Hierarchical MAP Denoising of Longitudinal Hamilton Depression Rating Scores

Jonathan Koss
Dept. of Electrical Engineering
Yale University
New Haven, CT
jonathan.koss@yale.edu

Christine DeLorenzo
Dept. of Psychiatry
Dept. of Biomedical Engineering
Stony Brook University
Stony Brook, NY
christine.delorenzo@stonybrookmedicine.edu

Hemant D. Tagare
Dept. of Radiology and Biomedical Imaging
Dept. of Electrical Engineering
Yale University
New Haven, CT
hemant.tagare@yale.edu

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669362

***Abstract*—The Hamilton Depression Rating Scale provides ordinal ratings for evaluating different aspects of depression. These ratings are usually quite noisy, and longitudinal patterns in the ratings can be difficult to discern. This paper proposes a hierarchical maximum-a-posteriori (MAP) method for denoising the ordinal time series of such ratings. Real-world data from a clinical trial are analyzed using the model. Denoising reveals subject-specific longitudinal patterns, predicts future ratings, and reveals progression patterns via principal component analysis.**

***Index Terms*—Time Series, Ordinal Regression, Hierarchical Modeling, Hamilton Depression Rating Scale.**

I. Introduction

The Hamilton Depression Rating Scale (HDRS-17) [1] is commonly used to assess depression in clinical studies. The HDRS-17 contains 17 questions that physicians use to assess depressed mood, agitation, sleep, and other symptoms. Subjects are evaluated for every question using ordinal scores $0, 1, 2, 3, 4$, with 0 being absence of a symptom, and 4 being extreme severity of the symptom.

All HDRS-17 scores are usually summed into a total score representing the severity of depression. In clinical trials, HDRS-17 is administered at every visit, and the longitudinal time series of the total score is used to assess the effect of a drug or placebo on the enrolled subjects.

There are three problems with using HDRS-17 total scores as described above. First, the total-score time series are quite noisy (see Fig. 4). Second, subject-specific time series have few data points because subjects can only be evaluated so often. Fitting a statistical model to these time series is difficult. Third, because the total score is a sum of all ratings in the HDRS-17, it hides the more nuanced progression of different aspects of depression (mood, work, sleep etc.).

This paper seeks to develop a method to overcome the above problems and to reveal underlying longitudinal progression patterns in HDRS-17. We propose a hierarchical ordinal regression model to denoise the subject-specific time series of each question in the HDRS-17. It is developed in a classical maximum-a-posteriori (MAP) framework. The model explicitly takes into account the ordinal nature of the time series, and pools information from all subjects in the study to overcome the time series' short duration. The resulting subject-specific denoised time series clearly reveals the progression of all subjects in the study.

The real-world data we use come from a Stony Brook University study [2] seeking to relate Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) images of the subjects' brains to clinical response, as described in Section II. We use our model to denoise the longitudinal HDRS-17 ratings in the study. The results are reported in Section IV.

In [3], we proposed a more complex model for denoising ordinal time series and applied it to clinical scores of Parkinson's disease. The model used in this paper is a simplified version of that model.

The rest of the paper is organized as follows. Section II contains a brief literature review. The simplified model is presented in Section III. The results of using this model are presented in Section IV. Section V concludes the paper.

II. Literature Review and Background

A. HDRS-17

The HDRS-17, proposed in [1], is one of the standard research assessment tools for depression. Of its 17 questions, 10 questions have ordinal scores from 0 to 4, and 7 have ordinal scores from 0 to 2. HDRS-17 total scores have been used in many clinical trials, e.g. [4], [5].

There has been some effort to extend the HDRS-17 analysis beyond just using the total score by using linear models [6]–[8]. These do not account for the ordinal nature of the data.

B. The Stony Brook Study

The Stony Brook study was approved by the Institutional Review Board of Stony Brook University. All participants were recruited as a community sample and provided informed consent [2]. Briefly, the goals of the study were to determine: (1) whether antidepressant treatment response at eight weeks could be predicted by pretreatment imaging using FDG-PET images of the subjects' brains, and (2) whether changes in FDG-PET with treatment correlated with clinical response. The study was double-blinded and had a placebo and a

This research was supported by the grant R01NS107328 from the NIH (NINDS) and the grant R01MH104512 from the NIH (NIMH).

treatment (Selective Serotonin Reuptake Inhibitor, SSRI) arm. HDRS-17 ratings were collected every week for the 8 week study except weeks 5 and 7. A total of 85 subjects were enrolled in the study split between the two arms. Three subjects from the SSRI branch and five from the placebo branch were excluded due to excessive missing data.

III. Model

We now turn to explaining the model used to denoise HDRS-17 ordinal time series. The explanation is guided by our previous manuscript [3].

A. Ordinal Time series

Suppose y is an ordinal random variable which takes values in $\{0, \cdots, C-1\}$. Then, the probability distribution of y can be modeled as [9], [10] (see Fig. 1)

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

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

$$\gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \cdots \leq \gamma_{C-1}. \tag{2}$$

The constraints of equation (2) guarantee that the probabilities in equation (1) are non-negative and sum to 1.

The constraints of equation (2) can be implicitly imposed by a change of variables from $\gamma = (\gamma_1, \cdots, \gamma_{C-1})$ to $u = (u_1, \cdots, u_{C-1})$ as follows:

$$\begin{aligned} \gamma_1 &= u_1 \\ \gamma_k &= u_1 + \sum_{i=2}^{k} \log(1 + e^{u_i}), \text{ for } k = 2, \cdots, C-1. \end{aligned} \tag{3}$$

Since $\log(1 + e^x) > 0$ for all x, and is a bijection from $\mathbb{R}$ to $\mathbb{R}_{++}$, the γ's of equation (3) automatically satisfy equation (2) for any $u \in \mathbb{R}^{C-1}$. Using this change of variables, we may write equation (1) as $p(y \mid \Gamma(u))$ with no constraints on u.

Next, consider an ordinal time series $Y = (y_1, \cdots, y_T)$ with $y_t \in \{0, \cdots, C-1\}$ for $t = 1, \cdots, T \in [0, 1]$. Then, a model for the series is

$$p(Y \mid u, v) = \prod_{t=1}^{T} p(y_t \mid \Gamma(u)(1-t) + \Gamma(v)t), \tag{4}$$

where $u, v \in \mathbb{R}^{C-1}$ are the parameters at $t = 0$ and $t = 1$. This model assumes a linear change in the γ's (equation (1)) over time. The model can be made more complex by making the change non-linear, but for limited-duration clinical studies, a linear change is sufficient. We set $\theta = (u, v)$ and write $p(Y \mid u, v)$ of equation (4) as $p(Y \mid \theta)$.

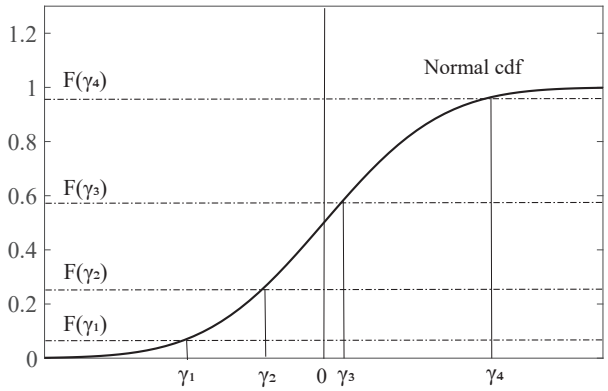

Fig. 1. A model for a five-class ordinal random variable. Class probabilities are $F(\gamma_1), F(\gamma_2) - F(\gamma_1), F(\gamma_3) - F(\gamma_2), F(\gamma_4) - F(\gamma_3)$, and $1 - F(\gamma_4)$.

Next, suppose there are N subjects, with time series $Y_i = (y_{i,1}, \cdots, y_{i,T_i})$, $i = 1, \cdots, N$. Assuming that the ith subject has the subject's own parameter $\theta_i = (u_i, v_i)$, the probability of Y_i is $p(Y_i \mid \theta_i)$, and that the time series of different subjects are independent, the probability of all time series is

$$p(\{Y_i\} \mid \{\theta_i\}) = \prod_{i=1}^{N} p(Y_i \mid \theta_i). \tag{5}$$

Recall that time series in clinical studies are short—often with the number of samples approaching the dimension of θ_i. Thus θ_i cannot be estimated reliably from y_i using equation (5).

B. Prior and Hyper-prior

Hierarchical modeling is one approach to overcoming the problems with short time series. The idea is to allow the estimate of any θ_i to draw statistical power from all other θ_i's, by using a normal prior for θ_i (mean μ and an isotropic variance-covariance matrix $\rho_\mu I$), i.e.

$$p(\theta_i \mid \mu, \rho_\mu) = \mathcal{N}(\mu, \rho_\mu I). \tag{6}$$

This model allows the θ_i's to be coupled via their mean μ. A large variance imposes the prior very weakly; conversely, a small variance imposes the prior more strongly.

We also assume a uniform hyper-prior on μ, and an inverse-gamma prior on ρ_μ (the usual conjugate prior on the variance of a normal distribution). Thus,

$$\begin{aligned} p(\mu) &= \text{const.}, \\ p(\rho_\mu \mid a, b) &= IG(a, b) = \frac{b^a}{\Gamma(a)} \frac{1}{(\rho_\mu)^{(a+1)}} e^{\frac{-b}{\rho_\mu}}. \end{aligned} \tag{7}$$

The parameter a is set to a fixed value, while the parameter b is found by cross-validation, as described later.

C. The full model

Grouping together all parameters to be estimated as $\phi = (\{\theta_i\}, \mu, \rho_\mu)$, we get the posterior likelihood

$$\begin{aligned} &p(\phi \mid \{Y_i\}) \propto p(\{Y_i\} \mid \phi) p(\phi) \\ &= \prod_{i=1}^{N} p(Y_i|\theta_i) p(\theta_i|\mu, \rho_\mu) \times p(\mu) \times p(\rho_\mu|\alpha, \beta). \end{aligned} \tag{8}$$

TABLE I
MAP-ITERATIONS. SUPERSCRIPT $[n+1]$, $[n]$ INDICATE ITERATION NUMBER. N = NUMBER OF SUBJECTS. $D = 2C-2$, WHERE C IS THE NUMBER OF ORDINAL CATEGORIES.

Parameter	Update
θ_i	$\theta_i^{[n+1]} = \arg\max_{\theta_i} \log p(Y_i \mid \theta_i) + \log p(\theta_i \mid \mu^{[n]}, \rho_\mu^{[n]})$
μ	$\mu^{[n+1]} = \frac{\sum_{i=1}^{N} \theta_i^{[n]}}{N}$
ρ_μ	$\rho_\mu^{[n+1]} = \sum_{i=1}^{N} \frac{\frac{\|\theta_i^{[n]} - \mu_k^{[n]}\|^2}{2} + b}{\frac{DN}{2} + a + 1}$

D. MAP Estimate

A maximum-a-posteriori (MAP) estimate of ϕ is obtained by maximizing the log of the posterior likelihood:

$$\log p(\phi \mid \{Y_i\}) = \sum_{i=1}^{N} (\log p(Y_i|\theta_i) + \log p(\theta_i|\mu, \rho_\mu)) + \log p(\rho_\mu \mid a, b), \quad (9)$$

$$\hat{\phi} = \arg\max_{\phi} \log p(\phi \mid \{Y_i\}). \quad (10)$$

We have dropped the logarithm of the prior, $\log p(\mu)$, in the above equations since it is a constant. The maximization is carried out iteratively. Within each iteration $\{\theta_i\}$, μ, and ρ_μ are optimized one at a time. The iterative updates are shown in Table I. The θ_i's are updated numerically using Adam [11]. The updates of μ and ρ_μ have closed form solutions. The iterations are carried out till convergence.

E. Denoising

The MAP estimate $\hat{\theta}_i = (\hat{u}_i, \hat{v}_i)$ gives the estimated mean and variance for the ith subject's time series as:

$$\hat{m}_i(t) = \sum_{y_t=0}^{C-1} y_t \times p(y_t|\Gamma(\hat{u}_i)(1-t) + \Gamma(\hat{v}_i)t) \quad (11)$$

$$\hat{s}_i^2(t) = \sum_{y_t=0}^{C-1} (y_t - \hat{m}_i(t))^2 \times p(y_t|\Gamma(\hat{u}_i)(1-t) + \Gamma(\hat{v}_i)t).$$

The estimated mean is the denoised time series.

F. Determining Hyper-prior Parameters

There are two parameters in the hyper-prior: a and b. Numerical experimentation shows that parameter estimates and the estimated denoised time series are not sensitive to the value of a. We simply set $a = 2$. The parameter b is found by cross validation as follows:

Given a data-set of ordinal time series for multiple subjects, 90% of the series are used as training data and 10% as test data. For each training series, 10% of the data points are set aside for cross-validation, and $10-$fold cross validation is used to determine b. The log-likelihood of the cross-validation data are assessed using equation (4) for the value of t corresponding to the cross-validation time points. The value b maximizing the validation log-likelihood is chosen.

G. From Denoising to Prediction

The model can also be used for extrapolation. In section IV we assess the ability of the model to make predictions. Specifically, we leave out the last time point from every time series (this is the data point to be predicted). We fit the model to all of the remaining data in the time series. Then, using the estimated $\hat{\theta}_i$ for each subject, we calculate the estimated mean $\hat{m}_i(t)$ where t is the time of the left-out data point.

H. Comments

Some salient features of the model are:

1) Equation (1) allows for "noise" in the data without assuming that the data belongs to a vector space.
2) The change of variables in equation (3) allows for priors whose support is all of $\mathbb{R}^{C-1}$, such as our normal prior.
3) The time series is not required to be evenly sampled in time, nor are all subjects required to have the same number of samples.
4) When the model is applied to a study, all time points in the study are scaled so that the start of the study is time 0 and the end of the study is time 1.
5) Ground truth is not available for real-world data to compare the denoised time series with. Because of this, in [3] we carried out detailed simulations to show that the denoised time series does estimate the simulation ground truth closely, even when the data are noisy.

IV. RESULTS

We have four goals in applying the model to the Stony Brook University study: 1) to understand the behavior of the raw data and the denoised data, 2) to understand the difference between the placebo and SSRI arms, 3) to predict time series using the strategy in section III-G, and 4) to understand the coupled progression of all HDRS-17 questions using PCA.

Before presenting the results, we fix some terminology.

A. Terminology

Recall that the HDRS-17 has 17 questions. Our model can be fit to the ordinal time series of any one of these questions. Slightly extending our notation, let $Y_{i,j} = \{y_{i,j,1}, y_{i,j,2}, \cdots, y_{i,j,T}\}$ denote the time series for the jth HDRS-17 question for subject i. Then, the *total score* time series of subject i is

$$\sum_{j=1}^{17} Y_{i,j} = \{\sum_{j=1}^{17} y_{i,j,1}, \sum_{j=1}^{17} y_{i,j,2}, \cdots, \sum_{j=1}^{17} y_{i,j,T}\}.$$

The *mean total score* time series over all subjects is

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{17} Y_{i,j} = \{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{17} y_{i,j,1}, \cdots, \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{17} y_{i,j,T}\}.$$

By analogy, we can also define a standard deviation of total score time series (formula not shown).

Turning to the denoised version, for a fixed j (fixed HDRS-17 question) we use all observed time series $Y_{i,j}$, $i = 1, \cdots, N$ of N subjects to estimate $\hat{\theta}_{i,j}$ from which the

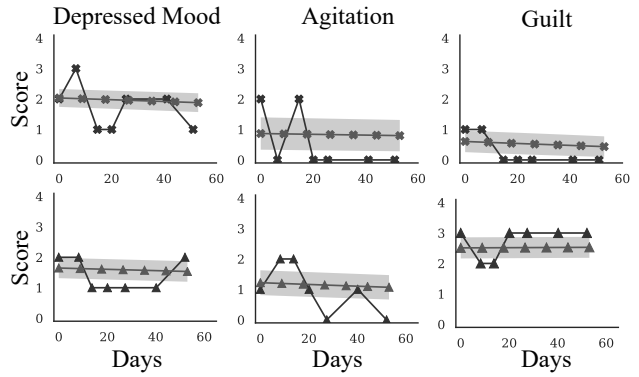

Fig. 2. Ordinal time series for three HDRS-17 scores. Top row: SSRI. Bottom row: Placebo. Solid lines are raw data. Gray lines are denoised data. Gray regions are standard deviations of denoised data.

denoised time series $\hat{m}_{i,j}(t)$ is calculated. As above, the *denoised total score* time series of subject i is $\sum_{j=1}^{17} \hat{m}_{i,j}(t)$, and the study-wide *mean denoised total score* time series is $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{17} Y_{i,j}$. Similarly, we can define a standard deviation denoised total score series.

B. Raw and Denoised Data

Fig. 2 shows examples of ordinal time series for Depressed Mood, Guilt, and Agitation questions from the Stony Brook University study.

The model from Section III was fit independently to the time series of each question for the placebo and SSRI arms separately. For each model fit, we used the cross-validation procedure described in section III-F to find the optimal b parameter. Then, the iterative MAP algorithm was used to denoise each subject-specific time series.

Fig. 2 shows the subject-specific denoised time series in gray along with the estimated standard deviation of the denoised estimate. The figure shows the ability of the model to deal with ordinal "noise", and the utility of the hierarchical structure to give good estimates despite the short length of the time series.

C. SSRI vs. Placebo

Next, we analyzed whether the subjects in the SSRI and placebo arms behaved similarly. Fig. 3 shows violin plots of total scores (dark gray) and denoised total scores (light gray) for placebo and SSRI arms for weeks 0, 4, and 8. The distribution of total scores for the placebo and SSRI arms are similar, as are the distributions of denoised total scores, suggesting similar progression in the two arms.

Table II provides further evidence, showing the mean and standard deviations of total scores and denoised total scores for placebo and SSRI arms for all weeks. Note that the placebo and SSRI means for all weeks are within one standard deviation of each other. Taken together with Fig. 3, this suggests that the SSRI and placebo arms of the study behave similarly. This fact has been observed in other studies as well, e.g. [12], [13]. Given this, from now on, we pool all subjects into a single group.

TABLE II
POPULATION MEANS AND STANDARD DEVIATIONS OF PLACEBO AND SSRI GROUPS FOR BOTH TOTAL SCORE AND DENOISED TOTAL SCORE.

Week	0	1	2	3	4	6	8
Total Score							
Placebo Mean (StDev)	16.8 (3.7)	14.4 (3.3)	14.4 (4.7)	12.9 (5.2)	12.5 (6.1)	10.4 (5.8)	9.7 (5.5)
SSRI Mean (StDev)	18.3 (5.7)	15.3 (5.4)	14.8 (5.9)	13.4 (6.5)	11.8 (6.7)	12.4 (6.7)	11.2 (6.7)
Total Denoised Score							
Placebo Mean (StDev)	15.5 (2.8)	15.0 (2.8)	14.6 (2.8)	14.2 (2.8)	13.5 (2.7)	12.7 (2.8)	12.3 (2.8)
SSRI Mean (StDev)	16.3 (4.1)	15.8 (4.2)	15.1 (4.2)	15.0 (4.1)	14.2 (4.0)	13.7 (4.1)	13.2 (4.1)

TABLE III
TOTAL SCORES AND DENOISED TOTAL SCORES FOR EACH WEEK AFTER POOLING PLACEBO AND SSRI SUBJECTS.

Week	0	1	2	3	4	6	8
Total Score							
Mean (StDev)	17.8 (5.0)	15.0 (4.5)	14.6 (5.5)	13.2 (5.9)	12.1 (6.3)	11.5 (6.5)	10.5 (6.2)
Denoised Total Score (8 Weeks)							
Mean (StDev)	16.0 (3.4)	15.2 (3.4)	14.4 (3.4)	13.9 (3.3)	13.0 (3.1)	11.4 (3.2)	9.7 (3.1)

D. Pooled Denoising

After pooling the subjects into a single group, we re-estimated the hyper-parameter b using cross validation, and used b to denoise the time series of every subject for each HDRS-17 question. Fig. 4a-b shows the total scores and denoised total scores for 38 randomly selected subjects. Note that the denoised total scores indicate progression far more clearly. This improvement is even more striking in Fig. 4c-d which show the change in the total score from baseline. The changes in the total score in Fig. 4c are obscured by noise, but are quite clear in the denoised series in Fig. 4d.

For a more detailed analysis, Fig. 5 shows a scatter plot of total denoised scores vs. total score for all subjects for all weeks. The correlation coefficient between the two is high, being 0.8.

E. Study-wide effect

Table III assesses the effect of denoising on study-wide statistics. The Table shows means and standard deviations for all subjects for each week. The first row shows total scores; the second row shows denoised total scores. Note that the denoised total score means are within one standard deviation of the total score means, showing that the effect of denoising on study-wide statistics is small.

F. Denoised Prediction

The high correlation between denoised total scores and total scores suggests that the denoised total scores can predict total scores. To investigate this, we followed the procedure in Section III-G: we fit the model to all data from week 0 to week 6, and used the model to predict the score for week 8.

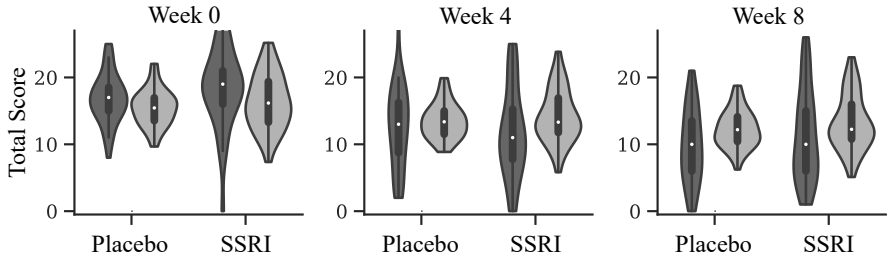

Fig. 3. Violin plots of the standard and denoised total scores for Week 0 till Week 8 for the total scores (dark gray) and the denoised total scores (light gray). The means of the placebo and SSRI groups were not found to be different for any of the weeks in the study when compared with a t-test. The resulting p-values were: 0.37 (Week 0), 0.27 (Week 4), and 0.23 (Week 8).

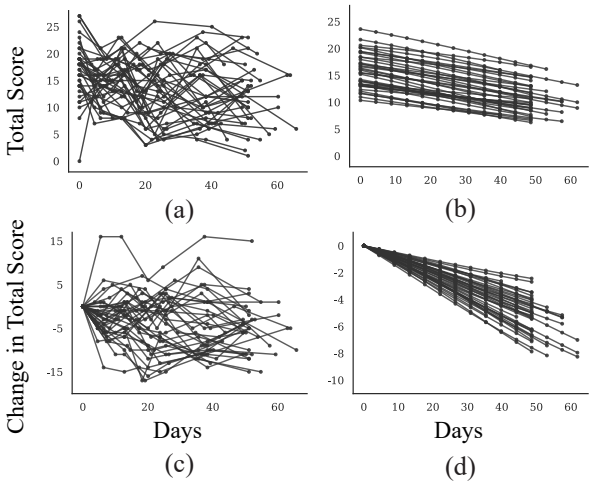

Fig. 4. Total scores for 38 randomly selected subjects. (a) Time series of total score. (b) Time series of denoised total scores. (c) Time series of change in total score (baseline subtracted). (d) Time series of change in denoised total score (baseline subtracted)

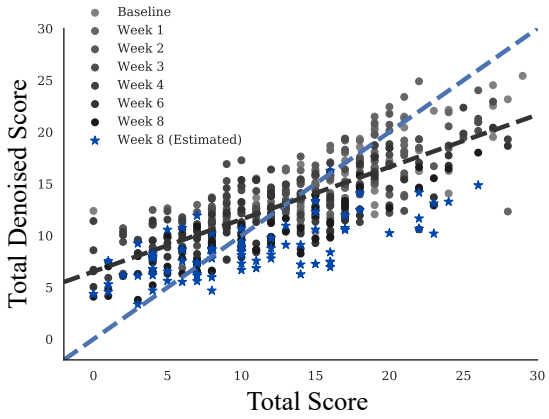

Fig. 5. Scatter plot of total score vs. denoised total score for every subject at every visit. Denoised scores learned from 8 weeks of data (black dots). Estimated denoised scores at week 8 learned from first 6 weeks of data (blue stars). A 45 degree line is shown in blue. The best linear fit is shown in black.

TABLE IV
CORRELATION COEFFICIENTS BETWEEN TOTAL SCORE AND TOTAL DENOISED SCORE FOR ALL SUBJECTS WEEK-BY-WEEK. DENOISED SCORES WERE OBTAINED USING ONLY 6 WEEKS OF DATA. DENOISED SCORES FOR WEEK 8 WERE PREDICTED.

Week	0	1	2	3	4	6	8(est.)
Correlations	0.75	0.73	0.81	0.81	0.81	0.76	0.70

Fig. 5 shows the scatter plot of the estimated week 8 denoised total scores vs. the measured total scores. Table IV shows the correlation coefficients between denoised total scores and total scores for all subjects in the study. The correlation coefficients for Week $0-6$ are in the range $0.73-0.81$. The correlation coefficient between the predicted Week 8 and the measured total scores is 0.70, only marginally outside the range of the coefficients from Weeks $0-6$. This demonstrates model's capacity for prediction.

G. Principal Component Analysis (PCA)

We next explored the relation between different HDRS-17 questions via PCA. We carried out PCA of the raw time series and of the denoised time series in the following way. For each subject, we took the scores of all questions for a visit as a 17-component vector. The vectors for all visits and subjects were gathered together in a data set and then PCA was carried out. Separate PCA's were performed for raw scores and for denoised scores. Fig. 7 shows the percentage variance explained by the principal components. Clearly, the first component is very significant.

Fig. 6a-b shows the PCA-loadings for the raw scores and the denoised scores for the first two principal components. The two sets of loadings are quite similar, again illustrating that study-wide statistics are not influenced by denoising. Fig. 6c-d shows the corresponding mean $\pm$ one standard deviation of the first two principal components. The first principal component corresponds to changes in Anxiety (Somatic, Psychic), Suicide, Guilt, General Somatic Symptoms, Insomnia, Work and Activities, and Depressed Mood (Figs. 6a and c). All loadings are positive. The second principal component (Figs. 6b and d) has significant loadings for Genital Symptoms and Insomnia (which are negative).

Fig. 6e-f shows the trajectories of individual subjects in the principal component space. The start of each trajectory is indicated by a black dot. The trajectories of raw data are not easily interpretable. On the other other hand, the trajectories of

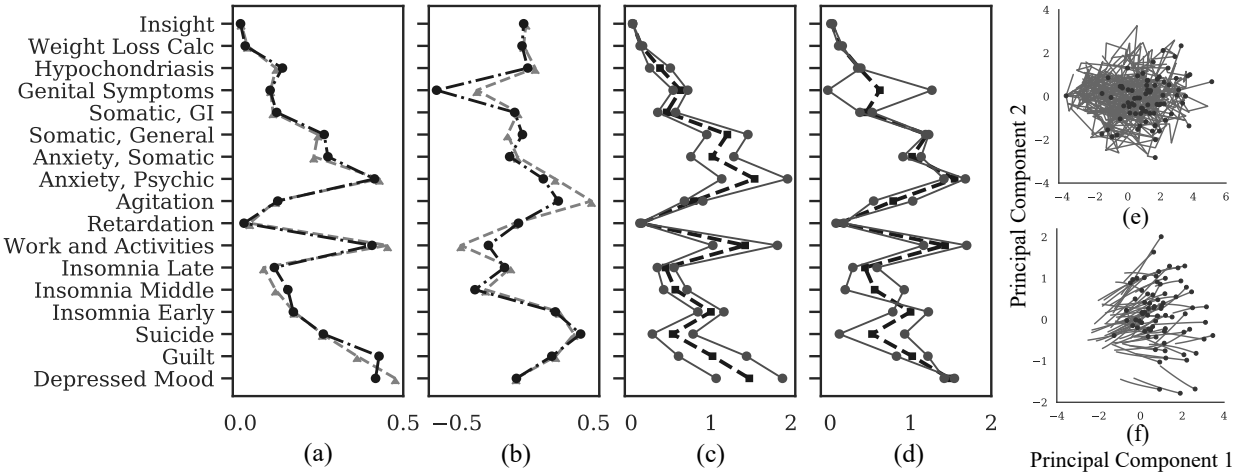

Fig. 6. (a) Principal component (PC) 1 calculated from raw ordinal scores (gray) and denoised data (black). (b) PC 2 calculated from raw ordinal scores (gray) and denoised data (black). (c) Mean denoised scores +/- PC 1. (d) Mean denoised scores +/- PC 2. (e, f) Subject-specific trajectories of HDRS-17 scores in Principal Component space. (e) Raw trajectories, (f) Denoised trajectories.

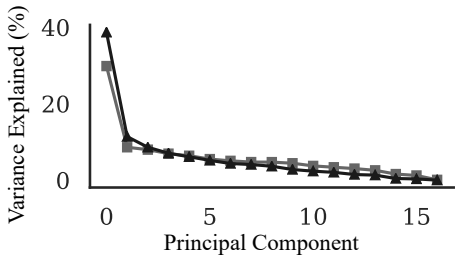

Fig. 7. Percent of variance explained by each principal component for raw ordinal data (grey) and denoised data (black).

denoised data provide useful information. The first component seems to decrease monotonically with time indicating that Depressed Mood, Suicide, Guilt etc. decrease as the study progresses. The second component converges towards a value slightly below 0. This suggests that any *differences* among the subjects in the second principal component scores decrease as the study progresses.

V. Conclusions

The hierarchical model proposed in this paper is effective in denoising HDRS-17 time series. This is clear in the results of Fig. 5 and in the principal component results of Fig. 6e-f. The denoised time series also has predictive capacity.

One limitation of the method is that it requires an entire time series before denoising can occur. We hope to develop an "on-line" method in the future.

Ordinal scores are also used to rate many other neurological diseases/disorders, e.g. Parkinson's disease [14]. Models, such as ours, can be quite useful in analyzing such studies.

Acknowledgements

This research was supported by the grant R01NS107328 (NINDS) and the grant R01MH104512 (NIMH).

References

[1] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, p. 56–62, Feb 1960.

[2] K. Hill, J. Gardus, E. Bartlett, G. Perlman, R. Parsey, and C. DeLorenzo, "Fully quantitative pretreatment brain metabolism does not predict depression response to escitalopram or placebo, a randomized trial," *Biological Psychiatry*, vol. 89, no. 9, pp. S357–S358, 2021.

[3] J. Koss, S. Tinaz, and H. D. Tagare, "Hierarchical denoising of ordinal time series of clinical scores," *IEEE Journal of Biomedical and Health Informatics*, submitted for publication.

[4] B. N. Gaynes, D. Warden, M. H. Trivedi, S. R. Wisniewski, M. Fava, and A. J. Rush, "What did star* d teach us? results from a large-scale, practical, clinical trial for patients with depression," *Psychiatric services*, vol. 60, no. 11, pp. 1439–1445, 2009.

[5] C. Goldberger, J. Guelfi, and D. Sheehan, "Assessment of anxiety in clinical trials with depressed patients using the Hamilton Depression Rating Scale," *Psychopharmacology bulletin*, vol. 44, no. 3, p. 34, 2011.

[6] E. Hamdi, Y. Amin, and M. Abou-Saleh, "Performance of the hamilton depression rating scale in depressed patients in the United Arab Emirates," *Acta Psychiatrica Scandinavica*, vol. 96, no. 6, pp. 416–423, 1997.

[7] M. S. Milak, R. V. Parsey, J. Keilp, M. A. Oquendo, K. M. Malone, and J. J. Mann, "Neuroanatomic correlates of psychopathologic components of major depressive disorder," *Archives of General Psychiatry*, vol. 62, no. 4, pp. 397–408, 2005.

[8] M. P. Broen, A. J. Moonen, M. L. Kuijf, K. Dujardin, L. Marsh, I. Richard, S. E. Starkstein, P. Martinez-Martin, and A. F. Leentjens, "Factor analysis of the hamilton depression rating scale in parkinson's disease," *Parkinsonism & related disorders*, vol. 21, no. 2, pp. 142–146, 2015.

[9] C. Cox, "Location—scale cumulative odds models for ordinal data: A generalized non-linear model approach," *Statistics in Medicine*, vol. 14, no. 11, p. 1191–1203, 1995.

[10] B. Peterson and F. E. Harrell, "Partial proportional odds models for ordinal response variables," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 39, p. 205–217, 1990.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980 [cs]*, Jan 2017. arXiv: 1412.6980.

[12] I. Kirsch, "Antidepressants and the placebo effect," *Zeitschrift für Psychologie*, 2015.

[13] B. T. Walsh, S. N. Seidman, R. Sysko, and M. Gould, "Placebo response in studies of major depression: variable, substantial, and growing," *Jama*, vol. 287, no. 14, pp. 1840–1847, 2002.

[14] C. G. Goetz, B. C. Tilley, and et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, p. 2129–2170, 2008.