Hierarchical Denoising of Ordinal Time Series of Clinical Scores

Jonathan Koss, Sule Tinaz, and Hemant D. Tagare, *Senior Member, IEEE*

***Abstract*—Clinical scores (disease rating scales) are ordinal in nature. Longitudinal studies which use clinical scores produce ordinal time series. These time series tend to be noisy and often have a short-duration. This paper proposes a denoising method for such time series. The method uses a hierarchical approach to draw statistical power from the entire population of a study's patients to give reliable, subject-specific results. The denoising method is applied to MDS-UPDRS motor scores for Parkinson's disease.**

***Index Terms*—Time series analysis, Parkinson's disease, machine learning, clinical neuroscience.**

I. Introduction

CLINICAL symptoms of neuropsychiatric and neurodegenerative disorders are often evaluated using ordinal ratings, or scores. For example, in Parkinson's Disease (PD), patients exhibit several motor disabilities, and each disability is rated using the ordinal scores: $Normal, Slight$, $Mild, Moderate, Severe$.

In longitudinal clinical studies, ordinal scores are collected at periodic intervals, producing one ordinal time series per symptom per patient. These time series are a rich source of information about how the disease is progressing, but their analysis is quite challenging. There are several reasons for this:

First, the time series are usually very noisy. Noise is partly due to the stochastic nature of the symptoms, and partly because it is difficult to differentiate between neighboring categories such as $Slight, Mild$, and $Moderate$ [1]. Noise can be so severe that it masks the underlying progression of the disease. Real-world examples of this are given in Section V.

Second, the data in the time series are ordinal. Thus, commonly used vector space techniques, such as additive noise models, cannot be used for creating denoising algorithms [2].

Third, the time series are short because most longitudinal studies have a limited time duration. A typical time series might have 10 samples. Fitting a statistical model to such a short time series becomes problematic, especially as the number of parameters in the model approaches the number of samples.

Finally, the time series are unevenly spaced and often have missing data.

A fine-grained and nuanced understanding of disease progression is possible if a method can be developed to denoise the ordinal time series of each symptom, taking all of the above-mentioned concerns into account. The goal of this paper is to propose such a denoising method.

The statistical model behind our method accounts for all of the above-mentioned characteristics of clinical scores. It explicitly takes the ordinal nature of the time series into account by using ordinal time regression. Further, the time regression is situated in a hierarchical framework, so that patients whose time series proceed similarly draw statistical power from each other, overcoming the limitation of the short duration of each series. And the model allows for uneven time spacing and missing data. In Section V, we show that using this model to denoise clinical time series helps substantially with understanding the progression of symptoms.

There is another important aspect of disease progression: Disease progression is often heterogeneous. That is, different patients progress along different trajectories at different rates. Heterogeneity is usually understood in terms of *progression subtypes*, which are prototypes of different trajectories of progression. The hierarchical design of our model automatically provides progression subtypes.

We evaluate the performance of our method by using synthetic data and also by applying the method to real-world PD data. The PD data come from the Parkinson's Progression Marker's Initiative (PPMI). Details of the PPMI data are given in Section II.

This paper is meant to be methodological. Our goal is to report a new denoising method and to demonstrate its utility by applying it to some aspects of PD. In forthcoming publications, we hope to report more comprehensive insights into PD and other diseases using our method.

The rest of the paper is organized as follows: Section II contains a literature review and background information on PD; Sections III and IV describe our model; Section V contains results of applying our method to synthetic and real-world data; Section VI contains a discussion and a conclusion.

Manuscript received August 2, 2021; revised February 13, 2022; accepted March 20, 2022. Date of publication March 29, 2022; date of current version July 4, 2022. This work was supported by the NIH (NINDS) under Grant R01NS107328. *(Corresponding author: Jonathan Koss.)*

Jonathan Koss is with the Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA (e-mail: jonathan.koss@yale.edu).

Sule Tinaz is with the Department of Neurology, Yale School of Medicine, New Haven, CT 06520 USA (e-mail: sule.tinaz@yale.edu).

Hemant D. Tagare is with the Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT 06520 USA (e-mail: hemant.tagare@yale.edu).

Digital Object Identifier 10.1109/JBHI.2022.3163126

II. Background and Literature Review

A. Ordinal Regression Models

Ordinal regression models have a long history, originating with the Cox model [3] from over 50 years ago. A complete review of ordinal regression models can be found in [4]. Textbooks such as [5] and [6] contain detailed explanations of ordinal data analysis.

Modern ordinal models can be divided into three categories [4]. In the first category are naive approaches which do not account for the ordinality of the data. They either treat clinical scores as cardinal numbers or as classes but without any order. These approaches then apply a traditional regression technique such as support vector machines (SVM) [7], [8]. Methods like [9] use a categorical classification technique but then alter the cost function to assume label distances corresponding to the cardinal values of the categories.

The second category of models decomposes the ordinal-regression problem into several binary regression problems which can be solved separately with multiple models, typically using SVMs [10] or using a single multi-output model such as a neural network [11], [12]. Common choices for decomposing the problem are comparing all pairs of neighboring categories, e.g. for ordinal classes $c_1, \ldots, c_4$ comparing $\{c_1 \text{ vs } c_2\}, \{c_2 \text{ vs } c_3\}, \{c_3 \text{ vs } c_4\}$, or comparing a single category with all the following (or previous) categories $\{c_2 \text{ vs } c_1\}, \{c_3 \text{ vs } c_{1,2}\}, \{c_4 \text{ vs } c_{1,2,3}\}$.

The last category of ordinal-regression models treat the ordinal variable as originating from a continuous underlying latent variable, which is divided into categories by a set of thresholds. One of the early and prominent threshold techniques is the proportional odds model [13], which is also the basis of our model. The proportional-odds model and other related models are discussed immediately below. Other threshold-based models like [14], [15] utilize SVMs while [16], [17] offer neural-network-based approaches.

Proportional Odds Model: The proportional odds model extends the linear logistic model [3] to handle more than two categories. For an ordinal response variable $Y \in \{0, 1, \ldots C-1\}$ and an explanatory variable X, the proportional odds model assumes that the log-odds ratio varies linearly with X:

$$\log\left(\frac{P(Y \leq c)}{P(Y > c)}\right) = \alpha_c + \beta X, \tag{1}$$

which in turn implies

$$P(Y \geq c) = \frac{1}{1 + \exp\left(-\alpha_c - \beta X\right)} \tag{2}$$

where α_c and β are scalar parameters learned from the data.

In the proportional odds model, β is independent of the category, c [18]. If the data do not meet this assumption, then the partial proportional odds model is likely to be more useful [19], [20].

Partial Proportional Odds Model: The partial proportional odds model takes the form

$$P(Y \geq c \mid X) = \frac{1}{1 + \exp\left(-\alpha_c - \beta X - \gamma_c X\right)}. \tag{3}$$

In this model, the β term of the proportional odds model is effectively set to $\beta + \gamma_c$, where γ_c is category dependent. If all $\gamma_c \approx 0$, then (3) simplifies to the proportional odds model.

The generalized ordered logit model [21] offers a model with similar properties.

B. Parkinson's Disease

Parkinson's disease (PD) is a neurodegenerative disorder that manifests with motor and non-motor symptoms. In the early stages of the disease, the motor symptoms tend to be prominent [22]. The current standard for clinically assessing disease severity in PD is the Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [23]. The MDS-UPDRS has four parts, each consisting of a set of items scored on an ordinal scale of 0, 1, 2, 3, 4 which correspond to *Normal, Slight, Mild, Moderate, and Severe* ratings. In the first two parts of the MDS-UPDRS, patients and/or caregivers rate symptoms associated with non-motor and motor aspects of daily living, respectively. Part 3 is the motor examination portion of the MDS-UPDRS, where a physician observes a patient performing a number of motor tasks and scores each on an ordinal scale. Part 3 contains 33 ordinal scores, which mainly focus on the cardinal motor features of PD including tremor, rigidity, bradykinesia (slowness of movement), and axial symptoms (e.g., gait impairment, postural instability). Part 3 scores are usually summed to create a total movement score. Higher total movement scores indicate greater overall motor severity. Part 4 assesses motor complications including dyskinesia and motor fluctuations in response to medication.

In this paper we focus on Part 3 scores. Rather than summing up the scores into a total movement score, we analyze scores of the cardinal PD symptoms independently.

PD is a heterogeneous disease with distinct clinical subtypes and progression trajectories [24]. These subtypes are traditionally determined by clinical observations (e.g., tremor-dominant or postural instability/gait difficulty subtypes) [25]. More recent data-driven clustering methods identify the following subtypes: 1) young-age onset and old-age onset subtypes with the latter progressing relatively faster than the former, but both progressing slowly 3) tremor-dominant, and 4) dominance of bradykinesia, rigidity, and postural instability/gait difficulty [26].

C. Parkinson's Progression Markers Initiative (PPMI)

The MDS-UPDRS data that we use comes from the de novo PD cohort of the Parkinson's Progression Markers Initiative (PPMI), a longitudinal study with the goal of finding biomarkers for PD (www.ppmi-info.org) [27].

PPMI contains MDS-UPDRS data for 423 patients who were observed for 8 years after the initial PD diagnosis. MDS-UPDRS scores are available for these patients for up to 17 visits. At the start of the study, visits are scheduled every three months, decreasing in frequency to every six months, and then to once a year. During the visits, MDS-UPDRS scores are gathered in the off-medication state, which is defined as the condition where the patient has not taken medication for ≥ 6 hours, so that their symptoms are not masked by medication.

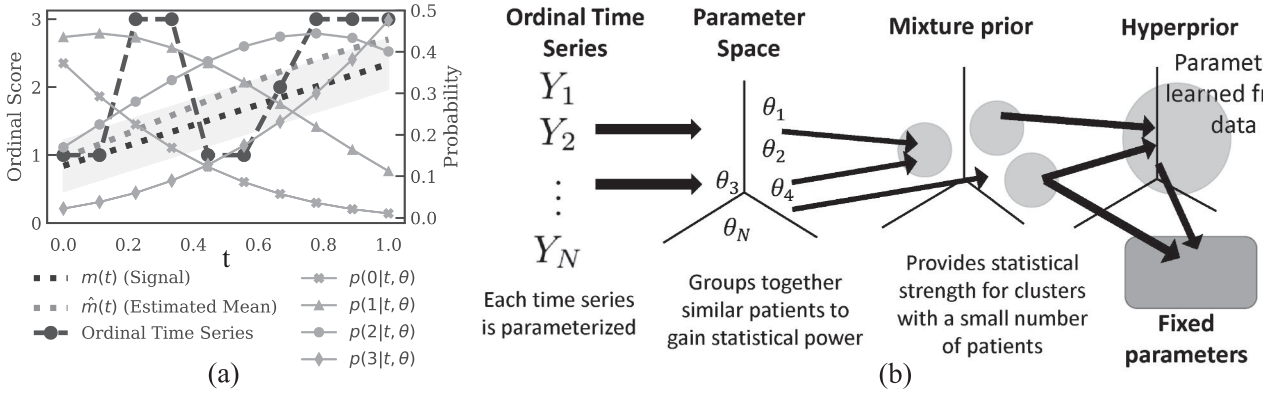

Fig. 1. (a) Ordinal time series sampled from probabilities ($p(c|t,\theta), c = 0,\ldots,3$) which vary over time along with the mean signal ($m(t)$) and an estimated mean ($\hat{m}(t)$) and standard deviation ($\hat{s}_i(t)$) shaded around it. (b) Overview of the hierarchical model proposed in this paper. Y_i is an ordinal time series for subject i and θ_i is its subject-specific set of parameters. The prior and the hyper-prior allow the model to draw statistical power from all subjects.

III. Denoising an Ordinal Time Series

When a time series takes values in a vector space, noise is usually assumed to be additive and Gaussian. Given a prior for the underlying signal, noisy measurements of the time series are used to calculate the posterior distribution of the signal, and the mean of the posterior distribution is taken as the denoised estimate of the signal.

This vector space idea does not carry over to ordinal time series in a straightforward manner because ordinal time series do not belong to a vector space and there is no simple additive noise model for such series. Nevertheless, there is a sense in which generalization is possible. The idea is illustrated in Fig. 1. Assuming 4 ordinal categories for the moment, Fig. 1(a) shows how the probabilities of observing an ordinal category change with time as the disease progresses. The probability of observing ordinal category $y = 0, 1, 2, 3$ at time t is $p(y|t,\theta)$, where θ is a set of parameters. Thus, $p(0|t,\theta)$ is initially large, followed by $p(1|t,\theta)$ being large, etc. The progress of the disease can be summarized by the mean $m(t) = \sum_{y=0}^{3} yp(y|t,\theta)$, which we may take as the underlying "signal" [See Fig. 1(a)]. Any realization of the time series is a sequence of ordinal values at times $t_1, t_2, \cdots$, drawn from the probabilities $p(y|t,\theta)$.

Assuming a prior on θ, we may estimate its value $\hat{\theta}$ from the observed time series, and use the estimate to obtain a denoised estimate $\hat{m}(t) = \sum_{y=0}^{3} yp(y|t,\hat{\theta})$ of the mean. This is the sense in which we can denoise an ordinal time series of clinical symptoms.

As mentioned in Section I, if the time series is short, we may not get a reliable estimate of θ. To ameliorate this problem, we use a hierarchical model in which θ_i for the time series of the ith subject is assumed to have a mixture of normals prior [see Fig. 1(b)]. All subjects with similar underlying disease progression are assumed to come from a single component of the mixture. Simultaneously estimating θ_i's and the parameters of the mixture allows the θ_i's to draw statistical power from similarly progressing subjects.

TABLE I
Descriptions of Variables Used in Our Model

Parameter	Update
θ_i	$\theta_i^{[n+1]} = \arg\max_{\theta_i} Q(\phi \mid \phi^{[n]})$
μ_k	$\mu_k^{[n+1]} = \frac{1}{\frac{1}{\rho_\mu^{[n]}} \sum_{i=1}^{N} w_{i,k}^{[n]} + \frac{1}{\rho_\zeta^{[n]}}}$ $\times\left(\frac{1}{\rho_\mu^{[n]}} \sum_{i=1}^{N} \theta_i^{[n]} w_{i,k}^{[n]} + \frac{1}{\rho_\zeta^{[n]}} \zeta^{[n]}\right)$
ρ_μ	$\rho_\mu^{[n+1]} = \sum_i \sum_k \frac{\frac{\|\theta_i^{[n]} - \mu_k^{[n]}\|^2}{2} w_{i,k}^{[n]} + b}{\frac{DN}{2} + a + 1}$
α_k	$\alpha_k^{[n+1]} = \frac{\sum_{i=1}^{N} w_{i,k}^{[n]}}{\sum_{i=1}^{N} \sum_{k=1}^{K} w_{i,k}^{[n]}}$
ζ	$\zeta^{[n+1]} = \frac{1}{K} \sum_{k=1}^{K} \mu_k^{[n]}$
ρ_ζ	$\rho_\zeta^{[n+1]} = \frac{1}{DK} \sum_{k=1}^{K} \|\mu_k^{[n]} - \zeta^{[n]}\|^2$

It is not uncommon in mixture models to have some components with only a few data points. To estimate these components reliably, we draw statistical power from the remaining components by introducing a hyper-prior which relates all components to each other [Fig. 1(b)].

The θ_i's of individual subjects, and parameters of the mixture prior, and of the hyper mixture prior, form the total set of parameters to be estimated from the set of all time series of a symptom for all subjects. After estimating the parameters, the time series for subject i is denoised as $\hat{m}_i(t) = \sum_{y=0}^{C-1} yp(y|t,\hat{\theta}_i)$.

IV. Model

We now explain mathematical details of the model. A glossary of variables used in this section can be found in Table I.

A. Ordinal Random Variables

Suppose y is an ordinal random variable taking values in ordinal categories $\{0,\ldots,C-1\}$. A common model [19], [21]

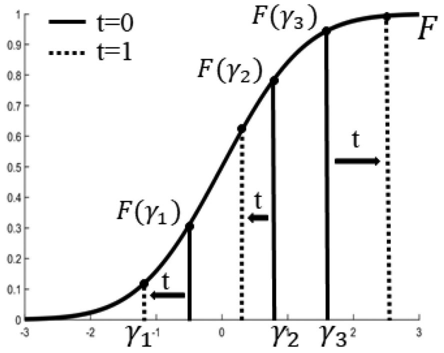

Fig. 2. Probabilities for an ordinal variable taking four values. The probabilities are created by partitioning [0,1] using a cdf and three thresholds $\{\gamma_1, \gamma_2, \gamma_3\}$ which vary over the time interval, [0,1].

for the probability distribution of y is

$$p(y \mid \gamma) = \begin{cases} F(\gamma_1) & \text{for } y = 0 \\ F(\gamma_2) - F(\gamma_1) & \text{for } y = 1 \\ \cdots & \\ F(\gamma_k) - F(\gamma_{k-1}) & \text{for } y = k \\ \cdots & \\ 1 - F(\gamma_{C-1}) & \text{for } y = C-1, \end{cases} \tag{4}$$

where F is the cumulative distribution function (cdf) of the standard normal density and $\gamma = (\gamma_1, \ldots, \gamma_{C-1}) \in R^{C-1}$ is a parameter of the distribution and is subject to the constraint

$$\gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \cdots \leq \gamma_{C-1}. \tag{5}$$

Fig. 2 illustrates the model. $F(\gamma_1), \ldots, F(\gamma_{C-1})$ are thresholds which divide the interval [0,1] into subintervals representing the probabilities of y. The constraints of (5) guarantee that the probabilities satisfy $p(y|\gamma) \geq 0$ for all y and $\sum_{y=0}^{C-1} p(y|\gamma) = 1$.

The set of γ satisfying the constraints of (5) form a proper subset $\Omega \subset R^{C-1}$. The boundary of Ω corresponds to one or more of the inequalities in (5) being exact. The interior of Ω, denoted int(Ω), is not empty. It corresponds to all γ that satisfy (5) with strict inequalities.

For reasons that will become clear below, it is useful to implicitly impose the constraints of (5) by a change of variables. Define a map $\Gamma : \mathbb{R}^{C-1} \to \text{ int } (\Omega)$ which maps $u = (u_1, \ldots, u_{C-1}) \in \mathbb{R}^{C-1}$ to $\gamma = (\gamma_1, \ldots, \gamma_{C-1}) \in \text{ int } (\Omega)$ as follows:

$$\begin{aligned} \gamma_1 &= u_1 \\ \gamma_k &= u_1 + \sum_{i=2}^{k} \log(1 + e^{u_i}), \text{ for } k = 2, \ldots, C-1. \end{aligned} \tag{6}$$

Since the function $\log(1 + e^x) > 0$ for all x, and is a bijection from $\mathbb{R}$ to $\mathbb{R}_{++}$, the γ's on the left hand side of (6) automatically satisfy (5) for any $u \in \mathbb{R}^{C-1}$. Thus, we may write the ordinal probability model of (4) as $p(y \mid \Gamma(u))$ with no constraints on u.

B. Ordinal Time Series

The above model extends easily to a time-series model. Suppose $Y = (y_1, \ldots, y_T)$ is a discrete ordinal time series with $y_t \in \{0, \ldots, C-1\}$ for $t = 1, \ldots, T \in [0, 1]$. Then, the model for the time series is

$$p(Y \mid u, v) = \prod_{t=1}^{T} p(y_t \mid \Gamma(u)(1-t) + \Gamma(v)t), \tag{7}$$

where $u, v \in \mathbb{R}^{C-1}$ generate the thresholds γ that divide the probabilities at the start and end of the time interval [0,1]. This model effectively assumes a linear change over time of the u value used in $p(y \mid \Gamma(u))$. The model can be made more complex by making the change non-linear, but for slowly progressing diseases such as PD, a linear change is sufficient. Note that although the u's change linearly with time, the probability values themselves change non-linearly with time, because the cdf is a nonlinear function. Also note that (7) does not assume t is evenly spaced allowing for uneven time sampling and missing data.

We define $\theta = (u, v)$. Since u and v have dimension $C - 1$, θ has dimension $D = 2C - 2$.

C. Multiple Time Series

Suppose there are N subjects, with time series $Y_i = (y_{i,1}, \ldots, y_{i,T_i})$, $i = 1, \ldots, N$. Assuming that the ith subject has the subject's own parameter $\theta_i = (u_i, v_i)$,

$$p(Y_i \mid \theta_i) = \prod_{t=1}^{T_i} p(y_{i,t} \mid \Gamma(u_i)(1-t) + \Gamma(v_i)t) \tag{8}$$

D. Prior on θ_i

As mentioned above, we set the prior for θ_i to a normal mixture with K components. To limit the number of parameters in the model, all components of the mixture are assumed to have the same diagonal covariance matrix $\rho_\mu I$, with $\rho_\mu > 0$. Letting z_i be the latent variable used to identify which component of the mixture θ_i comes from, the prior is given by

$$p(\theta_i \mid z_i = k, \{\mu_k\}, \rho_\mu) = \mathcal{N}(\mu_k, \rho_\mu I), \tag{9}$$

where $z_i = k$ indicates that θ_i belongs to the kth component. The set of means $\{\mu_k\}$ and ρ_μ are the parameters of the prior distribution. The dimension of μ_k is D.

As is standard with mixture models, we set the prior on the latent variable z_i to $p(z_i = k \mid \{\alpha_k\}) = \alpha_k$, where $\alpha_k \geq 0$, and $\sum_k \alpha_k = 1$. Here α_k's are the mixing coefficients.

Each component of the normal mixture has support over all of $\mathbb{R}^D$. The change of variables proposed in (6) allows us to use such infinite support priors, which is the reason for introducing that change of variables.

E. Hyper-Prior on $\{\mu_k\}$ and ρ_μ

We set the hyper-prior on $\{\mu_k\}$ to a normal distribution, and the hyper-prior on ρ_μ to an inverse-gamma distribution:

$$p(\mu_k \mid \zeta, \rho_\zeta) = \mathcal{N}(\zeta, \rho_\zeta I), \tag{10}$$

$$p(\rho_\mu \mid a, b) = IG(a, b) = \frac{b^a}{\Gamma(a)} \frac{1}{(\rho_\mu)^{(a+1)}} e^{\frac{-b}{\rho_\mu}}. \tag{11}$$

The hyper-prior parameters a, b, ρ_ζ are fixed and set by a procedure described later. This normal-inverse-gamma hyperprior is conjugate to the components of the normal mixture.

F. The Full Model

Grouping together all parameters to be estimated as $\phi = (\{\theta_i\}, \{\mu_k\}, \rho_\mu, \{\alpha_k\}, \zeta)$, we get:

$$\begin{aligned} p(\{Y_i\}\{z_i\}|\phi) = &\prod_{i=1}^{N} p(Y_i|\theta_i) p(\theta_i|z_i, \{\mu_k\}, \rho_\mu) \\ &\times \left(\prod_{k=1}^{K} p(\mu_k|\zeta, \rho_\zeta) \right) \\ &\times p(\rho_\mu|\alpha, \beta) \times \prod_{i=1}^{N} p(z_i|\{\alpha_k\}). \end{aligned} \tag{12}$$

Marginalizing out the latent variables $\{z_i\}$ gives

$$p(\{Y_i\} \mid \phi) = \sum_{z_1=1}^{K} \cdots \sum_{z_N=1}^{K} p(\{Y_i\}\{z_i\} \mid \phi), \tag{13}$$

from which the maximum-likelihood estimates of ϕ are obtained as $\hat{\phi} = \arg\max_\phi \log p(\{Y_i\} \mid \phi)$. The maximum-likelihood estimates are easily calculated by the EM-algorithm whose details are given below.

The maximum likelihood estimates of $\hat{\theta}_i = (\hat{u}_i, \hat{v}_i)$ give the estimated mean and variance:

$$\begin{aligned} \hat{m}_i(t) &= \sum_{y=0}^{C-1} y p(y_t|\Gamma(\hat{u}_i)(1-t) + \Gamma(\hat{v}_i)t) \\ \hat{s}_i^2(t) &= \sum_{y_t=0}^{C-1} (y_t - \hat{m}_i(t))^2 p(y_t|\Gamma(\hat{u}_i)(1-t) + \Gamma(\hat{v}_i)t). \end{aligned} \tag{14}$$

The estimated mean is the denoised signal.

G. The EM Algorithm

The Q-function for the EM algorithm which maximizes the log-likelihood for the model of (13) is:

$$\begin{aligned} Q(\phi \mid \phi^{[n]}) = &\sum_{i=1}^{N} \log P(Y_i|\theta_i^{[n]}) \\ &+ \sum_{i=1}^{N} \sum_{z_i=1}^{K} \log p(\theta_i, z_i|\{\mu_k\}^{[n]}, \rho_\mu^{[n]}) \\ &p(z_i|\theta_i^{[n]}, \{\mu_k\}^{[n]}, \rho_\mu^{[n]}) \\ &+ \sum_{k=1}^{K} \log p(\mu_k|\zeta^{[n]}, \rho_\zeta^{[n]}) + K \log p(\rho_\mu \mid a, b), \end{aligned} \tag{15}$$

TABLE II
EM-ITERATIONS SUPERSCRIPT $\theta_i^{[n]}$ INDICATES THE VALUE OF θ_i AT ITERATION n

Variable	Description
Y_i	Ordinal Time Series
$\hat{m}_i$	Denoised time series
$\hat{s}_i$	Standard deviation of denoised time series
u_i	Parameters in eq. (6) at t=0
v_i	Parameters in eq. (6) at t=1
θ_i	(u_i, v_i)
μ_k	Mixture prior means
ρ_μ	Mixture prior variance
K	Number of clusters in the mixture prior
z_i	Latent variable indicating mixture component
α_k	Mixture coefficients
ζ	Hyperprior mean
ρ_ζ	Hyperprior variance
a, b	Inverse-gamma prior parameters

where

$$\begin{aligned} p(z_i \mid \theta_i^{[n]}, \{\mu_k\}^{[n]}, \rho_\mu^{[n]}) &= \frac{p(\theta_i, z_i = k|\{\mu_k\}^{[n]}, \rho_\mu^{[n]})}{\sum_{k=1}^{K} p(\theta_i, z_i \mid \{\mu_k\}^{[n]}, \rho_\mu^{[n]}))} \\ &= w_{i,k}^{[n]}, \end{aligned} \tag{16}$$

and the superscript $[n]$ refers to values at the nth iteration.

The iterations of the EM algorithm are given in Table II. All updates have closed form solution except those of θ_i. The θ_i updates are obtained by numerical gradient ascent using Adam [28] on those terms of the Q-function that depend on θ_i. The EM updates are iterated until convergence.

H. Initialization

To initialize the EM algorithm, the parameters $\theta_i = (u_i, v_i)$ are estimated for each subject by maximizing the log-likelihood $\log p(Y_i|\theta_i)$ of (8) using Adam. That is, the hierarchical structure is ignored. Then the k-means algorithm is used to group the estimates $\{\hat{\theta}_i\}$ into K clusters whose means and isotropic variance are taken as initial estimates of μ_k and ρ_μ. The hyperparameters ζ and ρ_ζ are initialized as

$$\hat{\zeta} = \frac{1}{N} \sum_{n=1}^{N} \hat{\theta}_i, \text{ and } \hat{\rho_\zeta} = \frac{1}{N} \sum_{n=1}^{N} (\hat{\zeta} - \hat{\theta}_i)^2. \tag{17}$$

I. Cross Validation

We set the hyper parameter $a = 2.0$ (informal experimentation showed that the results are insensitive to changes in a). Then, the number of clusters K and the parameter b are determined by cross-validation. The number of clusters varies from 0 to 5. The 0 cluster corresponds to ignoring the prior and fitting (8) individually to each subject.

Given a data-set of ordinal time series of a symptom for multiple subjects, 90% of the time series are set aside as training data and 10% set aside as test data. For each series in the training data, 10% of the data points are set aside for cross-validation, and 10$-$fold cross validation is used to determine K and b. The log-likelihood of the cross-validation data are assessed using (8) for the value of t corresponding to the cross-validation

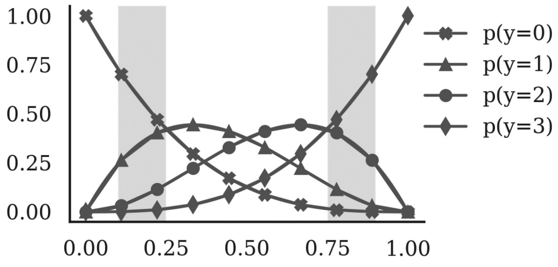

Fig. 3. Bernstein polynomials of degree 3 over the interval [0,1]. u and v are selected from the gray regions for map 1 and map 2.

TABLE III
CROSS VALIDATION RESULTS FOR THE SIMULATION AND ALGORITHM EXECUTION TIMES

K	0	1	**2**	3	4	5	Linear
b	-	270	**245**	200	215	250	Fit
ℓ_{train}	-2.06	-1.28	**-1.24**	-1.30	-1.40	-1.41	-
ℓ_{test}	-2.00	-1.28	**-1.20**	-1.20	-1.24	-1.29	-
$fMSE_{train}$	0.073	0.060	**0.027**	0.056	0.062	0.063	0.066
$fMSE_{test}$	0.067	0.073	**0.021**	0.021	0.036	0.056	0.051
EM Iter.	-	43	**81**	84	45	58	-
Time (s)	6.3	19.6	**22.2**	21.7	18.6	18.1	-

K is the number of clusters in the mixture prior ($K = 0$ represents no prior). b is a parameter of the inverse-gamma hyperprior. $\ell_{train}, \ell_{test}$ are the log-likelihood of the probability model of left-out data from (8). $fMSE_{train}$ and $fMSE_{test}$ are the normalized mean squared error values From the ground truth for the training and test sets. performance of linear regression is also included as a benchmark. the column with $k = 2$ is in bold font because it has the highest training and test Log-Likelihood.

time points. The values of K and b maximizing the validation log-likelihood are chosen.

J. Denoising

After the EM iterations converge, the $\hat{u}_i, \hat{v}_i$ estimates (obtained from $\hat{\theta}_i$ estimates) are used to calculate the denoised time series and its variance from (14).

K. Processing the Test Set

After all parameters are estimated from the training set, the estimated prior and hyperprior parameters are fixed and used with the test set. 10% of the data points in the test set are set aside to evaluate the performance of the model. The θ_i's for the test set, which are subject specific, are estimated via the update rule shown in Table II, using Adam.

L. Evaluating the Model

Note that 10% of the data are left out in the training set and in the test set. After fitting the model to the training and the test set as described above, we evaluate the performance of the model by calculating the log likelihood of the left-out data (using (8) on the left-out data) for both sets. This gives a fair comparison of how well the model fits the data, irrespective of whether the prior is used ($K = 0$) or not ($K > 0$).

V. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our hierarchical model we apply it to two data sets. The first is a simulated data set, which allows us to test the model's performance against ground truth. The second data set comes from PPMI MDS-UPDRS Part 3 scores.

A. Simulated Data

The idea behind the simulation is to create ordinal time series with two progression subtypes: one in which the probability of observing higher ordinal classes increases with time, and one in which the probability decreases with time. To achieve this, we used Bernstein polynomials of degree three defined over $I = [0, 1]$ as shown in Fig. 3 [29]. The polynomials, denoted $b_0, \ldots, b_3$ have the property that they are non-negative over [0,1] and at every point in I they sum to 1.

The polynomials were converted into the two types of ordinal time series by mapping $t \in I$ into the interior of I by one of two affine maps ϕ:

$$\phi(t) = \begin{cases} u(1-t) + vt \text{ (map 1) or,} \\ v(1-t) + ut \text{ (map 2),} \end{cases} \tag{18}$$

where u is randomly chosen from [0.1,0.25] and v is randomly chosen from $[0.75, 0.9]$. Thus map 1 maps I to the interior of I in monotonically increasing fashion, while map 2 maps in a monotonically decreasing fashion. Therefore $p_i(t) = b_i(\phi(t)),\ \tilde{}i = 0, \ldots, 3$ are ordinal class probabilities where the chances of observing higher ordinal classes increase with time for map 1, and decrease with time for map 2. Finally $t \in [0, 1]$ was sampled at 11 uniform points and the probabilities p_i at these points are used to generate the ordinal time series.

A total of 300 ordinal time series were used to train our model with 150 time series coming from each map. Another 50 time series from each map were used as a test set.

Results from cross-validation are given in Table III. The l_{train} row shows the log-likelihood of the left-out data in the training set. The l_{test} row shows the the log-likelihood of the left-out data in the test set. Note that l_{test} is similar to l_{train} showing that the model does not overfit for any K. The maximum value of l_{train} is obtained for $K = 2$, showing that the method correctly identifies the number of progression subtypes.

Using $K = 2$ and the corresponding b value, the EM-algorithm was used to estimate the rest of the parameters from the training set. Then, the prior and hyper-prior parameters $(\{\mu_k\}, \rho_\mu, \{\alpha_k\}, \zeta)$ were fixed to their estimates from the training set, and θ_i determined for the test set. Fig. 4 shows two typical ordinal time series from the simulation, their true and estimated probabilities, true and estimated mean, and true and estimated standard deviation.

The algorithm results were evaluated further for the model with the best K and b from cross-validation. First, we evaluated in detail whether the model overfit the training data. For this, we calculated the L_1 norm between the true and the estimated probabilities p_i for the left-out points in the training and test sets. Fig. 5 shows the histogram of the L_1 norms. Clearly, the two histograms are quite similar. Second, we calculated the histograms of the log-likelihood of the left-out points. These histograms are

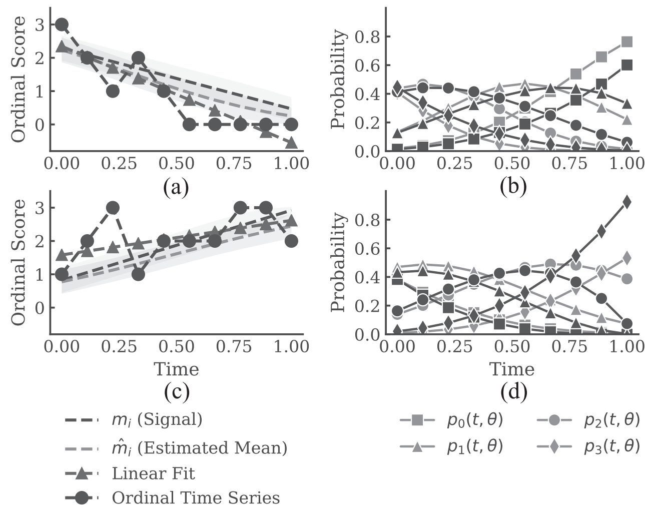

Fig. 4. Trajectories and probabilities of simulated time series. (a, c) Comparison of the synthetic time series with the expected ground truth trajectory, the denoised trajectory, and a linear fit. The shaded regions represent the standard deviation of the ordinal time series and $\hat{s}_i(t)$, the standard deviation of the estimated ordinal probabilities, $\hat{m}_i(t)$. (b, d) Comparison of the learned probabilities (black) and the true probabilities (gray).

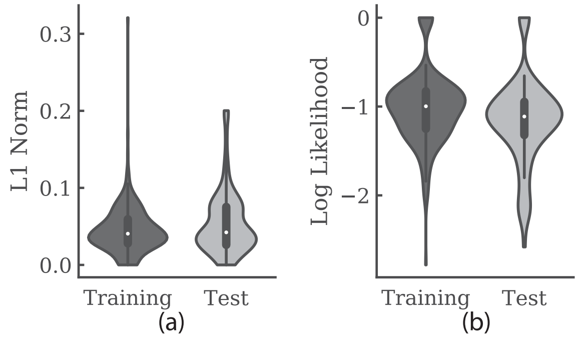

Fig. 5. (a) L_1 norm between estimated and real probabilities for the left-out points in the simulated training and test sets. (b) Log-likelihood as calculated from (8) for the left-out points in simulated training and test sets.

shown in Fig. 5(b). The training and test histograms are similar as well, showing that the model does not overfit the data.

Third, we evaluated the relative accuracy of our model by comparing it to linear regression as a control. To do this, we fit a linear regression model to each ordinal time series. Then we calculated the fractional mean-square-error of our model and of the linear regression. The fractional mean-square-error (fMSE) is defined as the ratio of the mean-square-error of the model-based denoised mean, or the linear regression, to the true mean divided by the L_2 norm square of the true mean ($\frac{\sum_{t=0}^{10}\|\mu(t)-\hat{\mu}(t)\|^2}{\sum_{t=0}^{10}\|\mu(t)\|^2}$). The training and test fMSEs for the left-out points are given in Table III. As the table shows, our model with the optimal K and b values has an average $fMSE$ of 0.027 for the training set and 0.021 for the test set, each corresponding to a nearly 60% reduction in error compared to linear regression. Fig. 4 shows linear fits to two simulated time series.

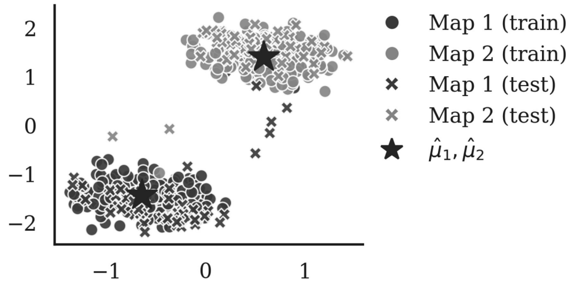

Fig. 6. Each point represents a θ embedded in 2-dimensional space using multi-dimensional scaling. Black points correspond to time series drawn from map 1 while black points are drawn from map 2. The shape of the points represent whether they are training or test time series. Gray stars indicate the values of μ_k, the means of the mixture prior, after multi-dimensional scaling.

Finally, we evaluated the model's ability to capture the two subtypes of ordinal time series. For this, we took the estimated $\hat{\theta}_i$'s for each time series and embedded them in a 2-dimensional space using multi-dimensional scaling [30]. Fig. 6 shows the embedded $\hat{\theta}_i$'s shaded with the type of map (of (18)) that the series was generated with. Clearly, the estimated $\hat{\theta}_i$'s have captured the "bi-modality" of the simulated data generation. Fig. 6 also shows that the estimated means, $(\hat{\mu}_1, \hat{\mu}_2)$ of the mixture of normals prior lie within the two clusters indicating natural progression subtypes for the time series.

B. Parkinson's Disease

Next, we analyzed data from PPMI. We used motor exam scores from MDS-UPDRS Part 3. After excluding patients with fewer than four visits, there were 377 PD patients with 10.4 ± 3.1 (mean ± stdev.) visits. Instead of summing all scores into a total movement score, we separately analyzed the time series of one symptom for each of the three cardinal symptoms of PD (rigidity, tremor, and bradykinesia). We analyzed scores for left-arm rigidity, right-leg rest tremor amplitude and right-leg agility (a measure of bradykinesia). Since PPMI data focuses on early-stage PD patients, scores of 4 were uncommon, and thus we combined categories 3 and 4 into a single category named 3.

Cross validation revealed optimal values of $K = 2$ and $b = 1167$ for left-arm rigidity, $K = 2$ and $b = 750$ for right-leg rest tremor amplitude, and $K = 1$ (single mixture component) and $b = 950$ for right-leg agility. Using the cross-validated parameters, the model was fit to the training set for each question. After training, we retained the trained prior and hyperprior parameters, and the model was again fit to the test set. The algorithm typically converged in 50-100 EM iterations, which took less than a minute for a MATLAB implementation. Fig. 7(a,c,e) show histograms of the log-likelihoods of the training left-out data and of the test data. The histograms indicate that there is no overfitting.

Fig. 7(b, d, f) show scatter plots of training and test $\hat{\theta}_i$'s along with the prior means $\hat{\mu}_k$ all projected via multi-dimensional scaling onto a 2-dim plane. Note that the $\hat{\theta}_i$'s have a greater dispersion between the clusters than the simulated data. Because

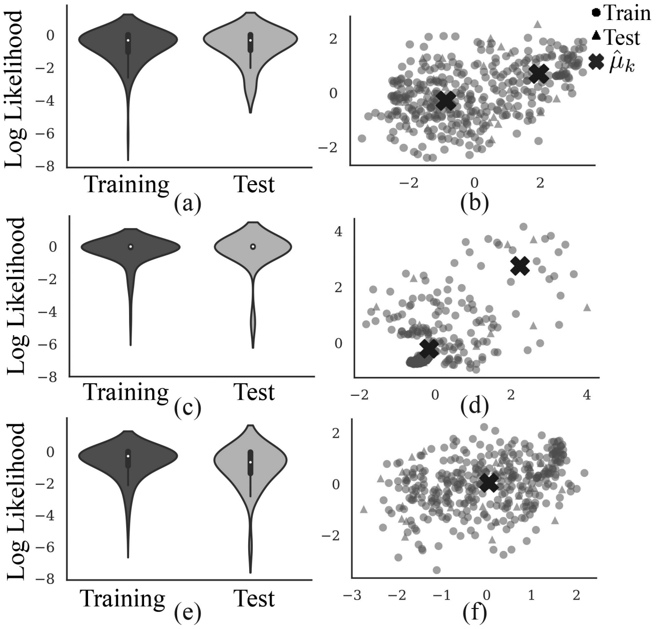

Fig. 7. (a, c, e) Comparisons of the distributions of log-likelihood of left-out data for training and test data. (b, d, f) θ_is embedded in 2-dim space with multi-dimensional scaling. (a, b) Right-Leg Rest Tremor Amplitude. (c, d) Left-Arm Rigidity. (e, f) Right-Leg Agility.

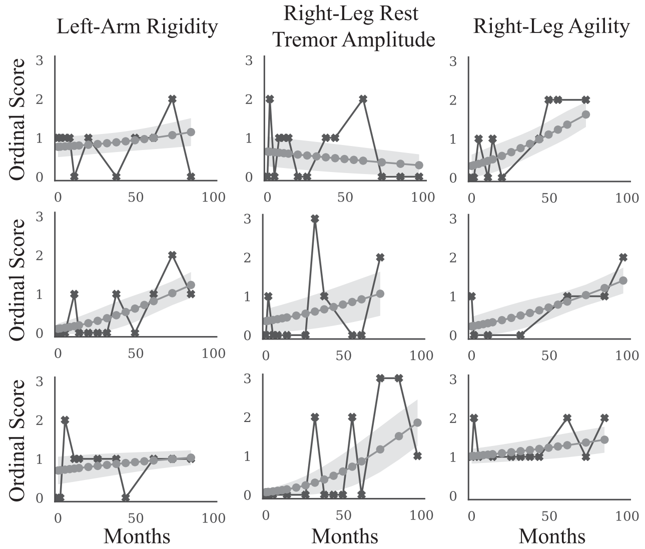

Fig. 8. Example ordinal time series (black) and their corresponding estimated means, $\hat{m}_i(t)$ (gray). The shaded region represents the standard deviation of the estimated ordinal probabilities, $\hat{s}_i(t)$.

the cluster variance is estimated using the data by our method, such dispersion is handled automatically.

Fig. 8 shows typical time series along with their estimated means and standard deviations. It is clear from the figure that the denoised time series capture the progression of the symptom. Some patients appear to progress more rapidly compared to others.

The patterns revealed by denoising are displayed in Figs. 9–11. The (a) part of Fig. 9 shows raw left-arm rigidity scores of 20 randomly chosen subjects. Any disease progression pattern in this cohort is difficult to discern. The (b) part of the figure shows denoised scores for the same subjects along with trajectories corresponding to the means of the two components in the mixture prior. The latter are the progression subtypes for this score. The disease progression is quite clear in the denoised scores. Also, the subtypes clearly show that this symptom steadily progresses with one group of subjects starting at a higher value relative to the other.

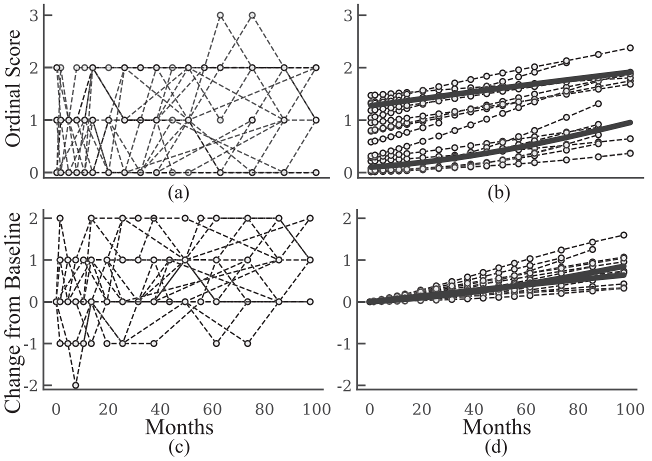

Fig. 9. Time series of left-arm rigidity scores for 20 randomly selected patients. (a) Ordinal time series. (b) Estimated means, $\hat{m}_i$ (gray dashed lines), and estimated trajectories of the prior means, $\hat{\mu}_k$ (black lines). (c) Change from baseline using raw scores. (d) Change from baseline using denoised scores.

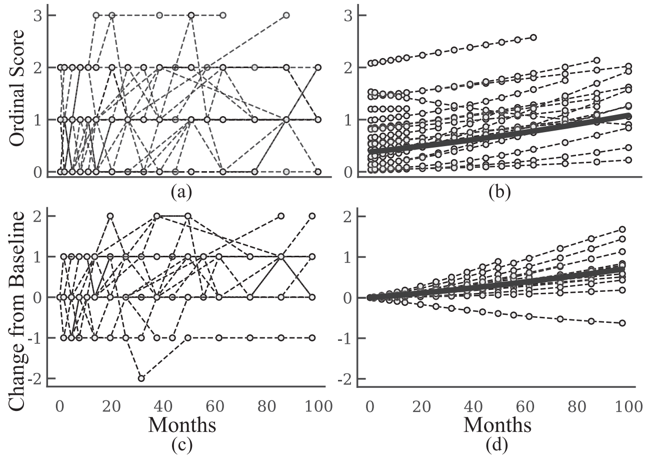

Fig. 10. Time series of right-leg rest tremor amplitude scores for 20 randomly selected patients. (a) Ordinal time series. (b) Estimated means, $\hat{m}_i$ (gray dashed lines), and estimated trajectories of the prior means, $\hat{\mu}_k$ (black lines). (c) Change from baseline using raw scores. (d) Change from baseline using denoised scores.

The effect of denoising is even more striking when we analyze the changes in the score from baseline. For each subject, the change from baseline at time t is the score at time t minus the score at baseline. For a steadily progressing disease, such as PD, the change from baseline should be positive for all t. Fig. 9(c) shows the change from baseline for raw left-arm rigidity scores for the subjects whose scores are plotted in Fig. 9(a). Note that the noise in the scores overwhelms the perception of any pattern.

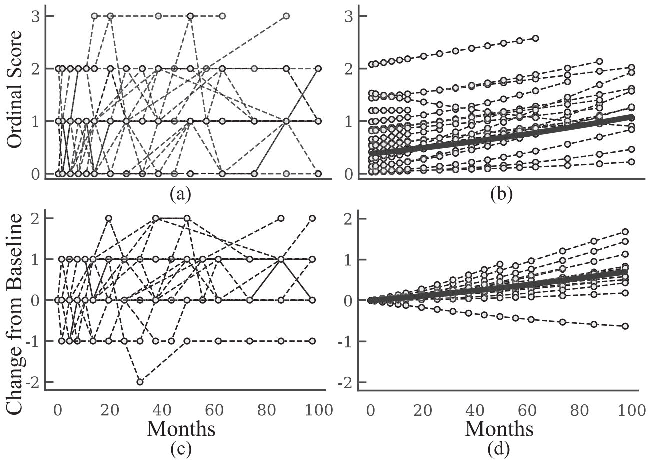

Fig. 11. Time series of right-leg agility scores for 20 randomly selected patients. (a) Ordinal time series. (b) Estimated means, $\hat{m}_i$ (gray dashed lines), and estimated trajectories of the prior means, $\hat{\mu}_k$ (black lines). (c) Change from baseline using raw scores. (d) Change from baseline using denoised scores.

TABLE IV
RATIO OF THE MEAN TO THE STANDARD DEVIATION OF THE CHANGE-FROM-BASELINE SCORES ACROSS THE POPULATION

Mean/Standard Dev. of change from baseline	Year 1	Year 2	Year 3	Year 4	Year 5
Left-arm Rigidity:					
Raw Ordinal Score	0.31	0.24	0.31	0.53	0.50
Denoised Score	0.88	0.89	0.98	0.99	0.98
Right-leg Rest Tremor Amplitude:					
Raw Ordinal Score	0.05	0.11	0.16	0.19	0.14
Denoised Score	0.25	0.28	0.31	0.35	0.41
Right-leg Agility:					
Raw Ordinal Score	0.11	0.33	0.39	0.52	0.51
Denoised Score	0.87	0.97	0.92	0.91	0.95

Fig. 9(d) shows the change from baseline for denoised scores. The progression of the disease is quite clear in Fig. 9(d).

Fig. 9(c) and (d) suggest that the denoised change-from-baseline scores are more useful to understand the underlying progression patterns than the raw change-from-baseline scores. This suggestion can be quantitatively evaluated as follows: Suppose we calculate the ratio of the mean to the standard deviation (across all subjects) of the change-from-baseline scores. If this ratio is higher, it is easier to quantitatively detect disease progression. The top two rows of Table IV show the change-from-baseline mean to standard deviation ratio for raw and denoised left-arm rigidity scores. Note the significantly larger ratios for denoised scores, clearly pointing to the utility of denoising.

The results for right-leg rest tremor amplitude and right-leg agility are similar. Fig. 10(a) and (b) show the raw and denoised scores for right-leg rest tremor amplitude for 20 randomly selected subjects. Again, progression patterns are difficult to discern in the raw scores. Denoising reveals that most patients' trajectories are relatively flat over time, while there is a small group for whom tremor typically begins a little higher and rises more rapidly. This is also reinforced by the trajectories of the prior component means (subtypes), $\hat{\mu}_k, \tilde{}k = 1, 2$. The raw and denoised change from baseline [Fig. 10(c) and (d)] behave similar to that of left-arm rigidity. The raw change from baseline does not indicate any pattern, while the denoised change from baseline shows mild progression. The mean to standard deviation ratios of change from baseline in Table IV (middle two rows) show quantitative improvement due to denoising.

Right-leg agility results (raw scores, denoised scores, raw change from baseline, denoised change from baseline) are shown in Fig. 11 for 20 randomly selected subjects. The denoising of the time series shown in Fig. 11(b) suggests that for most patients, this symptom gradually increases with time following the trajectory of the single prior mean, $\hat{\mu}_1$. Variation from this prior seems evenly distributed above and below the mean trajectory. Gradual progression with time is also evident in the denoised change-from-baseline in Fig. 11(d). The last two rows of Table IV show the improved mean to standard deviation ratio due to denoising.

VI. Discussion and Conclusion

The experimental results with simulated data clearly show that the hierarchical model is able to closely estimate the ground truth signal and does not overfit data. The number of clusters in the hierarchical model was successfully determined through cross-validation and matches the number of underlying clusters in the simulated data. The denoised mean provided by the model was similar to the ground truth. The mixture prior means, $\hat{\mu}_1, \hat{\mu}_2$, provide a good indication of a prototypical trajectory from each progression subtype (map).

The experimental results also show that the EM algorithm converges reliably in spite of the multi-level hierarchy for both the simulated data and the MDS-UPDRS data. For the MDS-UPDRS data, the hierarchical model also does not overfit, providing similar results on both training and test examples.

Analysis of the MDS-UPDRS data clearly shows the power of the method to elucidate underlying patterns in noisy time series of real-world clinical scores. The patterns exposed by the denoised series are more informative than those in the raw series. This is significant given the large number of studies which gather such data.

Identification of clinical subtypes and disease progression patterns in neuropsychiatric and neurodegenerative disorders is important for research and clinical trials. Different clinical subtypes may have different underlying pathological mechanisms and may respond to therapeutic interventions differently. Our methodology identifies disease progression subtypes from noisy real-world clinical data and is a promising tool for use in clinical studies.

Our model has some limitations which we discuss now: First, the model requires the entire ordinal time series before denoising is attempted. Thus, in its current form, the model cannot be used in an on-line fashion to update the estimate of the denoised series as the data arrive. However, it is not too difficult to modify the model to create such an on-line algorithm. Second, the mixture prior has some limitations. If the θ_i parameters have a very wide spread, then some other form of a prior with a wider support may be necessary. While processing the simulated and MDS-UPDRS data, we have not noticed the need for wider-support priors. However, if the need arises, the Gaussian components may be replaced with heavier-tailed t-distributed or other wide-support components. Finally, we note that there may be reason to use methods other than cross-validation to determine K, the number of components in the prior. For example, in some applications, it may be useful to just use $K = 1$ if there is reason not to cluster the θ_i's.

In conclusion, we have reported a method for denoising noisy ordinal time series of clinical scores which are common not only in PD but also in studies of other neurological diseases such as multiple sclerosis [31], [32], and in psychiatric conditions such as depression [33], [34] and anxiety [35], [36]. The method explicitly takes the ordinal nature of time series into account and draws statistical power from similarly progressing time series to overcome the short duration of the series. When applied to real-world clinical data, the method clarifies the underlying progression patterns.

Acknowledgment

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database.[1] For up-to-date information on the study, visit.[2] PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners. The full list of funding partners can be found at.[3]

References

[1] P. Martinez-Martin *et al.*, "Validation study of the Hoehn and Yahr scale included in the MDS-UPDRS," *Movement Disord.*, vol. 33, no. 4, pp. 651–652, 2018.

[2] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[3] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc. Ser. B. (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[4] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.

[5] P. McCullagh and J. A. Nelder, *Generalized Linear Models.* London, U.K.: Routledge, 1983.

[6] A. Agresti, *Analysis of Ordinal Categorical Data*. vol. 656. New York, NY, USA: Wiley, 2010.

[7] H. Wu, H. Lu, and S. Ma, "A practical SVM-based algorithm for ordinal regression in image retrieval," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 612–621.

[8] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[9] S. B. Kotsiantis and P. E. Pintelas, "A cost sensitive technique for ordinal classification problems," in *Methods and Applications of Artificial Intelligence*, (Lecture Notes in Computer Science), G. A. Vouros, and T. Panayiotopoulos, Eds., Berlin, Germany: Springer, 2004, pp. 220–229.

[10] E. Frank and M. Hall, *A Simple Approach to Ordinal Classification*. vol. 2167. (Lecture Notes in Computer Science), Berlin, Germany: Springer, 2001, pp. 145–156.

[11] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1279–1284.

[12] Y. Liu, A. W. K. Kong, and C. K. Goh, "A constrained deep neural network for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 831–839.

[13] P. McCullagh, "Regression models for ordinal data," *J. Roy. Stat. Soc. Ser. B. (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[14] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Adv. Large Margin Classifiers*, vol. 88, no. 2, pp. 115–132, 2000.

[15] W. Chu and S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, pp. 792–815, Apr. 2007.

[16] M. J. Mathieson, "Ordinal models for neural networks," in *Proc. 3rd Int. Conf. Neural Netw. Capital Markets*, 1996, pp. 523–536.

[17] F. Fernández-Navarro, A. Riccardi, and S. Carloni, "Ordinal neural networks without iterative tuning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2075–2085, Nov. 2014.

[18] R. Brant, "Assessing proportionality in the proportional odds model for ordinal logistic regression," *Biometrics*, vol. 46, no. 4, pp. 1171–1178, 1990.

[19] B. Peterson and F. E. Harrell, "Partial proportional odds models for ordinal response variables," *J. Roy. Stat. Soc. Ser. C. (Appl. Statist.)*, vol. 39, pp. 205–217, 1990.

[20] C. Cox, "Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach," *Statist. Med.*, vol. 14, no. 11, pp. 1191–1203, 1995.

[21] R. Williams, "Generalized ordered logit/partial proportional odds models for ordinal dependent variables," *Stata J.*, vol. 6, pp. 58–82, 2006.

[22] M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression and mortality," *Neurology*, vol. 17, no. 5, pp. 427–442, May 1967.

[23] C. G. Goetz *et al.*, "Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disord.*, vol. 23, no. 15, pp. 2129–2170, 2008.

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

[25] J. Jankovic *et al.*, "Variable expression of Parkinson's disease: A baseline analysis of the DAT ATOP cohort," *Neurology*, vol. 40, no. 10, pp. 1529–1529, 1990.

[26] S. M. Van Rooden, W. J. Heiser, J. N. Kok, D. Verbaan, J. J. Van Hilten, and J. Marinus, "The identification of Parkinson's disease subtypes using cluster analysis: A systematic review," *Movement Disord.*, vol. 25, no. 8, pp. 969–978, 2010.

[27] K. Marek, D. Jennings, S. Lasch, and A. E. A. Siderowf, "The Parkinson progression marker initiative (PPMI)," *Prog. Neurobiol.*, vol. 95, no. 4, pp. 629–635, 2011.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[29] G. G. Lorentz, *Bernstein Polynomials*. Providence, RI, USA: American Mathematical Soc., 2013.

[30] G. A. Seber, *Multivariate Observations*. vol. 252. New York, NY, USA: Wiley, 2009.

[31] B. Vickrey, R. Hays, R. Harooni, L. Myers, and G. Ellison, "A health-related quality of life measure for multiple sclerosis," *Qual. Life Res.*, vol. 4, no. 3, pp. 187–206, 1995.

[32] J. F. Kurtzke, "Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS)," *Neurology*, vol. 33, no. 11, pp. 1444–1444, 1983.

[33] M. Hamilton, "A rating scale for depression," *J. Neurol., Neurosurgery, Psychiatry*, vol. 23, pp. 56–62, Feb. 1960.

[34] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *Brit. J. Psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.

[35] M. Hamilton, "The assessment of anxiety states by rating," *Brit. J. Med. Psychol.*, vol. 32, pp. 50–55, 1959.

[36] A. T. Beck, N. Epstein, G. Brown, and R. A. Steer, "An inventory for measuring clinical anxiety: Psychometric properties," *J. Consulting Clin. Psychol.*, vol. 56, no. 6, pp. 893–897, 1988.

[1] [Online]. Available: www.ppmi-info.org/data

[2] [Online]. Available: www.ppmi-info.org

[3] [Online]. Available: https://www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors/