



A Bayesian Disease Progression Model for Clinical Trajectories

Yingying Zhu^(✉) and Mert R. Sabuncu^(✉)

Schools of ECE and BME, Cornell University, Ithaca, USA
{yz2377,ms3375}@cornell.edu

Abstract. In this work, we consider the problem of predicting the course of a progressive disease, such as cancer or Alzheimer’s. Progressive diseases often start with mild symptoms that might precede a diagnosis, and each patient follows their own trajectory. Patient trajectories exhibit wild variability, which can be associated with many factors such as genotype, age, or sex. An additional layer of complexity is that, in real life, the amount and type of data available for each patient can differ significantly. For example, for one patient we might have no prior history, whereas for another patient we might have detailed clinical assessments obtained at multiple prior time-points. This paper presents a probabilistic model that can handle multiple modalities (including images and clinical assessments) and variable patient histories with irregular timings and missing entries, to predict clinical scores at future time-points. We use a sigmoidal function to model latent disease progression, which gives rise to clinical observations in our generative model. We implemented an approximate Bayesian inference strategy on the proposed model to estimate the parameters on data from a large population of subjects. Furthermore, the Bayesian framework enables the model to automatically fine-tune its predictions based on historical observations that might be available on the test subject. We applied our method to a longitudinal Alzheimer’s disease dataset with more than 3,000 subjects [1] with comparisons against several benchmarks.

1 Introduction

Many progressive disorders, such as Alzheimer’s disease (AD) [2], begin with mild symptoms that often precede diagnosis, and follow a patient-specific clinical trajectory that can be influenced by genetic and/or other factors. Therapeutic interventions, if available, are usually more effective in the earliest stages of a progressive disease. Therefore, tracking and predicting disease progression, particularly during the mild stages, is one of the primary objectives of personalized medicine.

In this paper, we are motivated by the real-world clinical setting where each individual is at risk and thus monitored for a specific progressive disease, such as AD. Furthermore, we assume that each individual might pay zero, one, or more

visits to the clinic. In each clinical visit, various biomarkers or assessments (correlated with the disease and/or its progression) are obtained. Example biomarker modalities include brain MRI scans, PET scans, blood tests, and cognitive test scores. The number and timing of the visits, and the exact types of data collected at each visit can be planned to be standardized, but often vary wildly between patients in practice. An ideal clinical prediction tool should be able to deal with this heterogeneity and compute accurate forecasts for arbitrary time horizons.

We present a probabilistic disease progression model that elegantly handles the aforementioned challenges of longitudinal clinical settings: data missingness, variable timing and number of visits, and multi-modal data (i.e., data of different types). The backbone of our model is a latent sigmoidal curve that captures the dynamics of the unobserved pathology, which is reflected in time-varying clinical assessments. Sigmoid curves are conceptually useful abstractions that fit well a wide range of dynamic physical and biological phenomena, including disease progression [3–5], which exhibit a floor and ceiling effect. In our framework, the sigmoid allows us to model the temporal correlation in longitudinal measurements and capture the dependence between the different tests and assessments, which are assumed to be generated conditionally independently from the latent state. We implemented an approximate Bayesian inference strategy on the proposed model and applied it to a large-scale longitudinal AD dataset [1].

In our experiments, we considered three target variables, which are widely used cognitive and clinical assessments associated with AD: the Mini Mental State Examination (MMSE) [6], the Alzheimer’s Disease Assessment Scale Cognitive Subscale (ADAS-COG) [7], and the Clinical Dementia Rating Sum of Boxes (CDR-SB) [8]. We trained and evaluated the proposed model on a longitudinal dataset with more than 3,000 subjects that included healthy controls (cognitively normal elderly individuals), subjects with mild cognitive impairment (MCI, a clinical stage that indicates high risk for dementia), and patients with AD. We provide a detailed analysis of prediction accuracy achieved with the proposed model and alternative benchmark methods under different scenarios that involve varying the past available visits and future time windows. In all our comparisons, the proposed model achieves significantly and substantially better accuracy for all target biomarkers.

2 Methods

2.1 Model

Let us first describe our notation and present our model. Assume we are given n subjects. $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$ denotes subject i ’s d -dimensional attribute vector. In our experiments, this vector contains APOE genotype (encoded as number of E4 alleles, which can be 0, 1 or 2) [9], education (in years) [10], sex (0 for female and 1 for male) [11] and two well-established neuroanatomical biomarkers of AD computed from a baseline MRI scan (namely total hippocampal [12] and ventricular volume [13] normalized by brain size). The MRI biomarkers capture so-called “brain reserve” [14]. Let $\mathbf{y}_i^k \in \mathbb{R}^{v_i \times 1}$ represent the values of the the

k 'th dynamic (i.e., time-varying) target variable at v_i different clinical visits. $\mathbf{t}_i = [t_{i1}, \dots, t_{iv_i}] \in \mathbb{R}^{v_i \times 1}$ denotes a vector of the age of subject i at these visits. The number and timing of the visits can vary across subjects. In general, we will assume $k \in \{1, \dots, m\}$. In our experiments, we consider 3 target variables: MMSE, ADAS-COG or CDRSB and thus $m = 3$. We use $\mathbf{d}_i^k = [d_{i1}^k, \dots, d_{iv_i}^k]$ to denote subject i 's latent trajectory values associated with the k 'th target variable. We assume each $d_{ij}^k \in [0, 1]$, with lower values corresponding to milder stages. As we describe below, the target variable, which is a clinical assessment, will be assumed to be a noisy observation of this latent variable. We model the latent trajectory of \mathbf{d}_i^k as a sigmoid function of time (i.e., age), parameterized by a target- and subject-specific inflection point $p_i^k \in \mathbb{R}$ and a subject-specific slope parameter $s_i \in \mathbb{R}$. We assume that the slopes of the latent sigmoids associated with each target are coupled for each subject, yet the inflection points differ, which correspond to an average lag between the dynamics of target variables. This is consistent with the hypothesized biomarker trajectories of AD [3]. However, it would be easy to relax this assumption by allowing each target variable to have its own slope.

We assume the inflection points $\{p_i^k\}$ and slopes $\{s_i\}$ are random variables drawn from Gaussian priors with means equal to linear functions of subject-specific attributes \mathbf{x}_i : $p_i^k \sim \mathcal{N}(\mathbf{v}^T \mathbf{x}_i + a_k, \sigma_p^2)$, $s_i \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_i + b, \sigma_s^2)$, where $a_k \in \mathbb{R}$ is associated with the k 'th target (accounting for different time lags between target dynamics), while $\mathbf{v}, \mathbf{w} \in \mathbb{R}^{d \times 1}$, and $b, \sigma_p, \sigma_s \in \mathbb{R}$ are general parameters. Here and henceforth $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian with mean μ and variance σ^2 . Given s_i and p_i^k , the latent value d_{ij}^k associated with the k 'th target is computed by evaluating the sigmoid at t_{ij} , $d_{ij}^k = \frac{1}{1 + \exp(-(t_{ij} - p_i^k) s_i)}$. The inflection point p_i^k marks the age at which the rate of change achieves its maximum, which is equal to $s_i/4$.

Finally, we assume that the target variable value y_{ij}^k is a linear function of the latent state d_{ij}^k corrupted by additive zero-mean independent Gaussian noise:

$$y_{ij}^k \sim \mathcal{N}(c_k d_{ij}^k + h_k, \sigma_k^2), \quad (1)$$

where c_k, h_k , and $\sigma_k \in \mathbb{R}$ are universal (not subject-specific) parameters associated with the k 'th target variable. We refer to Eq. (1) as an observation model.

2.2 Inference

In this section, we discuss how to train the proposed model and apply it during test time.

Training: Let us use Θ to denote the parameter set of our model:

$$\Theta = \{\mathbf{w}, b, \sigma_p, \sigma_s, \mathbf{v}, \{a_k, c_k, h_k, \sigma_k\}_{k=1, \dots, m}\}.$$

The goal of training is to estimate the model parameters Θ given data from n subjects: $\{\mathbf{y}_i, \mathbf{x}_i, \mathbf{t}_i\}_{i=1, \dots, n}$. Here, $\mathbf{y}_i = [\mathbf{y}_i^1 \dots \mathbf{y}_i^m] \in \mathbb{R}^{v_i \times m}$ denotes m target values of the i th subject for v_i visits.

We estimate Θ via maximizing the likelihood function:

$$\prod_{i=1}^n P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{t}_i; \Theta).$$

We use the standard notation of $p(y|x)$ to indicate the probability density function of the random variable Y (evaluated at y) conditioned on the random variable X taking on the value x . Also, parameters not treated as random variables are collected on the right hand side of “;”.

Now, let us focus on the likelihood of each subject:

$$P(\mathbf{y}_i | \mathbf{x}_i, \mathbf{t}_i; \Theta) = \int \int \left[\prod_{j=1}^{v_i} p(\mathbf{y}_{ij} | s_i, \mathbf{p}_i, \mathbf{t}_{ij}) \right] p(s_i, \mathbf{p}_i | \mathbf{x}_i; \Theta) ds_i d\mathbf{p}_i,$$

with $p(s_i, \mathbf{p}_i | \mathbf{x}_i; \Theta) T = p(s_i | \mathbf{x}_i; \Theta) p(\mathbf{p}_i | \mathbf{x}_i; \Theta) T$.

Instead of the computationally challenging Eq. (2), we use variational approximation [15] and maximize the expected lower bound objective (ELBO):

$$\begin{aligned} F(\Theta, \{\gamma_i\}) &= \sum_{i=1}^n \mathbb{E}_q \left(\sum_{j=1}^{v_i} \sum_{k=1}^m T \log p(y_{ij}^k | s_i, p_i^k, \mathbf{t}_{ij}; \Theta) \right) \\ &\quad - \mathbb{E}_q (\log q(s_i; \gamma_i)) - \mathbb{E}_q (\log q(\mathbf{p}_i; \gamma_i)), \end{aligned} \quad (2)$$

where $q(s_i; \gamma_i) = N(\mu_{si}, \sigma_{si}^2)$ and $q(\mathbf{p}_i; \gamma_i) = N(\mu_{pi}, \Sigma_{pi} = \Gamma_{pi}^T \Gamma_{pi})$ are proxy distributions that approximate the true posteriors $p(s_i | \mathbf{y}_i, \mathbf{x}_i; \Theta)$ and $p(\mathbf{p}_i | \mathbf{y}_i, \mathbf{x}_i; \Theta)$, respectively. During training, we use gradient-ascent to iteratively optimize Eq. 2 and solve for the optimal model parameters Θ^* and the optimal parameters of the proxy distributions $\{\gamma_i^*\}$. The expectation in the first term is with respect to the proxy distributions and can be approximated via Monte Carlo sampling:

$$\mathbb{E}_q \left(\sum_k \sum_j \log p(y_{ij}^k | s_i, p_i^k, \mathbf{t}_{ij}; \Theta) \right) \approx \frac{1}{S} \sum_j \sum_{s=1}^S \log p(\mathbf{y}_{ij} | s_i^{(s)}, \mathbf{p}_i^{(s)}, \mathbf{t}_{ij}; \Theta), \quad (3)$$

where $s_i^{(s)}$ and $\mathbf{p}_i^{(s)}$ are samples drawn using the “re-parameterization trick.” I.e., $s_i^{(s)} = \eta^{(s)} \sigma_{si} + \mu_{si}$ and $\mathbf{p}_i^{(s)} = \Gamma_{pi}^T \epsilon^{(s)} + \mu_{pi}$, where $\eta^{(s)} \in \mathbf{R}$ and $\epsilon^{(s)} \in \mathbf{R}^{m \times 1}$ are realizations of the auxiliary random variables, independently drawn from zero-mean standard Gaussians, $N(0, 1)$ and $N(\mathbf{0}, \mathbf{I})$, respectively. The “re-parameterization trick” allows us to differentiate the ELBO (or more accurately, its approximation that uses Eq. 3) with respect to γ_i .

E.g.:

$$\frac{\partial s_i^{(s)}}{\partial \sigma_{si}} = \eta^{(s)}, \text{ and } \frac{\partial s_i^{(s)}}{\partial \mu_{si}} = 1.$$

Testing. During test time, we are interested in computing the posterior distribution of \mathbf{y}_{n+1} for a new subject with \mathbf{x}_{n+1} at an arbitrary time-point (age) t .

We drop the second sub-script, i.e., j index, of \mathbf{y}_{n+1} to emphasize that we will be computing these posterior probabilities at many different (often future) time-points. There are two types of test subjects: those with no history of visits (scenario 1), and those with at least one prior clinical visit (scenario 2). For scenario 2, we will use $\{\mathbf{y}_{(n+1)j}, t_{(n+1)j}\}_{j=1, \dots, v_{n+1}}$ to collectively denote the v_{n+1} historical observations and their corresponding visit times. We fix Θ^* to the values obtained from training. In scenario 1, we use Eq. (eq:ELBO) to compute the posterior. In the second scenario, we will first maximize the ELBO of Eq. (2) with respect to γ_{n+1} and evaluated for the observations on the new subject $\{\mathbf{y}_{(n+1)j}, t_{(n+1)j}\}$ and attribute vector: \mathbf{x}_{n+1} . We then proceed to use these approximate q distributions in Eq. (2), replacing $p(s|\mathbf{x}_i; \Theta^*)$ and $p(p^k|x_i; \Theta^*)$, to evaluate the posterior distribution for an arbitrary time-point t conditioned on past observations.

3 Experiments

Dataset. We use a dataset of 3,057 subjects (baseline age 73.3 ± 17.2 years) collected by ADNI [1] to empirically validate and demonstrate the proposed model. This dataset contained multiple clinical visits per subject, during which thorough cognitive and symptomatic assessments were conducted. In our experiments, we used MMSE, ADAS-COG and CDR-SB as three target variables. MMSE has a range between 0 (impaired) and 30 (healthy), whereas ADAS-COG takes on values between 0 (healthy) to 70 (severe), and CDR-SB varies from 0 (healthy) to 18 (severe). The first two (MMSE and ADAS-COG) are general cognitive assessments that track and predict dementia, while CDR-SB is a clinical score that measures the severity of dementia-associated symptoms. In addition to the target variables, we utilized individual-level traits associated with AD: age, APOE genotype (number of E4 alleles), sex, and education (in years). We also used baseline brain MRI scans to derive two anatomical biomarkers of AD: total hippocampal and ventricle volume normalized by brain size. These imaging biomarkers were automatically computed with FreeSurfer [16] and quality controlled as previously described [17].

3.1 Experimental Setup

Benchmark Methods. In our experiments, we compare the proposed method to the following benchmarks:

1. **Global:** A 4-parameter (scale, bias, inflection, and slope) sigmoidal model that was fit on all training data (least-squares).
2. **Sex-specific:** Same as “Global” but separate for males and females.
3. **APOE-specific:** Same as “Global”, but separate for three groups defined by APOE-E4 allele count $\{0, 1, 2\}$.
4. **Sex- and APOE-specific:** Same as “Global”, but separate for each sex and APOE group.

5. **Linear mixed effects (LME) model:** A linear regression model with subject-specific attributes (\mathbf{x}_i) as fixed effects, and time and bias term as a random effects. This LME model, commonly used to capture longitudinal dynamics, allows each subject to deviate from the average trajectory determined by its attributes by shifts in slope and offset.
6. **Subject-specific linear model:** Least-squares fit of a linear model on each subject’s historical data. When there is only one past visit, we adopt a carry-forward extrapolation.

Implementation of Proposed Method. We coded in Python ¹, using the Edward library [18], which is in turn built on TensorFlow [19]. We used a 20-fold cross-validation strategy in all our experiments. We first partitioned the data into 20 non-overlapping, roughly equally-sized sets of subjects. In each of the 20 folds, we reserved one of the partitions as the independent test set. Out of the remaining 19 partitions, one was set aside as a validation set, while the rest were combined into a training set. The training set was used to estimate the model parameters, i.e., Θ^* , while performance on the validation set was used to select hyper-parameters, such as step size in the optimization and evaluate random initializations. Finally, test performance was computed on the test set. We report results averaged across 20 folds.

3.2 Results and Discussion

We first show quantitative prediction results for all methods and target variables (MMSE, ADAS-COG, and CDRSB). In the following, we consider several prediction scenarios. In the first scenario, we vary the number of past visits available on test subjects (i.e., v_{n+1}). In general, we expect this variation to influence the LME and subject-specific linear model benchmarks, in addition to the proposed model. These methods fine-tune their predictions based on historical observations available on test data. With more test observations, we expect them to achieve better accuracy. All other benchmarks are fixed after training and thus their performance should not improve with increasing number of past observations. In the second scenario, we fix the number of past observations on test subjects and vary the prediction horizon. In general, all models’ predictions should be less accurate for more distant future time-points.

Varying the Number of Past Visits. Figure 1 shows the MMSE, ADAS-COG and CDRSB prediction accuracies (mean and standard deviation of absolute error). We observe that the performance of the training-fixed benchmarks (1–4) worsen slightly as the number of past visits increases. This is likely because the training data contains more samples at early times (i.e., relatively younger ages), partially because most subjects drop out by their 4th visit. Therefore, a model trained on these data is expected to be less accurate for older ages.

The adaptive benchmarks (5–6) and the proposed model, on the other hand, overcome this handicap to achieve better accuracy with more past visits. As we

¹ The code of this work is available at <https://github.com/zyy123jy/kdd>.

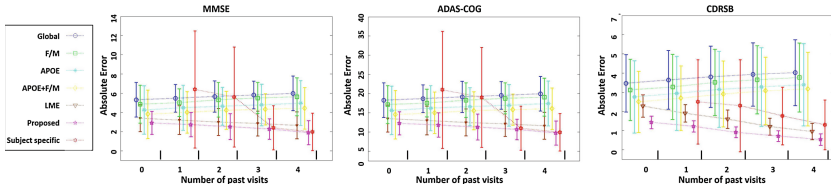


Fig. 1. Absolute error (mean and standard deviation) of all methods for predicting MMSE, ADAS-COG and CDRSB, as a function of number of past visits available on test subjects.

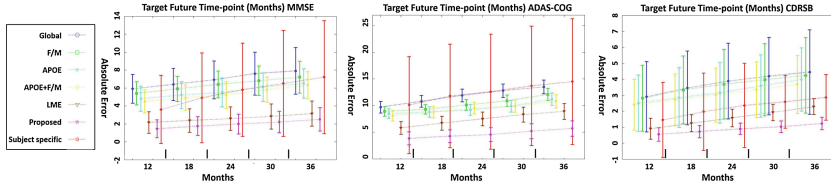


Fig. 2. Absolute error (mean and standard deviation) of all methods for predicting MMSE, ADAS-COG and CDRSB. We used two points from each test subject as past observations and varied the time horizon for prediction.

discussed above, this is largely because these techniques exploit test observations to fine-tune their models. The subject-level linear model (benchmark 6), in fact, is an extreme example, where the predictions are computed merely by extrapolating from historical observations without relying on training data.

Finally, the proposed model achieves a significantly and substantially better accuracy than all benchmarks (all paired permutation p-values $< p_{max} = 0.04$). The subject-specific benchmark (6) exhibits the largest variance implying the quality of performance varies wildly across subjects. Overall, the training-fixed benchmarks perform the worst. In general the proposed model’s variance is among the smallest, indicating consistency in prediction accuracy.

Varying the Time Horizon. In order to evaluate how prediction performance changes as a function of the time horizon, we evaluated the methods for different future time-points. In this empirical scenario, we assume that each test subject has 2 past clinical assessments (obtained at baseline and month 6). Our goal is to predict MMSE, ADAS-COG and CDRSB scores at later time-points (starting at 12 months after baseline, up to 36 months). Based on the longitudinal study protocol, we considered 6 month intervals and assigned the actual visits to the closest 6-month bucket.

Figure 2 shows prediction accuracies of all considered methods. The proposed method performs significantly (all paired permutation p-values $< p_{max} = 0.03$) and substantially better than all other methods, with the difference increasing from the short term (12 months) to long term (36 months). For the benchmark models, prediction accuracy tends to drop more dramatically for longer time horizons. As above, training-fixed benchmarks perform the worst.

4 Conclusion

We presented a probabilistic, latent disease progression model for capturing the dynamics of the underlying pathology that is often shaped by risk factors such as genotype. Our work was motivated by real-world clinical applications, where irregular visiting patterns, missing variables, and inconsistent multi-modal assessments are ubiquitous. We applied the proposed method on a large dataset of Alzheimer’s disease for predicting clinical scores at varying time horizons with promising results. Future work will conduct a more detailed analysis of our proposed model. We are also interested in exploring the use of modern neural network based methods, such as Recurrent Neural Networks [20], for this application.

Acknowledgements. This work was supported by NIH grants R01LM012719, R01AG053949, and 1R21AG050122, and the NSF NeuroNex grant 1707312. We used data from Tadpole 2017 Challenge (<https://tadpole.grand-challenge.org/home/>).

References

1. Petersen, R.: Alzheimer’s Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* **74**, 201–209 (2010)
2. Alzheimer’s Association: Alzheimer’s disease facts and figures. *Alzheimer’s Dement.* 158–194 (2010)
3. Jack, C., et al.: Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade. *Lancet Neurol.* **9**, 119–128 (2010)
4. Dalca, A.V., Sridharan, R., Sabuncu, M.R., Golland, P.: Predictive modeling of anatomy with genetic and clinical data. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015 Part III. LNCS*, vol. 9351, pp. 519–526. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_62
5. Haxby, J., Raffaele, K., Gillette, J., Schapiro, M., Rapoport, S.: Individual trajectories of cognitive decline in patients with dementia of the Alzheimer type. *J. Clin. Exp. Neuropsychol.* **14**(4), 575–592 (1992)
6. Marta, M.: Modelling mini mental state examination changes in Alzheimer’s disease (2000)
7. Cano, S.J., et al.: The ADAS-COG in Alzheimer’s disease clinical trials: psychometric evaluation of the sum and its parts. *J. Neurol. Neurosurg. Psychiat.* **81**, 1363–1368 (2010)
8. O’Bryant, S., et al.: Staging dementia using clinical dementia rating scale sum of boxes scores: a Texas Alzheimer’s research consortium study. *Archiv. Neurol.* **65**, 1091–1095 (2008)
9. Corder, E.H., et al.: Gene dose of Apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families. *Science* **261**(5123), 921–923 (1993)
10. Katzman, R.: Education and the prevalence of dementia and Alzheimer’s disease. *Neurology* **43**, 13–20 (1993)
11. Fratiglioni, L., et al.: Prevalence of Alzheimer’s disease and other dementias in an elderly urban population relationship with age, sex, and education. *Neurology* **41**, 1886–1892 (1991)

12. Jack, C., et al.: Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* **52**, 1397–1403 (1999)
13. Nestor, S., et al.: Ventricular enlargement as a possible measure of Alzheimer’s disease progression validated using the Alzheimer’s disease neuroimaging initiative database. *Brain* **131**, 2443–2454 (2008)
14. Stern, Y.: Cognitive reserve in ageing and Alzheimer’s disease. *Lancet Neurol.* **11**(11), 1006–1012 (2012)
15. Ranganath, R., et al.: Black box variational inference. In: *Artificial Intelligence and Statistics* (2014)
16. Fischl, B.: FreeSurfer. *Neuroimage* **62**, 774–781 (2012)
17. Mormino, E., et al.: Polygenic risk of Alzheimer disease is associated with early-and late-life processes. *Neurology* **87**, 481–488 (2016)
18. Tran, D., et al.: Edward: a library for probabilistic modeling, inference, and criticism. arXiv preprint [arXiv:1610.09787](https://arxiv.org/abs/1610.09787) (2016)
19. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: *OSDI* (2016)
20. Che, Z., Purushotham, S., Cho, K., Sontag, D.A., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8** (2018)