model with latent variables explicitly describing disease stages and individual changes in cognitive ability.

The model upon which our work builds is a disease course mapping model [30] based on a geometric mixed-effect framework. In this setting, the observations lie on a Riemannian manifold and individual trajectories are described by a bundle of parallel curves around a geodesic curve which can be thought of as a population average trajectory. Geometric models constitute a flexible approach to describe complex non-linear multidimensional data while providing interpretable parameters, which is crucial for clinical acceptation.

However such models are usually built on continuous variables such as fluid biomarkers, imaging or aggregated cognitive scores. Dealing with items has been at the heart of Item Response Theory (IRT), from which we will borrow tools to provide an extension of the disease course mapping framework. IRT has long been favoured to deal with this type of data since the first developments in the second half of the 20th century with the works of Lord and his collaborators [20]. IRT has mostly been used for cross-sectional data (even in Parkinson's disease [11]), and extensions to longitudinal data [10, 33] are less common, even though neurodegeneration is best highlighted by the collection of repeated measurements. In IRT each individual i is assumed to have a latent trait θ_i, which becomes a function of time (usually linear) in the case of longitudinal IRT. In disease modelling this latent trait stands for the disease progression score. Longitudinal IRT models have been used on medical cohorts to design new composite scores [1], however the model has too few latent parameters (only one per individual) and thus does little for the understanding of heterogeneity.

Multidimensional IRT models [13] use a multidimensional latent parameter θ_i for each individual, thus allowing for more heterogeneity in the model, but these have not been applied to disease progression modelling to our knowledge.

A recent attempt [5] at bridging IRT with other frameworks has been successful, where a Bayesian non-parametric approach was used to handle the dynamics of longitudinal observations. However the approach taken to model the longitudinal aspect was to discretize time and describe the evolution of the latent trait as unidimensional with an autoregressive scheme. We want to pursue this trend with the introduction of IRT to geometric mixed-effect models.

2 Method

2.1 Disease course mapping

This section will introduce disease course mapping (DCM), a particular disease modeling technique. This model was first mentioned in the original paper [30] and later extended and successfully applied to neurodegenerative diseases in subsequent papers [4, 16]. We will describe the structure of the model and the estimation method. Then we will focus on a specific variation of the model focused on logistic curves, as they will be convenient for the formulation of our ordinal model.

2.1.1 Generic framework: the disease course mapping

Our longitudinal dataset is composed of several measurements $(y_{ijk})_{1\leq i\leq n, 1\leq j\leq N_i, 1\leq k\leq d}$ where d measurements were observed on the patient i at N_i visits. We aggregate these measurements as the coordinates of a single vector in a multidimensional space. t_{ij} is the age of the patient i at the jth visit. Some observations can be missing, but the probabilistic framework of the model allows us to ignore those. The number of visits per patient is variable. In the first setting of this model, observations y_{ijk} are continuous values (not yet integers as for ordinal items). The values y_{ijk} are assumed to be points belonging to a Riemannian manifold $\mathcal{M}$. The model considers that repeated observations in time for a subject i follow a trajectory γ_i on the manifold. Each individual has its own trajectory, which it follows at its own pace. This modelling implies that the choice of the Riemannian manifold and metric defines the possible trajectories. One simple choice is to use $\mathbb{R}^n$ equipped with the euclidean metric, where all the trajectories will be simple straight lines. In figure 1, one can see another choice of manifold and metric, leading to trajectories being logistic curves.

In order to parametrize the distribution of those trajectories together, the model has a mixed-effects hierarchical structure that separates the population fixed parameters and the individual random parameters:

- **Population parameters**: the individual trajectories form a bundle centered around an "average" trajectory. This central trajectory is taken as a geodesic γ_0 in the manifold $\mathcal{M}$. This geodesic is

parametrized by its initial conditions at time t_0: the initial point on the manifold $\gamma_0(t_0) = \mathbf{p}$ and the initial speed $\dot{\gamma}_0(t_0) = \mathbf{v}$.

- **Individual parameters**: here we separate temporal and spatial parameters
  - **Spatial parameters**: each individual trajectory γ_i stems from the central geodesic γ_0 with a small shift on the manifold $\mathbf{w}_i$ called a space-shift, or inter-marker spacing. The individual follows a trajectory parallel to γ_0 in the sense of the Exp-parallelisation, a concept extending the notion of parallels to manifolds [30]. The space-shift $\mathbf{w}_i$ is a vector in the tangent space at $\mathbf{p}$ defined so that shooting with the Riemannian exponential from $\mathbf{p}$ at speed w_i yields a point of γ_i.
  - **Temporal parameters**: The individual temporality of the disease can be modelled with a time reparametrization $\psi_i(t)$ so that the observation $\mathbf{y}_{ij}$ is $\gamma_i(\psi_i(t_{ij}))$. The chosen time-warp is $\psi_i(t) = \alpha_i(t - t_0 - \tau_i) + t_0$, where τ_i is called the time-shift and models an early or late onset, and $\alpha_i = e^{\xi_i}$ is called the acceleration factor.

Figure 1 provides a visual understanding of the geometric model.

The space-shifts $(\mathbf{w}_i)_i$ have the same dimension as the observations and are a vector in the tangent space at $\mathbf{p}$. In order to have identifiability with the time delay τ_i which operates as a shift along the curve, the space-shifts are required to be orthogonal to $\mathbf{v}$, which we enforce by restraining $\mathbf{w}_i$ to $Span(\mathbf{v})^{\perp}$. For better interpretability the model reduces the dimension of the parameters to estimate with an independent component analysis (ICA decomposition) [12] with N_s independent sources $(\mathbf{s}_i)_{1 \leq i \leq N_s}$. The formulation unfolds as $\mathbf{w}_i = A\mathbf{s}_i$ where the columns of A are orthogonal to v.

2.1.2 The logistic curves model

The disease course mapping model is a generic framework, and we can apply it to different types of observations by choosing the adequate manifold. However because we will model binary observations using a logit model, we will focus on the manifold whose geodesics are logistic curves. Then, the Riemanninan manifold used in that case is set to be the product manifold of d times $(0, 1)$ equipped with the metric $g_p(u, v) = uG(p)v$ where $G(p) = \frac{1}{p^2(1-p)^2}$. The choice of logistic curves also echoes a well-known hypothesis about the progression of biomarkers in Alzheimer's disease [14], but the use of logistic curves is common in most neurodegenerative diseases. This is very convenient as we will see, since the logistic is also fundamental in IRT.

The specific formulation of the curve of one dimension k for one individual i at time t_{ij} in the model is the following:

$$\gamma_{ik}(t_{ij}) = \left(1 + (\frac{1}{p_k} - 1) \exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1 - p_k)}\right)\right)^{-1} \tag{1}$$

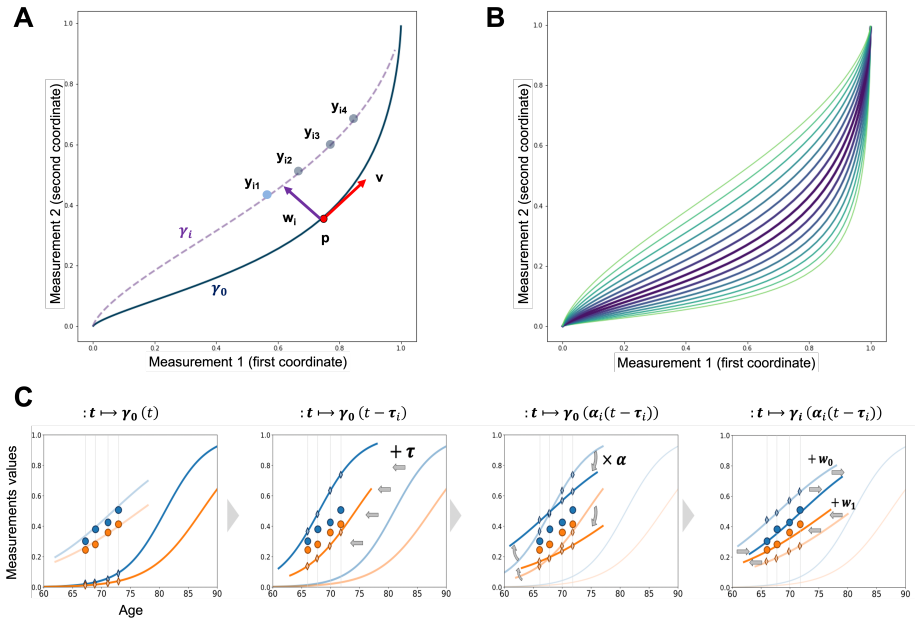

Figure 1: **A.** Generic model scheme. The average population trajectory γ_0 is a geodesic in the manifold. An individual trajectory γ_i is defined as an exp-parallellisation of γ_0 by a space-shift $\mathbf{w}_i$. $\mathbf{p}$ and $\mathbf{v}$ are shown on γ_0, with a space-shift $\mathbf{w}_i$ pointing to the resulting trajectory γ_i upon which lie the measurements $(\mathbf{y}_{ij})_j$. **B.** Logistic curves model. A bundle of trajectories corresponding to the logistic curves model in the 2-dimensional manifold is shown. **C.** Personalization process in the logistic curves model. The plots show the successive influence of the individual parameters on the average trajectory (left) until the trajectory is personalized and matches the measurements (right). Note that the x-axis introduces the time and the two dimensions of the measurements are shown as two different colors.

where (p_k, v_k, t_0) are the fixed-effects parameters for the average trajectory on item k defined by $\gamma_k(t_0) = p_k$ and $\dot{\gamma}_k(t_0) = v_k$. (τ_i, ξ_i, w_{ik}) constitute the random effects, ξ_i being the log-acceleration and τ_i the time-shift both modulating the individual disease timeline while w_{ik} is the space-shift parameter described in the previous section.

In figure 1 we can observe the typical trajectories on the manifold in 2 dimensions, as well as the trajectories as a function of time.

2.2 Binary model

Based on the logistic curves model we will now apply it to non-continuous data, starting from simple categorical data. The basic framework is when observations y_{ijk} are binary: either 0 or 1. The model adaptation is rather

simple: the observations y_{ijk} are not treated as a noisy measure of the true trajectory $\gamma_{ik}(\psi_i(t_{ij}))$ but as a realization of a random Bernoulli variable with probability $\gamma_{ik}(\psi_i(t_{ij}))$.

$$y_{ijk} \sim \mathcal{B}\left(\gamma_{ik}(\psi_i(t_{i,j}))\right) \tag{2}$$

As in the framework of IRT, if we consider the value to be the result of one item in a test as right or wrong (0 or 1), we can evaluate the observed result as a probabilistic outcome relying on a latent disease state passing through a logit model.

2.3 Ordinal model

Now let's assume the measurements lie on a discontinuous space with ordered values such as for a cognitive score. Without loss of generality we assume these values are integers between 0 and L included. As for the binary case, we will again consider that the model values belonging to the manifold (which is continuous) describe a form of probability (here through the cumulative distribution function). We now specify the logistic curves model for ordinal data similar to Samejima's model (also known as cumulative logit model) from IRT [28]:

$$\begin{aligned}
\forall l \in [|1, L|], \mathbb{P}(y_{ijk} \geq l) &= \gamma_{ik}(\psi_{ik}^l(t_{i,j})) \\
&= \left(1 + (\frac{1}{p_k} - 1)\exp\left(-\frac{v_k\psi_{ik}^l(t_{ij}) + w_{ik}}{p_k(1-p_k)}\right)\right)^{-1} \\
\psi_{ik}^l(t_{ij}) &= e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0 - \sum_{m=1}^{l} \delta_k^m \\
\forall l \in [|1, L|], \delta_k^l &> 0
\end{aligned} \tag{3}$$

where ψ_{ik}^l is the time reparametrization of individual i for the item k at level l. As in equation 1 the population parameters (p_k, v_k, t_0) define the base logistic curve (i.e. the curve for level 1 of the item) as the initial position, initial velocity and initial time respectively. τ_i and ξ_i are the time-related parameters, namely the time-shift and the log-acceleration. w_{ik} is the space-shift. δ_k^l are new parameters introduced to deal with the different levels of the item, as explained thereafter.

The idea is to model the cumulative distribution function with the logistic curves instead of the probabilities of each level directly. Each level is modelled with a logistic curve. However we do not want to overparametrize the model so we choose to enforce parallelism by setting the same velocity $\mathbf{v}$ for each level of a same item. This introduces only one parameter for each added level. The new parameter is δ_k^l, the delay in time between levels $l-1$ and l. This delay is a fixed effect, so it is shared by the whole population. Since it is added to the time reparametrization function ψ_{ik} after the affine transformation of time by τ and e^{ξ}, this time delay is a duration in the common disease timeline. This means that if one patient has an acceleration factor of 2, the expected time to jump from one level to

the next is divided by 2. With the δ_k^l being positive, we also ensure that the condition $\mathbb{P}(y_{ijk} \geq l) < \mathbb{P}(y_{ijk} \geq l-1)$ is verified. From an interpretability standpoint, these delays are helpful in understanding which levels of an ordinal scale last longer than the others, or on the contrary which levels represent only a short transition.

2.4 Bayesian estimation

The geometric model described previously can be plugged in a hierarchical probabilistic model in order to estimate the parameters by maximizing the likelihood. In this framework we assume that y_{ijk} are noisy observations, formulated as: $\mathbf{y}_{ij} = \gamma_i(\psi_i(t_{ij})) + \epsilon_{ij}$ with the noise $\epsilon_{ij} \sim \mathcal{N}(\mathbf{0}_d, \Sigma)$ with Σ being a d-dimensional diagonal matrix.

Then we assume that all of the previously described parameters are latent, and follow normal prior distributions except for the following ones: in the case of the logistic curves model, $\mathbf{p}$ is estimated as a logit transform of a gaussian: $p_k = \frac{1}{1+e^{g_k}}$ where g_k has a normal prior; for the ordinal model, the δ parameters must be positive, so we set their prior distribution as a log-normal.

We introduce the following notations: z_{pop} for the population parameters, z_i for the individual parameters, $\mathbf{z} = (z_{pop}, (z_i)_{1 \leq i \leq n})$ for the latent variables. Finally the statistical model parameters are noted θ.

For the estimation of all the parameters of the disease course mapping model, the Monte Carlo Markov chain stochastic approximation variant of the Expectation Maximization algorithm is used. The convergence has been proven in [18] for the curved exponential family. The algorithm alternates between the two steps of the classical EM, first by computing the expectation of the log-likelihood through stochastic sampling of latent parameters $\mathbf{z}$ and then by maximizing the model parameters θ in closed form.

As mentioned before, the probabilistic model allows us to deal with missing observations without needing to impute them as the likelihood is only computed for the points we observe. The robustness to the missing data as been evaluated in [4].

In terms of optimization of the log-likelihood, changing the model from a gaussian noise to a Bernoulli variable amounts to optimizing a crossentropy between the latent model and the trajectory instead of a mean squared error:

$$\hat{\mathbf{y}}_{i,j} = \gamma_i(\psi_i(t_{i,j})) \qquad \text{[simplified nota}$$

$$\log(p(\mathbf{y}|\mathbf{z};\theta)) = -N_{tot}\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{N_i}||\mathbf{y}_{i,j} - \hat{\mathbf{y}}_{i,j}||^2 \qquad \text{[continuous model with gaussian n}$$

$$\log(p(\mathbf{y}|\mathbf{z};\theta)) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{d} y_{i,j,k}\log(\hat{y}_{i,j,k}) + (1-y_{i,j,k})\log(1-\hat{y}_{i,j,k}) \qquad \text{[categorical model with Bernoulli realiza}$$

For the ordinal model, maximizing the log-likelihood is straightforward, as we can simply come back to $\mathbb{P}(y_{ijk} = l)$ by a difference $\mathbb{P}(y_{ijk} \geq$

$l) - \mathbb{P}(y_{ijk} \geq l+1)$. One thing to note here is that the log-likelihood cannot be expressed in the curved exponential family and we are outside the scope of the theoretical guarantees for the MCMC-SAEM convergence. However, convergence has always been reached in practice.

2.5 Initialization method

Since the MCMC-SAEM estimation algorithm will only converge towards a local optimum of the likelihood, we need to ensure that either we replicate the estimation with several initial points or we provide a starting point which is already a "good guess". For the population parameters of the logistic curves model, we compute the initial point for $(\mathbf{p}, \mathbf{v})$ by computing a regression on the individual progression above the first level, with $\mathbf{p}$ being deduced from the intercept and $\mathbf{v}$ from the slope. For the $(\delta_k^l)_{k,l}$, we computed them as the mean time to reach level l from level $l-1$. Indeed each δ_k^l can be understood as the time between the progression of the probabilities $\mathbb{P}(y_{ijk} \geq l)$ and $\mathbb{P}(y_{ijk} \geq l-1)$.

3 Results

All the methods are developed in Python by extending the open-source Leaspy library (https://leaspy.readthedocs.io) created for disease course mapping models and run on a 2.80GHz CPU with 16 GB RAM. All the code for the ordinal disease course mapping model is already available on Gitlab: https://gitlab.com/icm-institute/aramislab/leaspy

3.1 Simulation study

In this first experience we generate synthetic ordinal data in order to highlight two benefits of our modelling approach. On the one hand, we show how modelling items leads to a finer-grained description of the disease progression, as compared to modelling aggregated scores. In particular, we show that a model based on item response exhibits patients clusters that were not visible with a continuous model. On the other hand, we compare how the ordinal model fares as opposed to the logistic curves model. We show that the flexibility introduced with the δ parameters allows to learn a step function which is less rigid than an imposed template such as a logistic curve.

Generation process: We use our ordinal model in order to simulate synthetic patients and obtain their observations at randomly chosen time-points. For the initialization, we chose the statistical model parameters θ and the population parameters $\mathbf{z}_{pop}$. The values are chosen arbitrarily in the range of plausible values for a neurodegenerative-like disease progression model. In this simulation part, we kept a low dimensionality for the data with only $d = 3$ dimensions, which is the lowest number of dimensions that allows a meaningful ICA decomposition with at least 2 sources. We then adjusted the δ_k^l so that each of the three items has an atypical progression, by which we mean it can not be easily interpolated by a linear curve or a logistic curve. Each item has $L = 4$ levels, which sums

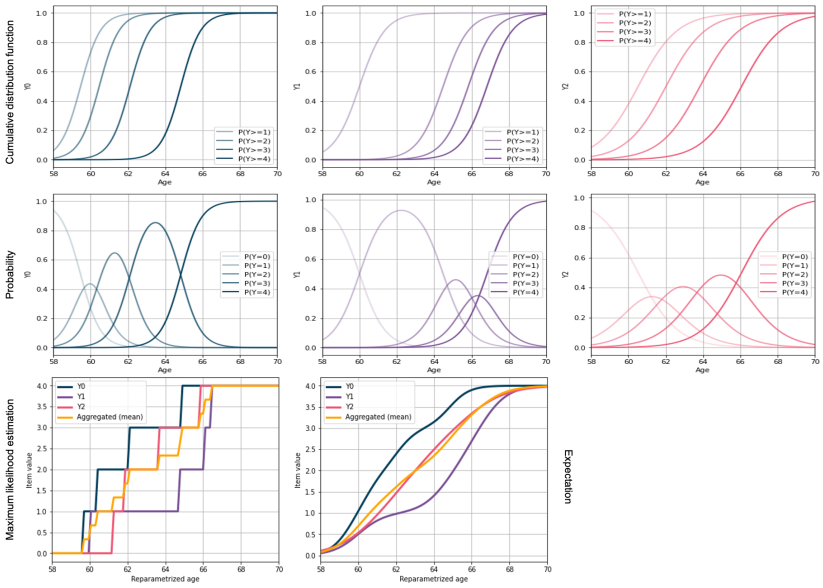

Figure 2: Average trajectory γ_0 for the ordinal model. Top row: for each item the cumulative distribution functions are shown. Note that δ_k^l correspond to the shift on the x-axis between the logistic curves. Middle row: raw probabilities are shown, directly obtained by difference between the logistic curves of top row. Last row: items are shown together by summarizing the multinomial distributions of items. On the left the value at each timepoint is the maximum likelihood; on the right the value is the expectation. We also plot the aggregated score computed as the mean of the items.

into an aggregated score ranging from 0 to 12. Figure 2 shows the average trajectory γ_0 of the model, as well as the average aggregated trajectory. Once the population parameters are fixed, we randomly generate new individuals: for each individual parameter, we sample the distribution mentioned in section 2.1.2. However to generate two different clusters, we choose to randomly attribute a class 1 or 2 to a new individual. Then for the individual parameters simulation, the only difference between individuals of the two clusters is that the first source in cluster 1 follows a gaussian distribution centered around -4 instead of 0 with a standard deviation of 1 and the source 1 in cluster 2 is oppositely centered around $+4$. This source acts as an inverter of the order of the curves. The mean trajectory in cluster 1 has the curve Y0 starting first, then Y1 follows and finally Y2, whereas for cluster 2 the order is switched: Y1 first, then Y2 and finally Y0. However the aggregated score is roughly the same in the two clusters. With an equal repartition of individuals between the two clusters, we therefore obtain a balanced dataset of individuals. Once this is done, we need to generate random timepoints for the visits of each individual. For this we decided to select the number of visits with a binomial distribution

with probability $p = 0.5$ and $n = 10$ to which we add 2 to guarantee at least two visits per individual. Then each of the timepoint is sampled with a gaussian distribution around the middle of the disease progression curve, with a time window calibrated to span about 5 years of follow-up for each subject.

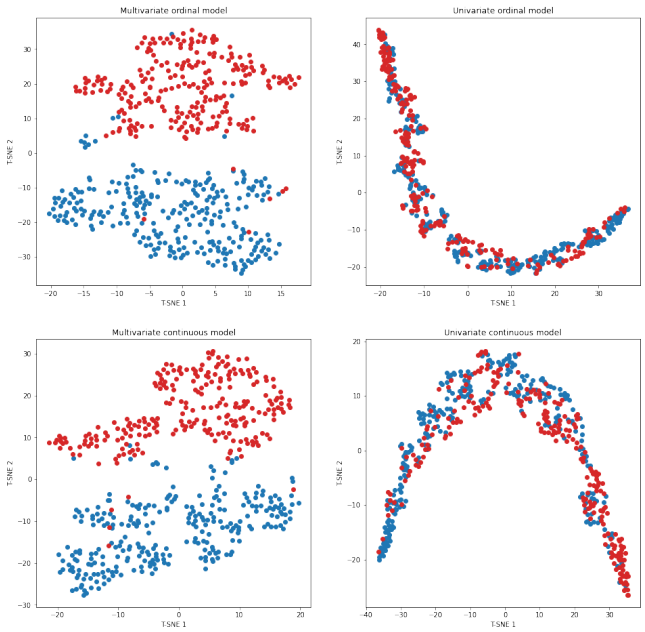

Figure 3: t-SNE plots for individual parameters. On the left side the parameters in the multivariate model are shown, on the right side the parameters in the univariate models. Top row shows the two ordinal models while bottom row shows the continuous models. Each color corresponds to one of the clusters (1 in blue, 2 in red).

We estimated four models on this synthetic dataset:

- a continuous model (equation 1) on items: to highlight the purpose of using a dedicated ordinal model for such scales, we fit a disease mapping course model with logistic curves as presented in the method section with observations considered as continuous with a gaussian noise. The model is three dimensional with two sources
- an ordinal model (equation 3) on items: we fit the model presented in the method section, here the observations are treated as ordinal. The model is also three dimensional with two sources
- a univariate continuous model on the aggregated score: to show how the loss of item information impacts the ability of a model to produce relevant information about patients, we fit a single logistic curve model on the sum of the three items. The model is thus one-dimensional without sources

- a univariate ordinal model on the aggregated score

The first point of our synthetic experiment was to show how the gain of finer-grain information allows to learn meaningful data patterns. We extracted the individual parameters of the four models. Each individual i thus has a set of parameters $(\tau_i, \xi_i, s_i^1, s_i^2)$ in the multivariate models and (τ_i', ξ_i') in the univariate models. For visualizing the points in figure 3, we decided to embed the individual in a two-dimensional space using a t-distributed stochastic neighbor embedding [21]. We see that the two clusters are clearly separated in the multivariate models, whereas the aggregated model is not able to disentangle them. A quantitative classification with a logistic regression taking in input the set of individual parameters and predicting the cluster of origin confirms this conclusion: with $(\tau_i, \xi_i, s_i^1, s_i^2)$ from either multivariate model the classifier gets close to 100% accuracy while with (τ_i', ξ_i') from the best univariate it only has 65%, which is much worse.

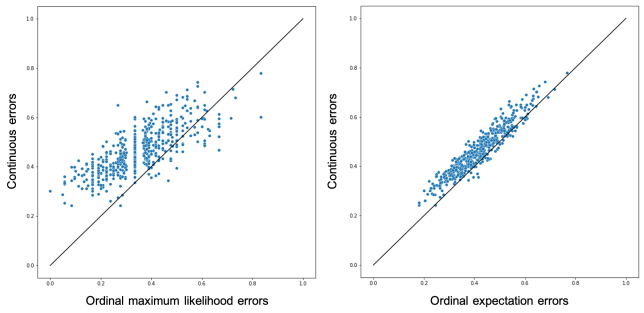

Figure 4: Error of reconstruction scatter plot. For each individual, we computed the mean absolute error of reconstruction over all visits and all items between the observations and the model prediction for the multivariate models. In the case of the continuous model the prediction is directly the outcome of the model, whereas for the ordinal model we computed the maximum likelihood for each value in the left plot, and the expectation in the right plot. The plots show that the ordinal model outperforms the continuous model for almost all patients in the reconstruction of the items.

While the multivariate continuous model also captures individual heterogeneity by identifying the two clusters, it is not as performant in terms of pure data reconstruction. Figure 4 illustrates the difference between the continuous and the ordinal model in terms of reconstruction errors on the items, with two different methods for predicting the outcome of the ordinal model. It is clear that the continuous model is worse, with a first burden being the lower bound of the noise due to the discretized scale (every 0.25 in this case). This can be mitigated in part by comparing the rounded predictions of the continuous model to the maximum likelihood predictions of the ordinal model, but even in this case we average 0.38 points of absolute error for continuous rounded predictions versus 0.34 for the ordinal maximum likelihood predictions. The second burden is the prior choice

of the form of the curve (here the logistic). It is important to note that for small scale items in general there is hardly a perfect continuous curve due to the variance of item levels. For instance, a standard Likert scale for a symptom would be a $0-4$ scale with each level corresponding to a severity measure: null, mild, moderate, severe, very severe. The difference between two successive levels may vary, which highlights the role of the δ_k^l parameters.

These preliminary results encourage the use of dedicated ordinal models on ordinal scales, and this proves true even for items with more levels (especially when levels are unevenly distributed). In the appendix we also provide experimental results with data generated with a continuous model rather than the ordinal model. These results show that the ordinal model is also able to outperform a continuous model even in this setting.

3.2 Application to Parkinson's disease data

3.2.1 Data set

Data availability statement: Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database [22] (www.ppmi-info.org/access-data- specimens/download-data). For up-to-date information on the study, visit ppmi-info.org. [1]

The cohort includes mainly Parkinson's disease (PD) patients within two years of their diagnosis. There are also prodromal individuals, but we excluded those as they are a minority and some of them do not convert to PD. We selected all PD diagnosed patients with at least 2 visits in the study in order to have a longitudinal history for our training set, which left 900 subjects for a total of 7918 visits. From the wide range of biomarkers and clinical assessments we chose to focus on the MDS-UPDRS score. It is composed of 65 items divided in 4 sections: non-motor aspects of experiences of daily living (I) which is filled half by the clinician and half by the patient in a self-report questionnaire; motor aspects of experiences of daily living (II) which is a form only filled by the patient; motor examination (III) by the clinician; motor complications (IV) which is assessed by the clinician.

Each item is rated on a Likert scale (0: Normal, 1: Slight, 2: Mild, 3: Moderate, 4: Severe). For the third part, two versions of the test are recorded depending on the treatment effect. A version "OFF" is assessed when the patient hasn't taken medication in the last 6 hours, which is a state when the treatment should have little effect and motor symptoms will usually be observed. A version "ON" corresponds to the same motor examination shortly after the patient has taken medication, which is a state where motor complications are alleviated. Therefore we used "OFF" measures when selecting features in order to build the disease natural history with as few mitigation of treatment effect as possible. We also removed the fourth part of the MDS-UPDRS since it focuses on treatment

[1] PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors].

secondary effects. In the end we obtained 59 items rated on scales from 0 to 4.

Most of the items have a low mean: 44 items have a mean between 0.4 and 1.25. Most of the observed values are 0 (66%), with 22% of 1 values, and only 12% of values above 2. For some items, the higher levels are very seldom seen so their progression cannot be modelled fully due to lack of data. The average total score on the parts I, II and III is 28.7 ± 22.9 with 90% of the values within the $[1, 60]$ range.

3.2.2 Qualitative experiment: a description of the MDS-UPDRS

As the previous summary shows, patients in PPMI do not display all of the possible symptoms, and most of their items remain at 0 during the course of their follow-up. However items with non-zero values often come as groups, which suggest that several items capture a similar biological phenomenon. We aim at disentangling these trends with the ICA on the space-shifts.

We calibrated an ordinal DCM model on the 900 subjects using their whole set of 59 items. Missing values do not need to be imputed since the model allows to only compute the likelihood for observed values, which is one of the perks of using the DCM. The estimation was performed with the MCMC-SAEM algorithm using a block Gibbs sampler (sampling was grouped for of all the 59 dimensions of the population parameters $\mathbf{v}, \mathbf{g}, \delta$) during the first 16000 iterations before switching to a more fine-tuned Gibbs sampler (each scalar coordinate was sampled separately) for the last 4000 iterations. The samplers used an oscillating tempered scheme. This temperature scheme allows alternating between the much needed exploration in high dimensions with high temperature and the local search of the optimum during low temperature phases. The overall time for calibration was 56h.

Due to the very high number of dimensions we chose 8 as the number of sources to reduce the dimension of individual parameters and prevent overfitting. Previous trials with less sources yielded significantly lower log-likelihood at the end of the calibration while increasing the number of sources becomes too computationally expensive. The ICA enforces correlations within the space-shifts due to the relation $\mathbf{w}_i = \mathbf{A}\mathbf{s}_i$, with $\mathbf{A}$ being rectangular. We can analyse the correlations across individuals resulting from these space-shifts values, which can be seen as a time delay in the progression of each item compared to the population average trajectory. Figure 5 shows the Pearson correlation matrix between the space-shifts corresponding to each item. We performed hierarchical clustering in order to isolate groups of highly correlated items.

As we can see in Figure 5, the first group of items is comprised of tremor related items. The next two clusters gather items of the motor part (III) of the MDS-UPDRS, with one cluster containing items for the left part of the body and the other being dedicated to the right part. This is explained by the fact that Parkinson's disease very often declares itself in one side of the body before the other. Then we have a small trio of gait, freezing of gait and posture, which can be included in a larger cluster of

items concerning motor aspects of daily life (essentially items of the part II related to movement and balance). Then the last large cluster includes all the elements of part I (non-motor aspects of daily life) and the last items of part II. This large cluster can be separated into two subclusters, one being the part I of the MDS-UPDRS and the second being the leftover items of part II which are more related to mouth tasks (eating, drooling...) and communication. Overall we can notice a strong correlation between items of part I and part II. These parts are both questionnaires and mostly filled by the patient, which can explain the correlation bias due to the self-assessment of one's abilities.

Figure 5: Pearson correlation matrix of space-shifts at the item level. Hierarchical clustering has been performed to order the items of the different parts of the MDS-UPDRS.

Parkinson's patients are commonly separated into two or three subtypes [31]: tremor dominant (TD), postural instability and gait dominant (PIGD) and mixed patients (at the limit between TD and PIGD). Here we see that the TD and PIGD tendencies are represented in the correlations of items of the first and fourth clusters.

Some items were not regrouped with others or were associated in different ways. For instance we can note that neck rigidity and facial expression were at the junction between the left and right motor clusters, suggesting that higher body motor impairment usually follows motor symptoms in one side of the body. However these two items are less correlated to motor impairment in the legs and the feet. This is very interesting as we can hypothesise that the disease implies a localised neurodegeneration of motor neurons, and then symptoms spread from their original onset, explaining the anticorrelation between top and bottom motor impairment. Speech (part III question 1) has been isolated, while spontaneity of movement (part III question 14) seems to correlate with all the motor assessments of part III, as a global appreciation of the motor state of the patient by the clinician.

We gave a specific look at the item called DDS (dopamine dysregulation syndrome) which is a disturbance of dopamine therapy with addictive patterns such as impulse control disorders (ICD) which include gambling, hypersexuality or compulsive eating [25]. This syndrome is a consequence of too much dopamine uptake (sometimes voluntarily on the patient's side). What appears is that the DDS is anticorrelated to all motor items of part III *except* for the freezing of gait (part III question 11). This backs the conclusion that freezing of gait is also a secondary effect of dopaminergic treatment rather than a disease symptom [8].

We then examinated the sources individually to identify which statistically independent linear combinations of the items were found, and tried to understand if they were linked to a pathological trend. We provide a description of the 8 sources:

- source 1 (variance explained 11%): this source induces a positive delay in all motor items of the left part of the body while it pushes for a negative delay in almost all of the other items. We can interpret this source as a left body delay
- source 2 (variance explained 10%): this source is almost the same as source 1 but for right body items. We call this one right body delay. These two first sources with equivalent prevalence show that Parkinson's disease is as likely to start in the left part of the body as the right
- source 3 (variance explained 12%): this source has a positive delay on all motor items linked to the limbs' movement as opposed to the other items. We name this one motor source
- source 4 (variance explained 15%): this source has a positive delay on the left part of the body and the rigidity items. This source embeds a left body and rigidity combined progression
- source 5 (variance explained 12%): this source has a positive delay on all the questions that are self-evaluated as opposed to all the motor evaluations and the non-motor assessment by the clinician. We can therefore say that this source is the patient's self report effect
- source 6 (variance explained 13%): this source regroup items of the part I (non-motor aspects of daily life) with tremor manifestations as opposed to motor issues. This source is seen as the brain-first source

- source 7 (variance explained 14%): this source is driven mostly by the motor items and more especially the gait, posture and balance items. We call this one the gait and balance source
- source 8 (variance explained 13%): this source gathers items from the part I with a strong emphasis on anxiety, depression and apathy, along with speech issues and opposed to tremor items. We coin this one the cognitive source

We observe that some sources are in practice not totally independent, for instance source 5 (self report effect) and source 4 (left body rigidity) are correlated. This suggest that the number of sources is not fully adequate. We see that sometimes a source seem to include several groups of items at the same time (source 4 is a mix of left body items and rigidity), which could be the consequence of too few dimensions to disentangle these effects.

If we come back to the classification of patients between tremor dominant and postural instability/gait dominant, we see that PIGD is solely driven by source 7, whereas tremor information is spread across sources 1 to 4.

We won't dwell on more clinical subtleties but hope this showcased the power of the ICA on the spatial effects unfolding from the Riemannian framework of the DCM. The results here could only be obtained thanks to the modelling of items with an ordinal method. The global model calibrated on the 59 items showed that some items had a very low progression profile (or could not be completely learned due to the few data available on the last levels), especially in the motor examination part. However the second part of the MDS-UPDRS scale was more reliable since it rarely stayed at zero and the progression was faster and steadier. We thus used the 13 items of the second part in the next experiment to show quantitatively how well the ordinal model performs.

3.2.3 Quantitative results: prediction task

In order to understand the value brought by our new take on the model, we compare the ordinal DCM to the logistic version of the DCM taking the cognitive scores as continuous values. We also compare to a linear mixed-effect model and a no-change prediction model as our baseline. We will have two different options in order to show what the model can bring to the table: one model for the total score of one MDS part (the second part as mentioned previously) which is thus an integer between 0 and 42; and one model taking the items into account, as 42 is the sum of 13 questions with a scale from 0 to 4. In this last case we will show that taking the score as a continuous value is not valid since the noise here is lower than the step between possible values. We present the several models used in this experiment:

Continuous univariate model

The MDS UPDRS part II (/42) is normalized between 0 and 1 in order for the logistic model to work. We then maximize the log-likelihood with gaussian noise, thus implying a fit in the least squares sense. This corresponds to formula 1.

Continuous multivariate model

Each item (/4) of the MDS UPDRS part II is normalized between 0 and 1. As for the univariate maximize the log-likelihood with gaussian noise, where the standard deviation of the noise is feature-dependent. For the prediction of the MDS UPDRS II total we sum the predictions of the items. This also corresponds to formula 1.

Ordinal models

In order to operate a fair comparison we use the same hyperparameters as for the continuous version, so that the ordinal model shares the same parameters as the continuous one except for the Δ. We fit by maximizing the log-likelihood, and we compute the predictions using the expectation of the ordinal model: $\mathbb{E}(\eta_{ij}) = \sum_l \mathbb{P}(\hat{y}_{ij} \geq l)$. As for the continuous case, we have a univariate model built with only the MDS UPDRS II total score and a multivariate model using the 13 items. For the multivariate case, we compute the prediction as the expectation of the sum of the items. The ordinal model corresponds to formula 3.

Baseline models

We also included a baseline constant model that predicts the last known value for the patient (hypothesising no change in the future) and a linear mixed-effect model for comparison (formula provided in appendix). Note that they perform very well: it is hard to beat these models when predicting in a very short future since the state of the patient evolves very slowly.

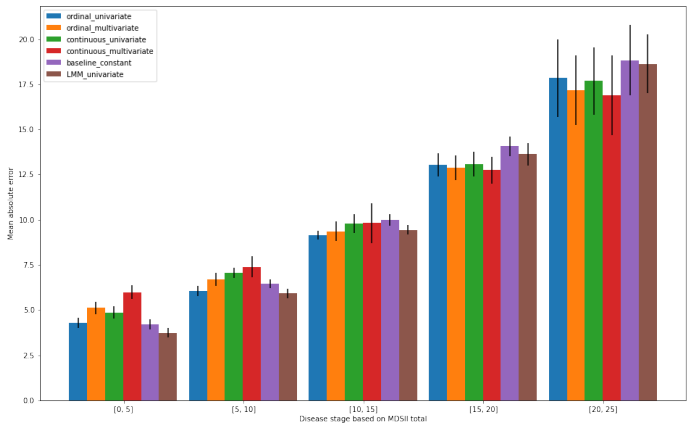

Figure 6: Mean absolute prediction error as a function of the current patient state. Bin ranges correspond to the MDSII value of the patient. Error bars for mean model error are estimated from cross-validation. LMM: linear mixed-effect model; continuous: disease course mapping model; ordinal: disease course mapping model, ordinal version; univariate: model learned of MDSII total directly; multivariate: model learned on MDSII items and then aggregated.

The prediction framework we use consists in two steps. First a training set is used to calibrate a model, i.e. learning the fixed effects for the mixed-effect models (linear and DCM). In a second step, we use the patients of

the test set. Out of each patient's set of visits we randomly select some past visits and a visit in the future for which we want to predict. We learn the random effects of the models on the past visits and use the personalized model to compute the value at the desired future visit's timepoint. We use a repeated 8-fold cross-validation to measure the models errors. The figure 6 shows the mean absolute prediction error.

As mentioned previously, we observe that constant prediction and linear mixed model perform best when the patient score is low, as they progress slowly at the beginning of the disease. However we see that the prediction from those simple models worsens as the score increases, meaning that linear and constant models are less able to capture disease progression. As a matter of fact, they largely under estimate the change for patients later in the disease.

On the ordinal side we observe a similar trend, although both ordinal models are more reliable at higher disease stages. They outperform continuous models on the low stages, due to the flexibility of the step parameters $(\delta_l)_l$. However when it comes to later disease stages continuous models catch up with ordinal models in performance. This is mainly due to the data being much sparser at these stages, leaving a larger role to the shape priors. In the long term setting it confirms that the logistic prior improves on linear and constant curve priors. Ordinal models show that their low granularity allows for better prediction at low levels when changes are small, while we lack data for higher disease stages thus leading to a mitigated performance.

Comparing univariate versus multivariate models is very interesting. At early disease stages the univariate models seem to better capture the essence of the progression. On the other hand, the higher the disease stage the better the multivariate models perform, showing that multidimensional models are better suited for disease progression prediction when the rate of change increases.

Figure 7 shows the average disease progression learned by univariate mixed-effect models. Spaghetti plots with wrapped individual trajectories are provided in the appendix. The figure 7 shows that the logistic and the ordinal provide a somehow similar description of the aggregate score up until age 85. After the logistic model is bound to the assumption that the score will evolve to 1, which in practice is highly unlikely as it is a complete state of degeneracy and patients usually do not reach this stage. This is why the ordinal curve slows compared to the logistic at the end. The logistic model assumes that the disease progresses faster after 85 years old. On the other hand the linear mixed-effect model tends to underestimate the rate of change at this stage. Note that the shape of the step function is mostly driven by the δ values learned, and thus is more flexible. This explains why it is more accurate for the values of the MDSII total between 0 and 15 where most of the observations lie according to the histograms.

The experiment was performed again on parts I and III of the MDS-UPDRS with continuous and ordinal models, and the results were essentially the same:

- For the MDS-UPDRS part I (/42), the ordinal univariate had a mean absolute error of 4.64 while the multivariate ordinal had 4.54;

continuous univariate was 4.76 and continuous multivariate was 5.41

- For the MDS-UPDRS part III (/132), the ordinal univariate had a mean absolute error of 12.8 while the multivariate ordinal had 12.3; continuous univariate was 15.8 and continuous multivariate was 13.9

What we seem to notice though is that the more items in a score, the better the multivariate approach performs, although it comes at a higher computational price.

4 Discussion

Our work shows the potential of leveraging items with a small ordinal scale over working with an aggregated score. We built upon a non-linear mixed-effect model used mainly for continuous markers and extended it to ordinal data. This allowed to address cognitive decline by processing items of the test instead of their aggregated value. We showed how this could help understand disease trends in Parkinson's disease within the MDS-UPDRS. We believe this could be extended to other cognitive scales in other neurodegenerative scales with the aim to better understand items dynamics.

In the prediction task, which is especially hard, ordinal models had a finer prediction for early disease stages, when disease evolution is slower. Our results highlight the need to rely on multivariate data to increase prediction performance at later disease stages, when rating scores changes are of higher magnitude.

Future applications could use this model as a tool to extract information from specific items after a multivariate analysis. We believe this can be the first step to building new composite scores such as the PACC5 in Alzheimer's disease [26]. One idea would be to use the Fisher information as the metric of choice to select meaningful items in a composite score, since in IRT it is common to compute Fisher information of items for the individual parameters.

On the downside we see the limitations of the ordinal method when used on data with few values for certain levels (in our case these were the high values of the MDS UPDRS II), as the added parameters of the ordinal DCM require more data in order to be properly learned. These extra parameters also come at a computational cost within the estimation algorithm, therefore we deem the method useful when the items are rated on short scales like the Likert one.

Acknowledgments

Financial disclosure

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-10-IAIHU-06 and ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and by ANR under the joint programme in neurodegenerative diseases (JPND) ANR-19-JPW2-000 (E-DADS). This

work was also funded in part by grant number 826421 (TVB-Cloud) from H2020 programme.

Conflict of interest

The authors declare no potential conflict of interests.

Supporting information

References

[1] Leticia Arrington, Sebastian Ueckert, Malidi Ahamadi, Sreeraj Macha, and Mats O. Karlsson. Performance of longitudinal item response theory models in shortened or partial assessments. *Journal of Pharmacokinetics and Pharmacodynamics*, 47(5):461–471, October 2020.

[2] Murat Bilgel, Bruno M. Jedynak, and Alzheimer's Disease Neuroimaging Initiative. Predicting time to dementia using a quantitative template of disease progression. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 11(1):205–215, 2019.

[3] Murat Bilgel, Jerry L. Prince, Dean F. Wong, Susan M. Resnick, and Bruno M. Jedynak. A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *NeuroImage*, 134:658–670, July 2016.

[4] Raphael Couronne, Marie Vidailhet, Jean Christophe Corvol, Stephane Lehericy, and Stanley Durrleman. Learning Disease Progression Models With Longitudinal Data and Missing Values. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1033–1037, April 2019. ISSN: 1945-8452.

[5] Andrea Cremaschi, Maria De Iorio, Yap Seng Chong, Birit Broekman, Michael J. Meaney, and Michelle Z. L. Kee. A Bayesian nonparametric approach to dynamic item-response modeling: An application to the GUSTO cohort study. *Statistics in Medicine*, 40(27):6021–6037, 2021.

[6] Stanley Durrleman, Xavier Pennec, Alain Trouvé, José Braga, Guido Gerig, and Nicholas Ayache. Toward a Comprehensive Framework for the Spatiotemporal Statistical Analysis of Longitudinal Shape Data. *International Journal of Computer Vision*, 103(1):22–59, May 2013.

[7] Stanley Durrleman, Xavier Pennec, Alain Trouvé, Guido Gerig, and Nicholas Ayache. Spatiotemporal Atlas Estimation for Developmental Delay Detection in Longitudinal Datasets. In Guang-Zhong Yang, David Hawkes, Daniel Rueckert, Alison Noble, and Chris Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, Lecture Notes in Computer Science, pages 297–304, Berlin, Heidelberg, 2009. Springer.

[8] A.J. Espay, A. Fasano, B.F.L. van Nuenen, M.M. Payne, A.H. Snijders, and B.R. Bloem. "On" state freezing of gait in Parkinson disease. *Neurology*, 78(7):454–457, February 2012.

[9] Christopher G. Goetz, Barbara C. Tilley, Stephanie R. Shaftman, Glenn T. Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B. Stern, Richard Dodel, Bruno Dubois, Robert Holloway, Joseph Jankovic, Jaime Kulisevsky, Anthony E. Lang, Andrew Lees, Sue Leurgans, Peter A. LeWitt, David Nyenhuis, C. Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A. Teresi, Jacobus J. van Hilten, and Nancy LaPelle. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results: MDS-UPDRS: Clinimetric Assessment. *Movement Disorders*, 23(15):2129–2170, November 2008.

[10] Rosalie Gorter, Jean-Paul Fox, and Jos W. R. Twisk. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, 15(1):55, July 2015.

[11] Gopichand Gottipati, Alienor Berges, Shuying Yang, Chao Chen, Mats Karlsson, and Elodie Plan. Item Response Model Adaptation for Analyzing Data from Different Versions of Parkinson's Disease Rating Scales. *Pharmaceutical Research*, 36, July 2019.

[12] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, June 2000.

[13] Jason C. Immekus, Kate E. Snyder, and Patricia A. Ralston. Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research. *Frontiers in Education*, 4:45, 2019.

[14] Clifford R. Jack, David S. Knopman, William J. Jagust, Ronald C. Petersen, Michael W. Weiner, Paul S. Aisen, Leslie M. Shaw, Prashanthi Vemuri, Heather J. Wiste, Stephen D. Weigand, Timothy G. Lesnick, Vernon S. Pankratz, Michael C. Donohue, and John Q. Trojanowski. Update on hypothetical model of Alzheimer's disease biomarkers. *Lancet neurology*, 12(2):207–216, February 2013.

[15] Bruno M. Jedynak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T. Wyman, David Raunig, C. Pierre Jedynak, Brian Caffo, and Jerry L. Prince. A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort. *NeuroImage*, 63(3):1478–1486, November 2012.

[16] Igor Koval, Alexandre Bône, Maxime Louis, Thomas Lartigue, Simona Bottani, Arnaud Marcoux, Jorge Samper-González, Ninon Burgos, Benjamin Charlier, Anne Bertrand, Stéphane Epelbaum, Olivier Colliot, Stéphanie Allassonnière, and Stanley Durrleman. AD Course Map charts Alzheimer's disease progression. *Scientific Reports*, 11(1):8020, April 2021.

[17] Jacqueline K. Kueper, Mark Speechley, and Manuel Montero-Odasso. The Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review. *Journal of Alzheimer's Disease*, 63(2):423–444.

[18] Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[19] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, December 1982.

[20] F. M. Lord. *Applications of Item Response Theory To Practical Testing Problems*. Routledge, New York, July 1980.

[21] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[22] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, Werner Poewe, Brit Mollenhauer, Paracelsus-Elena Klinik, Todd Sherer, Mark Frasier, Claire Meunier, Alice Rudolph, Cindy Casaceli, John Seibyl, Susan Mendick, Norbert Schuff, Ying Zhang, Arthur Toga, Karen Crawford, Alison Ansbach, Pasquale De Blasio, Michele Piovella, John Trojanowski, Les Shaw, Andrew Singleton, Keith Hawkins, Jamie Eberling, Deborah Brooks, David Russell, Laura Leary, Stewart Factor, Barbara Sommerfeld, Penelope Hogarth, Emily Pighetti, Karen Williams, David Standaert, Stephanie Guthrie, Robert Hauser, Holly Delgado, Joseph Jankovic, Christine Hunter, Matthew Stern, Baochan Tran, Jim Leverenz, Marne Baca, Sam Frank, Cathi-Ann Thomas, Irene Richard, Cheryl Deeley, Linda Rees, Fabienne Sprenger, Elisabeth Lang, Holly Shill, Sanja Obradov, Hubert Fernandez, Adrienna Winters, Daniela Berg, Katharina Gauss, Douglas Galasko, Deborah Fontaine, Zoltan Mari, Melissa Gerstenhaber, David Brooks, Sophie Malloy, Paolo Barone, Katia Longo, Tom Comery, Bernard Ravina, Igor Grachev, Kim Gallagher, Michelle Collins, Katherine L. Widnell, Suzanne Ostrowizki, Paulo Fontoura, Tony Ho, Johan Luthman, Marcel van der Brug, Alastair D. Reith, and Peggy Taylor. The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology*, 95(4):629–635, December 2011.

[23] Răzvan V. Marinescu, Arman Eshaghi, Marco Lorenzi, Alexandra L. Young, Neil P. Oxtoby, Sara Garbarino, Sebastian J. Crutch, and Daniel C. Alexander. DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage*, 192:166–177, May 2019.

[24] P. McCullagh. *Generalized Linear Models*. Routledge, October 2018.

[25] Sean S. O'Sullivan, Andrew H. Evans, and Andrew J. Lees. Dopamine dysregulation syndrome: an overview of its epidemiology, mechanisms and management. *CNS drugs*, 23(2):157–170, 2009.

[26] Kathryn V. Papp, Dorene M. Rentz, Irina Orlovsky, Reisa A. Sperling, and Elizabeth C. Mormino. Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: The PACC5. *Alzheimer's & Dementia : Translational Research & Clinical Interventions*, 3(4):668–677, November 2017.

[27] Lars Lau Raket. Statistical Disease Progression Modeling in Alzheimer Disease. *Frontiers in Big Data*, 3, 2020.

[28] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(1):1–97, March 1969.

[29] J.-B. Schiratti, S. Allassonnière, A. Routier, O. Colliot, and S. Durrleman. A Mixed-Effects Model with Time Reparametrization for Longitudinal Univariate Manifold-Valued Data. In Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin, and M. Jorge Cardoso, editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 564–575, Cham, 2015. Springer International Publishing.

[30] Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *The Journal of Machine Learning Research*, 18(1):4840–4872, January 2017.

[31] Glenn T. Stebbins, Christopher G. Goetz, David J. Burn, Joseph Jankovic, Tien K. Khoo, and Barbara C. Tilley. How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: Comparison with the unified Parkinson's disease rating scale. *Movement Disorders*, 28(5):668–670, 2013.

[32] Bachirou O. Taddé, Hélène Jacqmin-Gadda, Jean-François Dartigues, Daniel Commenges, and Cécile Proust-Lima. Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: Application to Alzheimer's disease. *Biometrics*, 76(3):886–899, 2020.

[33] Marc Vandemeulebroecke, Björn Bornkamp, Tillmann Krahnke, Johanna Mielke, Andreas Monsch, and Peter Quarg. A Longitudinal Item Response Theory Model to Characterize Cognition Over Time in Elderly Subjects. *CPT: Pharmacometrics & Systems Pharmacology*, 6(9):635–641, September 2017.

[34] Alexandra L. Young, Razvan V. Marinescu, Neil P. Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C. Firth, David M. Cash, David L. Thomas, Katrina M. Dick, Jorge Cardoso, John van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, James B. Rowe, Caroline Graff, Fabrizio Tagliavini, Giovanni B. Frisoni, Robert Laforce, Elizabeth Finger, Alexandre de Mendonça, Sandro Sorbi, Jason D. Warren, Sebastian Crutch, Nick C. Fox, Sebastien Ourselin, Jonathan M. Schott, Jonathan D. Rohrer, and Daniel C. Alexander. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature Communications*, 9(1):1–16, October 2018. Number: 1 Publisher: Nature Publishing Group.

[35] Alexandra L. Young, Jacob W. Vogel, Leon M. Aksman, Peter A. Wijeratne, Arman Eshaghi, Neil P. Oxtoby, Steven C. R. Williams, Daniel C. Alexander, and for the Alzheimer's Disease Neuroimaging Initiative . Ordinal SuStaIn: Subtype and Stage Inference for Clinical Scores, Visual Ratings, and Other Ordinal Data. *Frontiers in Artificial Intelligence*, 4, 2021.

A Synthetic experiment complementary results

These results complement the experiment in section 3.1. In the synthetic experiment we generated data using an ordinal model and showed that the ordinal model was better at reconstructing such data than a continuous model. It is important to check that this also holds for data not generated by the ordinal model.

We used a continuous model similar to the one in section 3.1, but without the δ_k^l parameters. For each individual i generated, at each visit j for each item k we compute a noisy observation $y_ijk \in [0, 1]$. We then transform this data into ordinal data. We choose a Likert scale (0-4), so we multiply the observation by 4 and round it. Both multivariate models, ordinal and continuous, are calibrated on the data generated and we compute the reconstruction errors as in figure 4. Results are presented in figure A1 below.

B Quantitative experiment formulas

We provide here the formulas of the models used in section 3.2.3.

The continuous univariate model computes the MDS-UPDRS part II total as:

$$y_{ij} = \left(1 + (\frac{1}{p_0} - 1) \exp\left(-\frac{v_0(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0)}{p_0(1 - p_0)}\right)\right)^{-1} + \epsilon_{ij}$$

with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$.

The multivariate continuous model computes the 13 items of the MDS-UPDRS part II as follows:

$$\forall k \in [|1, 13|], y_{ijk} = \left(1 + (\frac{1}{p_k} - 1) \exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0) + w_{ik}}{p_k(1 - p_k)}\right)\right)^{-1} + \epsilon_{ijk}$$

with $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma)$.

The ordinal univariate model computes the MDS-UPDRS part II total as:

$$\forall l \in [|1, 52|], \mathbb{P}(y_{ij} \geq l) = \left(1 + (\frac{1}{p_0} - 1) \exp\left(-\frac{v_0(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0 - \sum_{m=1}^{l} \delta_0^m)}{p_0(1 - p_0)}\right)\right)^{-1}$$

The multivariate ordinal model computes the 13 items of the MDS-UPDRS part II as follows:

$$\forall k \in [|1, 13|], \forall l \in [|1, 4|], \mathbb{P}(y_{ijk} \geq l) = \left(1 + (\frac{1}{p_k} - 1) \exp\left(-\frac{v_k(e^{\xi_i}(t_{ij} - t_0 - \tau_i) + t_0 - \sum_{m=1}^{l} \delta_k^m) + w_{ik}}{p_k(1 - p_k)}\right)\right)^{-1}$$

The constant prediction model for the MDS-UPDRS part II total formulates as $\hat{y}_{ij} = y_{ij_{final}}$ where $j_{final} < j$ is the last visit known for subject i in the training data.

The linear mixed-effect model for the MDS-UPDRS part II total formulates as :

$$y_{ij} = \beta + \alpha \times t_{ij} + \beta_i + \alpha_i \times t_{ij} + \epsilon_{ij}$$

with β being the fixed intercept, α the fixed slope, $\beta_i \sim \mathcal{N}(0, \sigma_\beta)$ the random intercept, $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha)$ the random slope and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$ the gaussian noise.

C Quantitative experiment supplementary plots

We provide a visualization of the fit of univariate models in section 3.2.3 with figure C2. The spaghetti plots show the individual trajectories mapped onto the average trajectories. We do this by removing the random effects, hence only leaving the noise of individual observations as a deviation from the mean trajectory, shown in black, which accounts for the fixed effects. Note that the score we are modelling is quite noisy, with some individuals experiencing very sharp and sudden surges or collapses. This is also one of the reasons for the use of an ordinal model: the noise is more flexibly modelled as a spread of probability weights across several levels rather than a noise distribution chosen in model design (here a simple gaussian noise).

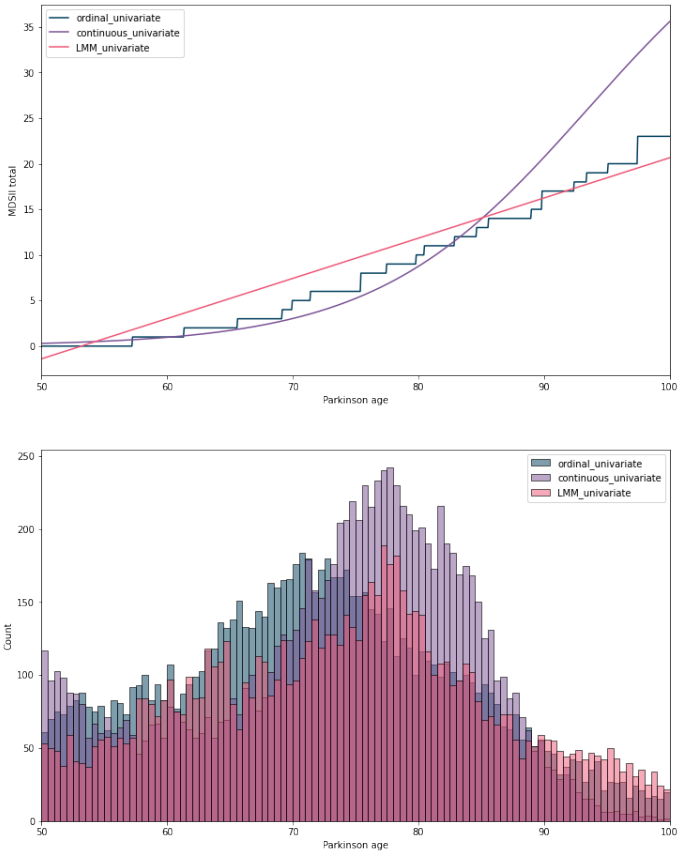

Figure 7: Top figure: average disease progression of the univariate mixed-effect models as a function of the age. Bottom figure: histograms of patients mapped onto the corresponding timeline of their models. Aligned with the top figure it shows where data points are distributed according to each model. Parkinson age corresponds to that mapping on the time axis, as a result of the individual random effects linked to time ((ξ_i, τ_i) for ordinal and continuous DCM, random slope and intercept for LMM). LMM: linear mixed-effect model; continuous: disease course mapping model; ordinal: disease course mapping model, ordinal version.

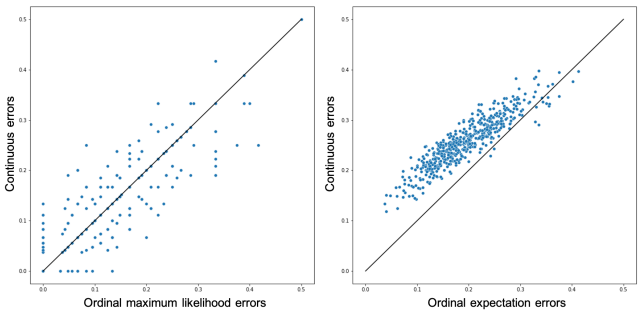

Figure 8: Error of reconstruction scatter plot. For each individual, we computed the mean absolute error of reconstruction over all visits and all items between the observations and the model prediction for the multivariate models. In the left plot the predictions of the ordinal model are the maximum likelihood estimation while the predictions of the continuous model are rounded. In the right plot for the continuous model the prediction is directly the outcome of the model, whereas for the ordinal model we computed the expectation. The right plots show that the ordinal model outperforms the continuous model for almost all patients in the reconstruction of the items.

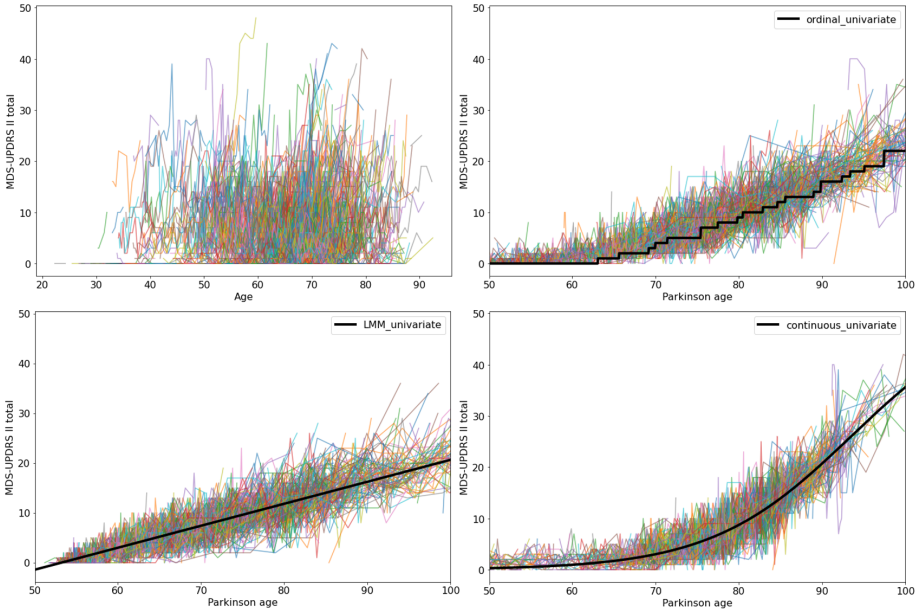

Figure 9: Spaghetti plots with individuals mapped onto the mean trajectory. Mapped trajectories are obtained by removing the random effects learned by each model. Top left: raw data; bottom left: linear mixed-effect model; top right: ordinal univariate model; bottom right: contiuous univariate DCM model with logistic curve.