An Exploration of Latent Structure in Observational Huntington's Disease Studies

Soumya Ghosh[1], PhD, Zhaonan Sun[1], PhD, Ying Li[1], PhD, Yu Cheng[1], PhD,
Amrita Mohan[2], PhD, Cristina Sampaio[2], MD, PhD, Jianying Hu[1], PhD
[1] IBM T.J. Watson Research Center, Yorktown Heights, NY
[2] CHDI Management/CHDI Foundation, Princeton, NJ

Abstract

Huntington's disease (HD) is a monogenic neurodegenerative disorder characterized by the progressive decay of motor and cognitive abilities accompanied by psychiatric episodes. Tracking and modeling the progression of the multi-faceted clinical symptoms of HD is a challenging problem that has important implications for staging of HD patients and the development of improved enrollment criteria for future HD studies and trials. In this paper, we describe the first steps towards this goal. We begin by curating data from four recent observational HD studies, each containing a diverse collection of clinical assessments. The resulting dataset is unprecedented in size and contains data from 19,269 study participants. By analyzing this large dataset, we are able to discover hidden low dimensional structure in the data that correlates well with surrogate measures of HD progression. The discovered structures are promising candidates for future consumption by downstream statistical HD progression models.

1 Introduction

Huntington's disease (HD) is a progressive, hereditary neurodegenerative disorder caused by an abnormal trinucleotide (CAG) repeat expansion in the *huntingtin* (HTT) gene.[1] Owing to its monogenic nature and 100% penetrance, predictive genetic tests are able to determine whether the disorder will manifest in an individual. Among genetically confirmed HD patients, a clinical diagnosis of HD is typically made when an individual exhibits overt, otherwise unexplained extrapyramidal movement disorders.[2] The mean age of clinical motor onset is strongly dependent on the length of the CAG repeat expansion, with longer expansions causing earlier onset. Statistical models[3] capturing this relationship have been developed to estimate years to clinical onset given CAG repeats. The availability of these estimates makes it possible to estimate an HD gene expansion carrier's (HDGEC) exposure to the toxic effects of the mutant HTT gene over time.

Along with motor disturbances, cognitive decline and psychiatric episodes are other typical characteristics of HD. Various clinical assessments have been designed to record the triad of motor, cognitive/behavioral and functional symptoms of HD. While motor impairment is currently considered the primary indicator of clinical onset, cognitive[4] and certain behavioral disorders[5] are known to surface years before motor onset. As such, clinical measurements along these dimensions are important for understanding the pre manifest progression of the disease. Functional assessments are responsible for measuring the quality of life of individuals with HD and prove useful for descriptive characterizations of post manifest HD progression.

Despite the availability of year to onset estimates and a plethora of clinical assessments, no fine grained staging of HD progression exists. Instead, researchers typically rely on coarse HD staging to characterize pre and post manifest subjects. Post-manifest subjects are often staged using the Shoulson and Fahn[6] rating scale. It divides HD patients into 5 stages (HD1 though HD5) based on an univariate summary of the subject's functional capabilities. Pre manifest HD individuals are sometimes categorized into early and late pre-manifest based on time to predicted motor onset. Here following,[5] we categorize subjects less than $T = 10.8$ years away from motor onset as late pre-manifest and those farther away from motor onset as early pre-manifest. However, staging based on such univariate criteria fails to account for HD progression along dimensions other than functional deterioration and time to onset. In order to capture the multi-faceted progression of HD, sophisticated statistical models[7] are necessary. The recent availability of large observational HD studies provides an unique opportunity for reliably learning such models from observed HD clinical assessments. However, the sheer number and variety of such assessments makes learning challenging. Furthermore, not all assessments are stable under repeated measurements or sensitive to HD progression. Noise, outliers, missing values, sparsity and heterogeneity among subjects further exacerbate the problem. Consequently, the development of computational models of HD progression based on clinical assessments remains a challenging open problem.

In this paper, we take the first steps towards this goal. We begin by cleaning, merging and aggregating data from four large observational HD studies. To the best of our knowledge, the aggregated dataset is the largest HD dataset studied to date. We posit that the observed clinical assessments in the merged dataset are a manifestation of some underlying low-dimensional disease process. We utilize a Bayesian latent variable model to recover this low dimensional structure. The employed model exhibits several desirable properties that make it well suited for our problem. First, it is able to seamlessly deal with data missing at random and does not require imputation of missing values. Next, it is robust to outliers, and obviates the need for any sort of outlier filtering. Finally, it avoids expensive cross validation based model selection procedures by simultaneously learning the dimensionality of the latent space along with other model parameters. We find that the discovered lower dimensional representations correlate well with surrogate measures of HD progression and are promising candidates for HD staging. Furthermore, by providing a dense, amalgamated representation of diverse clinical assessments they become attractive candidates for consumption by downstream statistical progression models.[7]

2 Data

In this study, we aggregate data from four prospective observational studies of HD, named Enroll-HD,[8] Registry,[9] Track-HD, Track-ON,[5] and Predict-HD.[10]

ENROLL-HD is a worldwide observational study of Huntingtons disease families. The study aims at providing a platform to support the design and conduct of future clinical trials, improving the understanding of the phenotypic spectrum and the disease mechanisms of HD and improving health outcomes for the participant/family unit. The study monitors how HD appears and changes over time in different subjects, and is open to either confirmed HD patients or those that are at-risk. Study participants were required to visit study sites annually, and undergo a comprehensive battery of clinical assessments. In this paper, we refer to the data generated from one visit of a participant as an *observation*. The Enroll-HD cohort used in this study contains data from 7614 subjects who made their baseline visits before October 2015, among which 5475 are HD gene carriers (*i.e.* CAG length ≥ 35), 1613 are control subjects (*i.e.* CAG length < 35), and the other 527 have unknown CAG length. Subjects have up to 4 visits in a year, with the average number of visits being 1.44.

REGISTRY is a multi-centre, multi-national observational study, managed by the European Huntington's Disease Network (EHDN), with no experimental intervention. REGISTRY aims at obtaining natural history data on many HD mutation carriers and individuals who are part of an HD family, relating phenotypical characteristics of HD, expediting the identification and recruitment of participants for clinical trials, developing and validating sensitive and reliable outcome measure for detecting onset and change over the natural course of pre-manifest and manifest HD. The REGISTRY cohort used in this study consists of 12108 participants, among which 7988 participants are HD gene mutation carriers (*i.e.* CAG ≥ 35), 758 are control participants with CAG length < 35, and the other 3894 participants do not have CAG length information. Participants have up to 15 annual visits, and the average number of visits equals to 2.90.

TRACK-HD is a multinational longitudinal HD study that examines clinical and biological findings of disease progression in individuals with pre-manifest and early-stage HD. Participants underwent annual clinical assessments for 36 months. At baseline, 366 participants were enrolled and 298 completed the 36 month study. Among them 97 were controls, 104 were pre manifest cases and 97 were post manifest HD patients.

TRACK-ON is a follow-up study to TRACK-HD, with the aim of testing for the presence of compensatory brain networks after structural brain changes in TRACK-HD pre-manifest participants. Participants in the study underwent annual clinical assessment for 24 months. At the baseline visit, 239 participants were enrolled, among them 106 were pre manifest HD, 22 were early stage HD and 111 participants were controls.

PREDICT-HD is another longitudinal observational study of subjects who chose to undergo predictive testing for the CAG expansion in the HD gene but did not meet criteria for a diagnosis of HD, i.e., pre manifest cases. Participants were recruited from multiple sites in the United States, Canada, Australia, and Europe beginning in October 2002. The goal of PREDICT-HD was to define the neurobiology of HD and to develop tools to allow clinical trials of potential disease-modifying therapies before at-risk individuals have diagnosable symptoms of the disease. It collected a variety

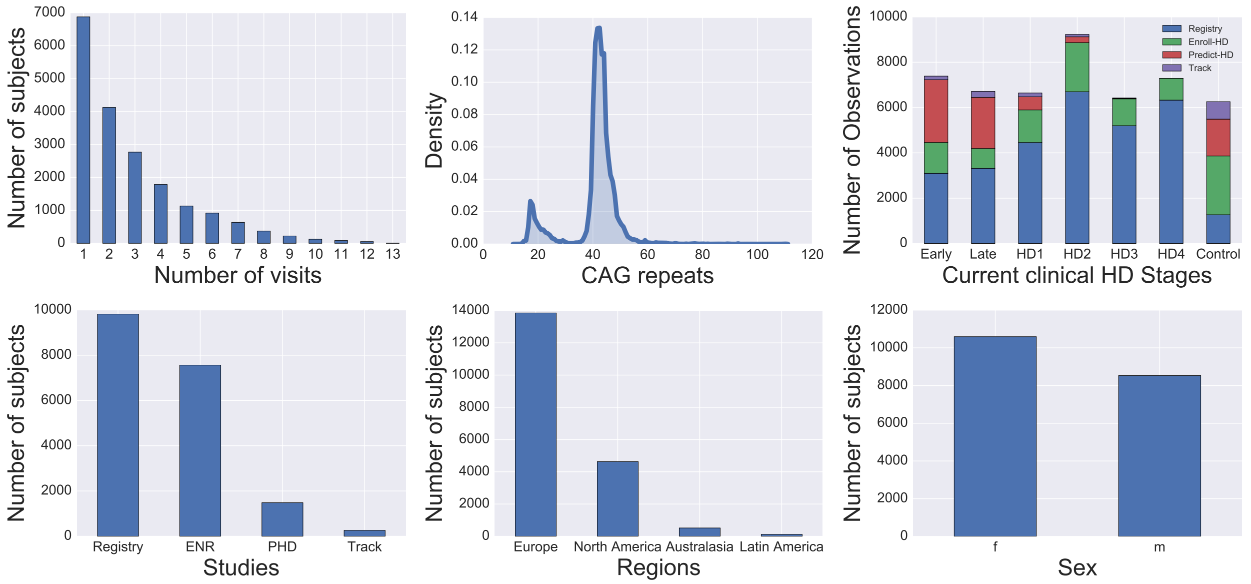

Figure 1: Descriptive statistics summarizing the aggregated HD dataset.

of biosamples including MRI, blood and urine samples, and several clinical assessments of cognitive, motor, functional and psychiatric outcomes to characterize the pre-manifest symptoms of HD, to document the rate of change of these variables during the years leading up to and following a clinical diagnosis of HD, and to investigate the relationship among neurobiologic factors, clinical diagnosis and CAG repeat length. The PREDICT-HD data used in this study consists of 1481 participants. Among them 316 are control subjects. Participants have up to 14 annual study visits, with the average number of visits equals to 5.23.

3 Materials and Methods

The four studies introduced in the previous section contain a diverse set of clinical assessments that span the gamut of clinical symptoms expressed by HD patients. Not all assessments across studies are compatible with each other or demonstrate appreciable sensitivity to HD progression. In this section, we briefly describe the process of aggregating information from the different studies and selecting a subset of assessments useful for tracking progression.

3.1 Multiple HD dataset aggregation

We began by matching subjects across studies using the unique Recorded HD participant ID. This unique identifier also allows us to recognize the small number of subjects that span multiple studies. Not all assessments are named consistently across studies. To cope, we analyzed study protocols and guidelines from the different studies and manually matched assessments and measurements across studies. We also corrected coding inconsistencies of certain categorical variables across studies. Finally, we performed a cross-study distributional check to filter out obvious erroneous measurements (for example, measurements outside valid ranges). After these steps we end up with an aggregate data set containing 2079 assessments and 55782 observations from 16553 HD subjects and 2716 controls. The Descriptive statistics summarizing various aspects of the combined dataset can be found in Figure 1.

3.2 Assessment Selection

Of the 2079 assessments not all are sensitive to HD progression. We select a smaller subset of assessments based on clinical feedback and statistical tests that measure correlation with surrogate measures of disease progression,[11]

Motor Features
Diagnostic confidence level (diagconf)
Ocular pursuit - Horizontal (ocularh)
Ocular pursuit - Vertical (ocularv)
Saccade initiation - Horizontal (sacinith)
Saccade initiation - Vertical (sacinitv)
Saccade velocity - Horizontal (sacvelh)
Saccade velocity - Vertical (sacvelv)
Dysarthria (dysarth)
Tongue protrusion (tongue)
Finger taps - Right (fingtapr)
Finger taps - Left (fingtapl)
Pronate/supinate-hands - Right (prosupr)
Pronate/supinate-hands - Left (prosupl)
Luria (luria)
Rigidity-arms - Right (rigarmr)
Rigidity-arms - Left (rigarml)
Bradykinesia-body (brady)
Maximal dystonia - Trunk(dysttrnk)
Maximal dystonia - RUE(dystrue)
Maximal dystonia - LUE(dystlue)
Maximal dystonia - RLE(dystrle)
Maximal dystonia - LLE(dystlle)
Maximal chorea - Face(chorface)
Maximal chorea - BOL(chorbol)
Maximal chorea - Trunk (chortrnk)
Maximal chorea - RUE (chorrue)
Maximal chorea - LUE (chorlue)
Maximal chorea - RLE (chorrle)
Maximal chorea - LLE (chorlle)
Gait (gait)
Tandem walking (tandem)
Retropulsion pull test (retropls)

Functional Features
Subject's independence scale (indepscl)
Occupation (occupatn)
Finances (finances)
Domestic chores (chores)
Activities of daily living (adl)
Care level (carelevl)
Cognitive Features
Symbol digit modality test, total number of correct responses (sdmt)
Stroop color naming test, total number of correct responses (scnt)
Stroop word recognition test, total number of correct responses in 45 seconds (swrt)
Stroop interference test, total number of correct responses (sit)
Verbal fluency test, total number of correct responses in 3 min, (verfl)
Mini-mental state examination, total score (mmse)
Behavior Features
HADS Anxiety subscore (hadsanx)
HADS Depression subscore (hadsdep)
HADS Irritability subscore (hadsirr)
HADS Outward irritability subscore (hadsout)
HADS Inward irritability subscore (hadsin)
Companion FrSBe Total (frsbef)
Companion FrSBe Apathy Subscale (apathyf)
Companion FrSBe Disinhibition Subscale (disinhibf)
Companion FrSBe Executive Subscale (execdyf)
Participant FrSBe Total (frsbes)
Participant FrSBe Apathy Subscale (apathys)
Participant FrSBe Disinhibition Subscale (disinhibs)
Participant FrSBe Executive Subscale (execdys)

Table 1: List of features used in analysis

while accounting for confounding factors such as age, gender and education level. The set of selected assessments categorized by the symptoms they measure are listed in Table 1. The motor assessments take values on a 0-4 ordinal scale, with 4 indicating severe impairment. The cognitive assessments are measured on an integer scale with *lower* values indicating higher degree of cognitive impairment. Functional and behavioral assessments are rated on an ordinal scale with higher scores indicating more intact functioning and more severe behavioral impairments. Detailed descriptions of these assessments can be found in the guidelines of the studies described in the previous section.

3.3 Robust Bayesian Latent Variable Analysis

In this paper, we restrict our attention to an observation level analysis. An observation $x_i \in \mathbf{R}^D$ is generated when a subject visits a study center and undergoes the D relevant assessments described in the previous section. We posit that these moderately high dimensional, noisy and sparse observations are a manifestation of an unobserved lower dimensional latent disease process. Here, we utilize a probabilistic generative latent variable model, a reformulation[12] of principal components analysis to recover this latent structure. The probabilistic formulation provides several advantages that are well suited to our application. First, owing to its generative nature, the model allows us to easily

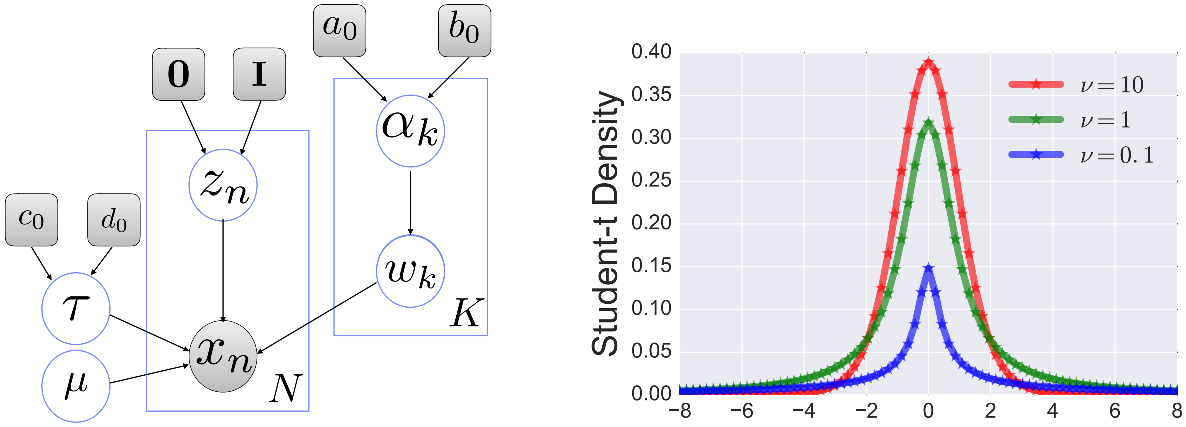

Figure 2: Left: Graphical model depicting the conditional dependencies assumed by the model. Shaded nodes indicate observed random variables. Plates indicate replication and arrows encode conditional dependencies. Right: Heavy tailed behavior of a zero mean, unit variance Student-t distribution for different degrees of freedom ν. The tails get heavier with decreasing ν.

marginalize out data missing at random. Second, the probabilistic formulation naturally deals with noisy data and allows us to easily incorporate robust likelihoods to deal with outliers. Finally, a further simple extension provides the benefits of automatic model selection — a data driven mechanism to infer the dimensionality of the latent space along with model parameters.

Let $\mathbf{x} = [x_1, \ldots, x_N] \in \mathbf{R}^{D \times N}$ denote a set of N observations and $\mathbf{z} = [z_1, \ldots, z_N] \in \mathbf{R}^{K \times N}$ be the corresponding latent representations with $K < D$. We assume that a observation x_n is generated according to the following generative process,

$$z_n \sim \mathcal{N}(0, I_K); \quad x_n \mid \mathrm{W}, \mu, z_n, \tau \sim \mathcal{N}(\mathrm{W}z_n + \mu, \tau^{-1}\mathbf{I}_D) \tag{1}$$

where $\mathcal{N}(\mathbf{m}, \Sigma)$ denotes a Gaussian distribution with mean $\mathbf{m}$ and covariance Σ, I_K and I_D denote K and D-dimensional identity matrices, $\mu \in \mathbf{R}^D$ is a bias term, $\mathrm{W} \in \mathbf{R}^{D \times K}$ is a linear mapping responsible for projecting the lower dimensional latent variables z_n to the observed data space. The model accounts for noise in the observed data, by assuming that the observed data is generated by adding isotropic Gaussian noise to the noise-free projections $\mathrm{W}z_n + \mu$. The scale of the observation noise is governed by the precision parameter τ.

Missing Values Dealing with data missing at random is straightforward under the proposed generative model. To see this, consider a data instance x_n with H missing dimensions and V observed dimensions, such that $V + H = D$. The conditional distribution of x_n then factorizes as follows,

$$\begin{aligned} p(x_n \mid \mathrm{W}, \mu, z_n, \tau) &= \prod_{d=1}^{D} \mathcal{N}(x_{dn} \mid (Wz_n)_d + \mu_d, \tau^{-1}) \\ &= \prod_{v=1}^{V} \mathcal{N}(x_{vn} \mid (Wz_n)_v + \mu_v, \tau^{-1}) \prod_{h=1}^{H} \int \mathcal{N}(x_{hn} \mid (Wz_n)_h + \mu_h, \tau^{-1}) dx_{hn} \\ &= \prod_{v} \mathcal{N}(x_{vn} \mid (Wz_n)_v + \mu_v, \tau^{-1}), \end{aligned} \tag{2}$$

where the second equality follows from the fact that $\int \mathcal{N}(x \mid \mu, \sigma^2) dx = 1$. Thus, missing values can be safely ignored and no imputation of data is required to apply the model.

Robust Likelihoods The set of clinical assessments analyzed in this paper contain many outliers. We employ robust likelihood models to prevent these outliers from biasing the analysis and the learned projections. To inject robustness in our Gaussian likelihoods, we place a Gamma prior over the noise precision τ,

$$\tau \sim \text{Gamma}(c_0 = \frac{\nu}{2}, d_0 = \frac{\nu}{2}). \tag{3}$$

Marginalizing over the nuisance noise precision τ,

$$\begin{aligned} p(x_n \mid W, z_n, \mu) &= \int_0^\infty \mathcal{N}(x_n \mid Wz_n + \mu, \tau^{-1} I_D) \text{Gamma}(\tau \mid \nu/2, \nu/2) d\tau \\ &= \prod_{d=1}^{D} \int_0^\infty \mathcal{N}(x_{nd} \mid (Wz_n)_d + \mu_d, \tau^{-1}) \text{Gamma}(\tau \mid \nu/2, \nu/2) d\tau \\ &= \prod_{d=1}^{D} t_\nu(x_{nd} \mid (Wz_n)_d + \mu_d, 1), \end{aligned} \tag{4}$$

we see that placing a Gamma prior on τ implies a heavy tailed distribution on the observations x_n — a Student-t distribution of unit variance and degree of freedom ν.[13] By retaining higher probability mass in the tails the Student-t distribution is more robust to outliers. The tails of the Student-t distribution get heavier as $\nu \rightarrow 0$, and as $\nu \rightarrow \infty$, the Student-t approaches a Gaussian distribution and ceases to be robust to outliers (Figure 2). The robust likelihoods obviate the need to perform extensive, error-prone outlier filtering.

Automatic Model Selection We further place an Automatic relevance determination (ARD)[14] prior on the linear mapping W, by placing independent Gaussian distributions on columns of W, ($\text{W}_k \in \mathbf{R}^D$),

$$\text{W}_k \mid \alpha_k \sim \mathcal{N}(0, \alpha_k^{-1} I_D) \tag{5}$$

along with a Gamma prior on the precision α_k, $\alpha_k \sim \text{Gamma}(a_0, b_0)$. ARD is a sparsity promoting[14] prior and it prunes away (sets to zero) columns of W that are not required to explain the observed data well by constraining the corresponding α_k to large values and thus restricting W_k to small values near zero.[12] The ARD priors obviate the need for expensive cross validation procedures often employed for determining K, allowing us to instead infer it from data simultaneously with other relevant model parameters.

The joint distribution of the resulting probabilistic model factorizes as follows,

$$\begin{aligned} p_0(\mathbf{x}, \text{W}, \mathbf{z}, \alpha, \tau \mid a_0, b_0, c_0, d_0, I_k) = \text{Gamma}(\tau \mid c_0, d_0) \prod_{k=1}^{K} \mathcal{N}(\text{W}_k \mid 0, \alpha_k^{-1} I_D) \text{Gamma}(\alpha_k \mid a_0, b_0) \\ \prod_{n=1}^{N} \mathcal{N}(z_n \mid 0, I_K) p(x_n \mid \text{W}z_n + \mu, \tau^{-1} I_D) \end{aligned} \tag{6}$$

The corresponding graphical model summarizing the conditional dependencies assumed by the model is shown in Figure 2.

Learning and Inference We are interested in learning the posterior distribution $p(\text{W}, \mathbf{z}, \alpha, \tau \mid \mathbf{x})$, where $\mathbf{z} = [z_1, \ldots, z_N] \in \mathbf{R}^{N \times K}$ and $\alpha = [\alpha_1, \ldots, \alpha_K]$. The marginal posterior $p(\mathbf{z} \mid \mathbf{x})$ is of particular importance for subsequent analysis of disease progression.

Unfortunately, the posterior distribution is intractable. Here we use learn an approximation to the intractable posterior using variational inference. Variational inference approximates the posterior $p(\text{W}, \mathbf{z}, \alpha, \tau \mid \mathbf{x})$ with a surrogate distribution $q(\text{W}, \mathbf{z}, \alpha, \tau \mid \phi)$, such that the Kullback-Leibler (KL) divergence between the two is minimized,

$$\hat{\phi} = \underset{\phi}{\text{argmin}} \; \text{KL}(q(\text{W}, \mathbf{z}, \alpha, \tau \mid \phi) || p(\text{W}, \mathbf{z}, \alpha, \tau \mid \mathbf{x})). \tag{7}$$

The set of parameter ϕ governing the variational distribution q are called the variational parameters. By turning the inference problem into an optimization problem these methods can leverage the large body of work on stochastic optimization[15] and scale to large N. However, an unconstrained optimization of Equation 7 will set $q = p$ and is not useful. To make progress, q is typically constrained to a family of tractable distributions $\mathcal{Q}$ and the optimization in Equation 7 causes us to approximate p with the closest (in KL sense) member of $\mathcal{Q}$. Here, we restrict $\mathcal{Q}$ to the family of the following fully factorized approximate distributions,

$$q(\mathrm{W}, \mathbf{z}, \alpha, \tau \mid \phi) = \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{N}(w_{dk} \mid \mu_{w_{dk}}, \sigma^2_{w_{dk}}) \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(z_{nk} \mid \mu_{z_{nk}}, \sigma^2_{z_{nk}}) \prod_{k=1}^{K} \mathrm{Gamma}(\alpha_k \mid a_k, b_k) \mathrm{Gamma}(\tau \mid c, d), \tag{8}$$

where ϕ is the set of all variational parameters $\{\{\mu_{w_dk}, \sigma^2_{w_{dk}}\}_{d=1,k=1}^{d=D,k=K}, \{\mu_{z_{nk}}, \sigma^2_{z_{nk}}\}_{n=1,k=1}^{n=N,k=K} \{a_k, b_k\}_{k=1}^{K}, c, d\}$. Thus, under this fully factorized variational approximation the posterior distribution of the k^{th} column of the linear mapping W are approximated using a diagonal Gaussian $p(W_k \mid \mathbf{x}) \approx \mathcal{N}(\mu_{w_k}, diag(\sigma^2_{w_k}))$, where $\mu_{w_k} = [\mu_{w_{1k}}, \dots \mu_{w_{Dk}}]^T$ and $diag(\sigma^2_{w_k})$ denotes a matrix whose diagonal is populated by the vector $\sigma^2_{w_k} = [\sigma^2_{w_{1k}}, \dots \sigma^2_{w_{Dk}}]^T$. The posterior $p(z_n \mid \mathbf{x})$ is also approximated by another diagonal Gaussian $\mathcal{N}(\mu_{z_n}, diag(\sigma^2_{z_n}))$.

It can be shown[16] that minimizing the KL divergence is equivalent to maximizing the following lower bound to the marginal likelihood $p(\mathbf{x} \mid \theta)$:

$$p(\mathbf{x} \mid \theta) \geq \mathcal{L}(\phi) = E_q[\log p(\mathbf{x} \mid \mathrm{W}, \mathbf{z}, \tau)] - \mathrm{KL}(q(\mathrm{W}, \mathbf{z}, \alpha, \tau \mid \phi) || p(\mathrm{W}, \mathbf{z}, \alpha, \tau \mid \theta)), \tag{9}$$

with respect to ϕ. Here, we denote the set of all hyper-parameters $\{a_0, b_0, c_0, d_0, I_K\}$ as θ and $p(\mathrm{W}, \mathbf{z}, \alpha, \tau \mid \theta) = \mathrm{Gamma}(\tau \mid c_0, d_0) \prod_{n=1}^{N} \mathcal{N}(z_n \mid 0, I_K) \prod_{k=1}^{K} \mathcal{N}(\mathrm{W}_k \mid 0, \alpha_k) \mathrm{Gamma}(\alpha_k \mid a_0, b_0)$ represents the prior distribution. $\mathcal{L}$ is sometimes called the Expected Lower BOund (ELBO) and is made up of two counter acting terms(Equation 9). The first term $E_q[\log p(\mathbf{x} \mid \mathrm{W}, \mathbf{z}, \tau)]$, measures the average reconstruction error and penalizes solutions that do not reconstruct the observations well. The KL term may be interpreted as a regularizer that penalizes solutions that deviate too strongly from the prior distribution.

In general, maximizing Equation 9 can be challenging. However, our robust Bayesian PCA model is a member of the conditionally conjugate[17] family of models. For models in this class, fixed point updates for the variational parameters are available and the ELBO ($\mathcal{L}(\phi)$) can be optimized by repeatedly applying the updates in a coordinate ascent algorithm. We used BayesPy (`http://bayespy.org`) to perform these updates. See[18] for the corresponding derivations. At convergence the fixed point updates provide us with a locally optimal set of variational parameters $\hat{\phi}$ that completely specify the variational approximation to the posterior.

4 Analysis and Results

In this section we discuss the latent structure recovered by the model developed in Section 3.3 when applied to the aggregated HD dataset. Since, we are primarily interested in exploring HD progression, we excluded control observations from our analysis. The model assumes that all observed data dimensions share a common scale. We accommodate this assumption by preprocessing all features to have unit variance. We place uninformative priors on α and noise precision τ by setting a_0, b_0, c_0, d_0 to 10^{-3}. This allows the observed data to easily overwhelm the prior. HD clinical assessment inventories are designed to measure progression of clinical symptoms along motor, cognitive, functional and behavioral domains. To carefully explore properties of the different domains, we model each domain separately with an independent robust latent variable model. We initialize these models with $D-1$ bases, where D is the number of assessments in the particular domain[1] and let the ARD prior prune away spurious bases and recover the optimal latent dimensionality K. For each domain, we performed five variational inference runs each from a different random initialization of the variational parameters and selected the solution that achieved the highest ELBO value.

The posterior means ($E[\mathrm{W} \mid \mathbf{x}] \approx \mu_{\mathrm{W}}$) of the discovered linear mappings are shown in Figure 3. In the motor domain, we discover a latent dimensionality of $K = 15$. We also find that the primary principal component accounts

[1] $D_{motor} = 32, D_{behavior} = 13, D_{cognitive} = 6, D_{functional} = 6$. See Table 1

for 60% of the total variance. Functional and Cognitive domains are best explained by three dimensional latent representations with the dominant principal component accounting for 85% and 87% of the total variance. We discover a five dimensional latent space for the behavioral domain, with 50% of the total variance being explained by the dominant principal component.

We also find (Figure 4) that in motor, functional and cognitive domains, the primary direction of variation correlates well with a genetic surrogate measure of disease progression — CAP score. It is defined as $y * (r - L)/S$, where y is the current age of the subject, r is the number of CAG repeats, L and S are constants. In this work we use $L = 30$ and $S = 6.27$.[2] On average, higher CAP scores indicate more advanced progression. To quantify the correlation between the dominant direction of variance and CAP, we computed the Pearson correlation coefficient (ρ), a measure of the strength of linear association between two variables, for each domain. For motor, cognitive and functional domains the computed correlation coefficients were $\rho_{\text{motor}} = 0.71$, $\rho_{\text{cog}} = -0.64$, $\rho_{\text{func}} = -0.63$. The behavioral domain exhibits much weaker correlation with $\rho_{\text{beh}} = 0.03$.

Further, cross-correlating with the coarse HD stages, we find that post-manifest progression along the dominant motor and cognitive principal components correlate well with the Shoulson and Fahn stages. However, the separation of these stages is clearest in the functional domain. This is unsurprising, since the Shoulson and Fahn staging is based on the total function score, an aggregate measure of functional capacity. More interestingly, in pre-manifest observations we find that the cognitive features are able to best distinguish early and late pre-manifest observations. This corroborates previous analysis on smaller studies[4,5] that found the cognitive assessments well suited for monitoring progression in pre-Manifest HD cases. We also find the behavioral assessments considered here to correlate poorly with both CAP progression and coarse HD staging.

In all domains, subsequent directions of variance show no significant correlation with CAP. The amount of variance not explained by CAP progression ranges from 13% in the cognitive domain to 40% in motor assessments. This significant residual variance likely stems from several factors — disease progression along dimensions not well characterized by CAP, heterogeneity in the subject population, variance between studies and study sites, and noise in the recording process. A detailed analysis of these factors is a promising direction of future research.

5 Discussion

The clinical symptoms of HD are multi-faceted. Tracking the progression of these symptoms along the diverse dimensions of motor, cognitive, behavioral and functional impairment is a challenging computational problem. In this paper, we take the first steps towards attacking this challenging problem. By analyzing data from four large HD observational studies, each containing a diverse set of clinical assessments, we are able to discover intermediate representations that correlate well with surrogate measures of HD progression and are amenable to downstream progression models.

6 Acknowledgements

Data used in this work was generously provided by the research participants in ENROLL-HD and CHDI Foundation, Inc that financially supports the study. Data used in this work was generously provided by the research participants from the TRACK-HD and TRACK-ON studies and made available by the TRACK Investigators. Data used in this work was generously provided by the the participants of the European Huntington's Disease Network (EHDN) REGISTRY study and made available by the EHDN REGISTRY Investigators. Data used in this work was generously provided by the participants in PREDICT-HD and made available by the PREDICT-HD Investigators and Coordinators of the Huntington Study Group, Jane Paulsen, Principal Investigator.

References

[1] Group HDCR, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72:971–983.

[2] Ross CA, Aylward EH, Wild EJ, Langbehn DR, Long JD, Warner JH, et al. Huntington disease: natural history, biomarkers and prospects for therapeutics. Nature reviews Neurology. 2014;10(4):204.

[3] Langbehn DR, Hayden MR, Paulsen JS. CAG-repeat length and the age of onset in Huntington disease (HD): a review and valida-

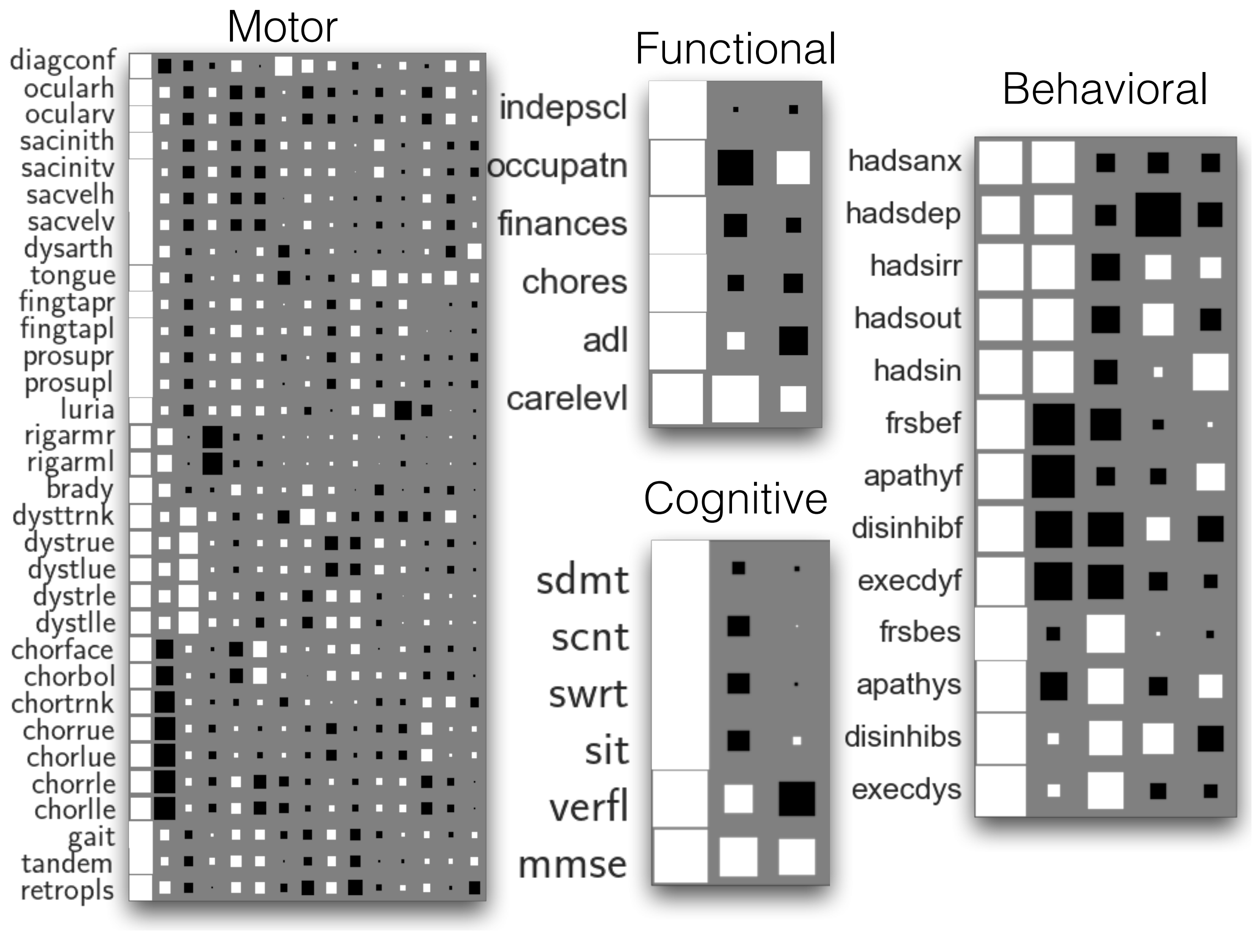

Figure 3: Hinton diagrams of posterior means ($E[\mathrm{W} \mid \mathbf{x}]$) of the loading matrices discovered by the model for the different domains. White squares indicate positive values and black indicates negative values. Larger squares indicate larger magnitude. Within each domain the columns of the matrix are sorted from left to right according to the proportion of variance explained by the column. The leftmost column corresponds to the direction of maximum variance within that domain.

tion study of statistical approaches. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2010;153(2):397–408.

[4] Stout JC, Paulsen JS, Queller S, Solomon AC, Whitlock KB, Campbell JC, et al. Neurocognitive signs in prodromal Huntington disease. Neuropsychology. 2011;25(1):1.

[5] Tabrizi SJ, Scahill RI, Owen G, Durr A, Leavitt BR, Roos RA, et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. The Lancet Neurology. 2013;12(7):637–649.

[6] Shoulson I, Kurlan R, Rubin A, Goldblatt D, Behr J, Miller C, et al. Assessment of functional capacity in neurodegenerative movement disorders: Huntington's disease as a prototype. Quantification of Neurologic Deficit, T Munsat (ed). 1989;p. 271–283.

[7] Wang X, Sontag D, Wang F. Unsupervised learning of disease progression models. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2014. p. 85–94.

[8] Mestre T, Fitzer-Attas C, Giuliano J, Landwehrmeyer B, Sampaio C. Enroll-HD: A Global Clinical Research Platform for Huntingtons Disease (S25. 005). Neurology. 2016;86(16 Supplement):S25–005.

[9] Orth M, Network EHD, et al. Observing Huntington's disease: the European Huntington's disease network's REGISTRY. Journal of Neurology, Neurosurgery & Psychiatry. 2010;p. jnnp–2010.

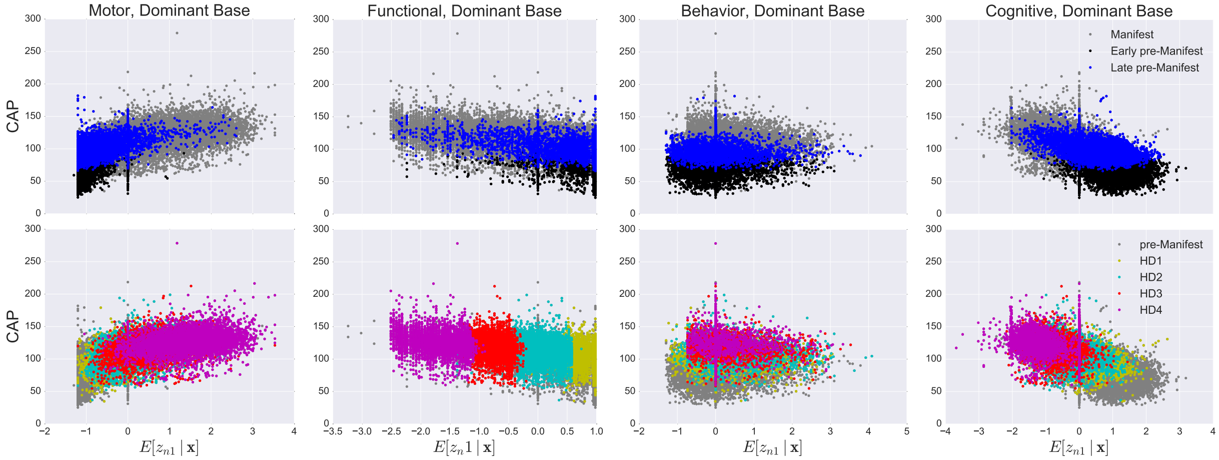

Figure 4: Correlation of the dominant direction of variation in the motor, functional, cognitive and behavioral domains with CAP and coarse HD stages. The x axis corresponds to the posterior means ($E[z_{n1} \mid \mathbf{x}]$) of projections of different observations along the dominant direction of variance (the leftmost column of the matrices displayed in Figure 4), the y axis represents the CAP score associated with the observation. The colors correspond to the coarse HD stage of the observation. In all but the behavior domain, the projections correlate well with CAP and HD stages. In the motor domain, increasing projections along the dominant variance detection indicate increasing motor impairment and thus correlate positively with CAP scores. In functional and behavioral domains, increasing scores increase decreasing impairment and thus correlate negatively with CAP.

[10] Paulsen J, Langbehn D, Stout J, Aylward E, Ross C, Nance M, et al. Detection of Huntingtons disease decades before diagnosis: the Predict-HD study. Journal of Neurology, Neurosurgery & Psychiatry. 2008;79(8):874–880.

[11] Sun Z, et al. Manuscript in preperation. 2017;.

[12] Bishop CM. Variational principal components. In: Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470). vol. 1. IET; 1999. p. 509–514.

[13] Archambeau C, et al. Probabilistic models in noisy environments: and their application to a visual prosthesis for the blind. UCL.; 2005.

[14] Tipping ME. Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research. 2001;1(Jun):211–244.

[15] Hoffman MD, Blei DM, Wang C, Paisley JW. Stochastic variational inference. Journal of Machine Learning Research. 2013;14(1):1303–1347.

[16] Bishop CM. Pattern recognition and Machine Learning. vol. 128; 2006.

[17] Wang C, Blei DM. Variational inference in nonconjugate models. Journal of Machine Learning Research. 2013;14(Apr):1005–1031.

[18] Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research. 2010;11(Jul):1957–2000.