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

Soumya Ghosh¹, Zhaonan Sun¹, Ying Li¹, Yu Cheng¹, Amrita Mohan², Cristina Sampaio², and Jianying Hu¹,

¹Center for Computational Health, IBM Research
²CHDI Foundation
Huntington’s Disease (HD)

Source: hdsa.org
HD prevalence

Minimum Prevalence of HD (per 100,000)
- >5
- 1-5
- 0.5-1
- 0.1-0.5
- No data available

Warby et al., 2009
HD symptoms

- Unsteady gait & involuntary movements (chorea)
- Slurred speech, difficulty in swallowing
- Forgetfulness & impaired judgment
- Personality changes, mood swings & depression
- Activities of daily living severely hampered

*HD affects the whole brain, but certain areas are more vulnerable than others. Pictured above are the basal ganglia - a group of nerves cell clusters, called nuclei. These nuclei play a key role in movement and behavior control and are the parts of the brain most prominently affected in early HD. Source: hdsa.org*
HD natural history

Ross et. al., Nature Reviews, 2004
Clinical assessments

Motor

Clinical Assessments

Cognitive & Behavioral

Functional
Clinical assessment examples

A Finger tapping

B Tongue force

Finger Tapping and Tongue Protrusion

Symbol Digits Modalities Test (SDMT)

source: Weir et. al., Lancet, 2011

source: clevelandclinic.org
Observational studies

TRACKHD

PREDICT-HD
An observational study of the earliest signs of Huntington disease

Enroll-HD

EUROPEAN HUNTINGTON’S DISEASE NETWORK
Combined dataset

- Largest HD dataset studied to-date,
  - 16,553 HD subjects and 2,716 Controls
  - ~ 2000 Assessments
Combined dataset
Assessment selection

- All put together there are ~ 2000 assessments.
  - Not all are available in all studies or even between centers in a study
  - Not all are stable under repeated measurements
  - Some are more noisy than others

- We selected a subset based on clinical feedback and,
  - Correlation with surrogate measures of HD progression
  - Ability to discriminate between clinical HD stages and controls
  - 57 assessments
Assessments are high dimensional, but clearly not independent.

We posit that there is a hidden lower dimensional structure underlying the assessments.

Discovering this structure is challenging, 
- Noisy
- High dimensional
- Missing values
Robust Probabilistic PCA

Unobserved Lower Dimensional Representation

Unobserved Lower Dimensional Basis

Observed Clinical Assessments
Robust Likelihoods:

\[ x_n \mid W, \mu, z_n, \tau \sim \mathcal{N}(Wz_n + \mu, \tau^{-1}I) \]
\[ \tau \sim \text{Gamma}(c_0, d_0) \]
Automatic Relevance Determination priors:
Sparsity promoting; turns off additional bases

\[
\alpha_k \sim \text{Gamma}(a_0, b_0)
\]

\[
w_k \mid \alpha_k \sim \mathcal{N}(0, \alpha_k^{-1}I)
\]
Learning

- We learn the model by maximizing the marginal likelihood of the data.

\[ p(\mathbf{x}; \theta) \geq \mathcal{L}(W, z; \theta) \]

- This is intractable. We maximize a *lower bound* to the marginal likelihood (*variational inference*)

- Generalization of EM; involves cycling over fixed point updates.
Inferred Embedding and Bases
Discovered Bases $E[W | x]$

Motor

% var

02 02 02 03 06 11 60

Occulo-motor  Dystonia  Chorea
<table>
<thead>
<tr>
<th></th>
<th>Cognitive</th>
<th>Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>% var</td>
<td>87 07 04</td>
<td>85 10 04</td>
</tr>
<tr>
<td>SDMT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCNT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word Recog.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal Fluency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indep. scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finances</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activities of Daily living</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carelevel</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HD severity vs embedding

$\rho_{\text{motor}} = 0.71 \quad \rho_{\text{cog}} = -0.64 \quad \rho_{\text{func}} = -0.63$

$\text{CAP} = \text{cag repeats} \times \text{age}$. A coarse measure of disease progression.

Higher CAP $\Rightarrow$ More advanced HD
Summary

- Curated the largest observational HD dataset to date

- Robust probabilistic latent variable analysis
  - Generates lower dimensional embeddings that track well with surrogate measures of HD progression.
  - Discovers interesting latent structure
    - Dominant base tracks well with CAP, subsequent bases don’t.
    - Non-negligible unexplained variance.
    - Behavior assessments appear less reliable.

- Follow up preliminary work using these embeddings has resulted in exciting new data driven HD stages.
Questions

ghoshso@us.ibm.com