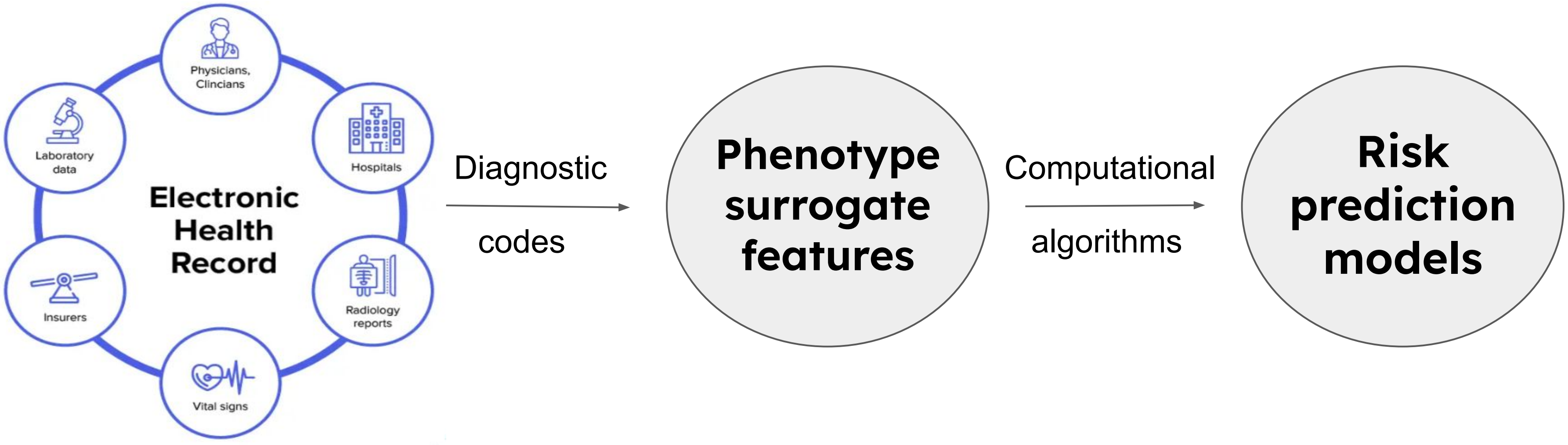
A brief overview of "A semi-supervised adaptive Markov Gaussian embedding process (SAMGEP) for prediction of phenotype event times using electronic health records (EHRs)"

by Ismail Benchekroun & Quynh (Christina) Vu

2022-12-06

Motivation: What is a phenotype event?

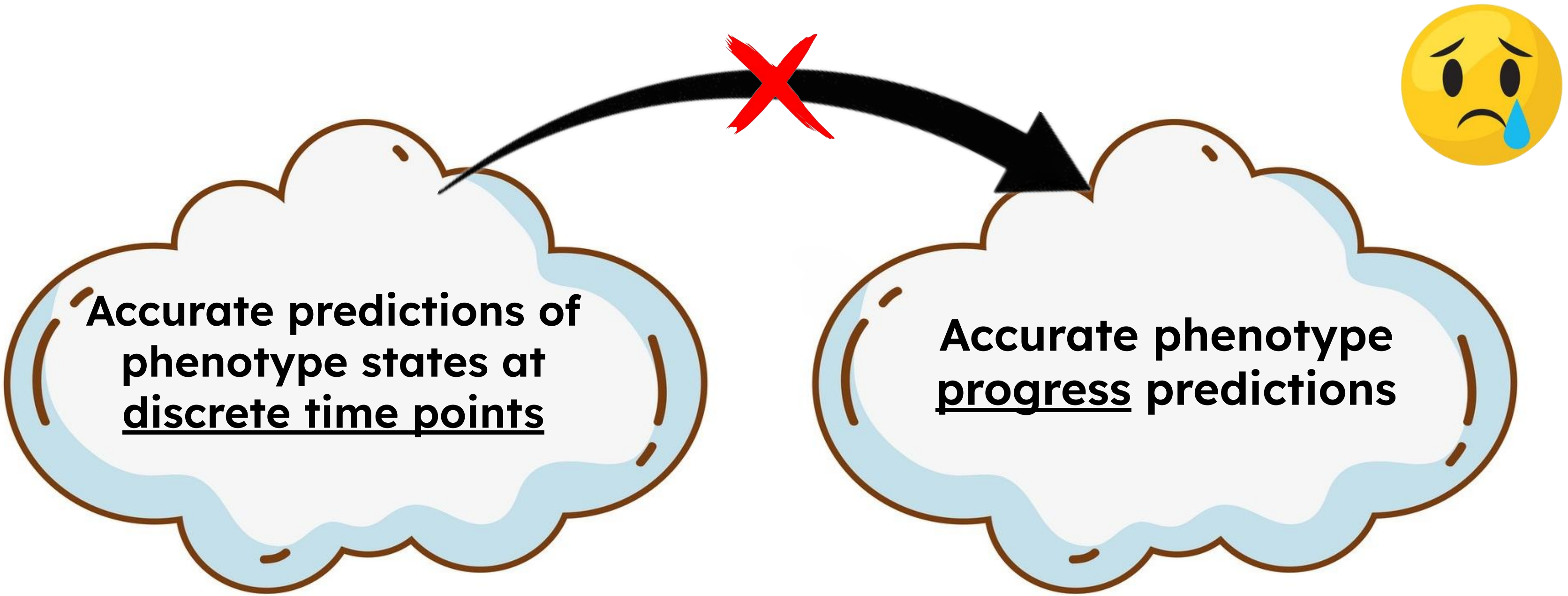


A phenotype event is a set of physical and latent health outcomes caused by a medical condition.

Motivation: Previous methods and Drawbacks

- ➢ **Unsupervised and semi-supervised methods**
  - ➥ Reliant on a set of designated codes
  - ➥ Sensitive to sparsity
- ➢ **Hidden Markov Models (HMMs) based models**
  - ➥ Not reflective or clinically relevant
- ➢ **Supervised learning methods (i.e. Reverse Time Attention Models (RETAIN))**
  - ➥ Reliant on large numbers of standardized labels for stable performance

Motivation: Main ground for a new EHRs implementation method



Introduction: A semi-supervised adaptive Markov Gaussian embedding process (SAMGEP)

➢ A state-dependent Gaussian process

- Capable of handling the inherent sparsity in EHRs data
- Requiring only a small number of standardized labels
- Reproducible and Generalizable
- Robust to model misspecification

enabling on-time allocation of interventions and treatments

Methods

Notation

- ➢ i^{th} patient, j^{th} feature, t^{th} time period
- ➢ T_i - **# of time periods** for patient i:
- ➢ $\mathbf{y_i} = (y_{i,1}, \ldots, y_{i,T_i})$ - **phenotype state sequence** for patient i, collected for n patients
- ➢ $\mathbf{C_{i,t}}$ - **feature vector** for patient i at time t, a p-dimensional vector, collected for N patients
- ➢ $H_i = \log(\text{mean healthcare encounter count per month} + 1)$
  - **healthcare utilization** for patient i, collected for N patients
- ➢ **n << N** (ie $\#_{labeled} << \#_{unlabeled}$)

Methods

Notation

- ➢ i^{th} patient, j^{th} feature, t^{th} time period
- ➢ T_i - **# of time periods** for patient i:
- ➢ $\mathbf{y_i} = (y_{i,1}, \ldots, y_{i,T_i})$ - **phenotype state sequence** for patient i, collected for n patients
- ➢ $\mathbf{C_{i,t}}$ - **feature vector** for patient i at time t, a p-dimensional vector, collected for N patients
- ➢ $H_i = \log(\text{mean healthcare encounter count per month} + 1)$
  - **healthcare utilization** for patient i, collected for N patients
- ➢ **n << N** (ie $\#_{labeled} << \#_{unlabeled}$)

Methods

Notation

- i^{th} patient, j^{th} feature, t^{th} time period
- T_i - **# of time periods** for patient i:
- $\mathbf{y_i} = (y_{i,1}, \ldots, y_{i,T_i})$ - **phenotype state sequence** for patient i, collected for n patients
- $\mathbf{C_{i,t}}$ - **feature vector** for patient i at time t, a p-dimensional vector, collected for N patients
- $H_i = \log(\text{mean healthcare encounter count per month} + 1)$
  - **healthcare utilization** for patient i, collected for N patients
- **n << N** (ie $\#_{labeled} << \#_{unlabeled}$)

Methods

Notation

- ➢ i^{th} patient, j^{th} feature, t^{th} time period
- ➢ T_i - **# of time periods** for patient i:
- ➢ $\mathbf{y_i} = (y_{i,1}, \ldots, y_{i,T_i})$ - **phenotype state sequence** for patient i, collected for n patients
- ➢ $\mathbf{C_{i,t}}$ - **feature vector** for patient i at time t, a p-dimensional vector, collected for N patients
- ➢ $H_i = \log(\text{mean healthcare encounter count per month} + 1)$
  - **healthcare utilization** for patient i, collected for N patients
- ➢ **n << N** (ie $\#_{labeled} << \#_{unlabeled}$)

Methods

Producing patient-timepoint embeddings

$$\mathbf{X_{i,t}} = \mathbf{C_{i,t}}\mathbf{W}_{p\times p}\mathbf{V}_{m\times p}^{T}$$

Methods

Producing patient-timepoint embeddings

$$\mathbf{X_{i,t}} = \mathbf{C_{i,t}} \mathbf{W}_{p\times p} \mathbf{V}^{T}_{m\times p}$$

Weight matrix

- maximizing L1-regularized linear discriminant analysis (LDA)

$$\mathbf{D}(\mathbf{W}) = (\mu_{\mathbf{1}} - \mu_{\mathbf{0}})^{\mathbf{T}} \sum_{\mathbf{X}}^{\mathbf{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{0}}) - \lambda ||\mathbf{W}||_1^1$$

Methods

Producing patient-timepoint embeddings

$$\mathbf{X_{i,t}} = \mathbf{C_{i,t}} \mathbf{W}_{p \times p} \mathbf{V}_{m \times p}^{T}$$

Weight matrix

- maximizing L1-regularized linear discriminant analysis (LDA)
- **Using labeled set only!!**

Methods

Patient embeddings follow a Gaussian Process

$$\boldsymbol{\mu_i}(\boldsymbol{t}) = E(\boldsymbol{X_{i,t}}) = \boldsymbol{\mu}_0(1 - Y_{i,t}) + \boldsymbol{\mu}_1 Y_{i,t} + \boldsymbol{\mu}_H H_i + \boldsymbol{\mu}_{YH} H_i Y_{i,t} + \boldsymbol{\mu}_2 t + \boldsymbol{\mu}_3 \log t + \boldsymbol{\mu}_4 Y_{i,t} t + \boldsymbol{\mu}_5 Y_{i,t} \log t$$

$$E\left[\epsilon_{i,t,k} | \epsilon_{i,t-1,k}\right] = r\tau_k \epsilon_{i,t-1,k}.$$

Phenotype state follows a Markov Process

$$P(Y_{i,t} = y | Y_{i,t-1} = y_{t-1}, H_i) = expit\left(\lambda_0(1 - y_{t-1}) + \lambda_1 y_{t-1} + \lambda_2 t + \lambda_3 \log t + \lambda_H H_i\right)$$

Methods

Patient embeddings follow a Gaussian Process

$$\boldsymbol{\mu}_{\boldsymbol{i}}(\boldsymbol{t}) = E(\boldsymbol{X}_{\boldsymbol{i},\boldsymbol{t}}) = \boldsymbol{\mu}_0(1 - Y_{i,t}) + \boldsymbol{\mu}_1 Y_{i,t} + \boldsymbol{\mu}_H H_i + \boldsymbol{\mu}_{YH} H_i Y_{i,t} + \boldsymbol{\mu}_2 t + \boldsymbol{\mu}_3 \mathrm{log}t + \boldsymbol{\mu}_4 Y_{i,t} t + \boldsymbol{\mu}_5 Y_{i,t} \mathrm{log}t$$

$$E\left[\epsilon_{i,t,k} | \epsilon_{i,t-1,k}\right] = r\tau_k \epsilon_{i,t-1,k}.$$

Phenotype state follows a Markov Process

$$P(Y_{i,t} = y | Y_{i,t-1} = y_{t-1}, H_i) = expit\left(\lambda_0 (1 - y_{t-1}) + \lambda_1 y_{t-1} + \lambda_2 t + \lambda_3 \mathrm{log}t + \lambda_H H_i\right)$$

We want to estimate conditional posterior $\hat{p_{it}} = E[Y_{i,t} | \mathbf{X}]$

How do we estimate phenotype?

Expectation-Maximization:

1. Initialize parameters
2. Compute probability of $\hat{p_{it}}$
3. Use new $\hat{p_{it}}$ to compute new estimates of parameters
4. Iterate steps 2 & 3 until convergence

How do we estimate phenotype?

Expectation-Maximization

1. **Initialize parameters (supervised learning)**
2. Compute probability of $\hat{p}_{it}$
3. Use new $\hat{p}_{it}$ to compute new estimates of parameters
4. Iterate steps 2 & 3 until convergence

Initialize parameters using MLE on <u>labeled set</u>

- Logistic regression for $Y_{it}|Y_{i(t-1)}, H_i, t$
- Generalised least squares for $X_i|Yi$'s Gaussian process

How do we estimate phenotype?

Expectation-Maximization

1. Initialize parameters
2. **Compute probability of $\hat{p}_{it}$ ($\hat{p}_{sup}$)**
3. Use new $\hat{p}_{it}$ to compute new estimates of parameters
4. Iterate steps 2 & 3 until convergence

Compute probability of $\hat{p}_{it}$ for <u>unlabeled set</u>

$$= \frac{\sum_{u=0}^{1}\sum_{w=0}^{1} P(Y_{i,t-1}=u)P(Y_{i,t}=1|Y_{i,t-1}=u)P(Y_{i,t+1}=w|Y_{i,t}=1)f(\boldsymbol{X}_{i,t-1},\boldsymbol{X}_{i,t},\boldsymbol{X}_{i,t+1}|Y_{i,t-1},Y_{i,t},Y_{i,t+1})}{\sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{w=0}^{1} P(Y_{i,t-1}=u)P(Y_{i,t}=v|Y_{i,t-1}=u)P(Y_{i,t+1}=w|Y_{i,t}=v)f(\boldsymbol{X}_{i,t-1},\boldsymbol{X}_{i,t},\boldsymbol{X}_{i,t+1}|Y_{i,t-1},Y_{i,t},Y_{i,t+1})}$$

How do we estimate phenotype?

Expectation-Maximization

1. Initialize parameters
2. Compute probability of $\hat{p_{it}}$ $(\hat{p}_{sup})$
3. **Use new $\hat{p_{it}}$ to compute new estimates of parameters $(\hat{p}_{semisup})$**
4. Iterate steps 2 & 3 until convergence

Use new $\hat{p_{it}}$ to update parameter estimates on <u>unlabeled set</u>

- Weighted logistic regression for $\hat{p_{i1}}|H_i$
- Generalised least squares for $X_i|Y i$'s Gaussian process

How do we estimate phenotype?

Expectation-Maximization

1. Initialize parameters
2. Compute probability of $\hat{p_{it}}$ $(\hat{p}_{sup})$
3. Use new $\hat{p_{it}}$ to compute new estimates of parameters $(\hat{p}_{semisup})$
4. ~~**Iterate steps 2 & 3 until convergence**~~

No need to iterate!

- ➢ Initial parameters are consistent estimators already
- ➢ Reduces computational cost
- ➢ Performance not sensitive to max # of iterations

How do we estimate phenotype?

Expectation-Maximization

1. Initialize parameters
2. Compute probability of $\hat{p_{it}}$ ($\hat{p}_{sup}$)
3. Use new $\hat{p_{it}}$ to compute new estimates of parameters ($\hat{p}_{semisup}$)
4. ~~Iterate steps 2 & 3 until convergence~~

4. Weighted sum of $\hat{p}_{sup}$ and $\hat{p}_{semisup}$:

$$\hat{\mathbf{p}} = \alpha \hat{\mathbf{p}}_{\mathbf{sup}} + (1 - \alpha)\hat{\mathbf{p}}_{\mathbf{semisup}}$$

Results: About the datasets

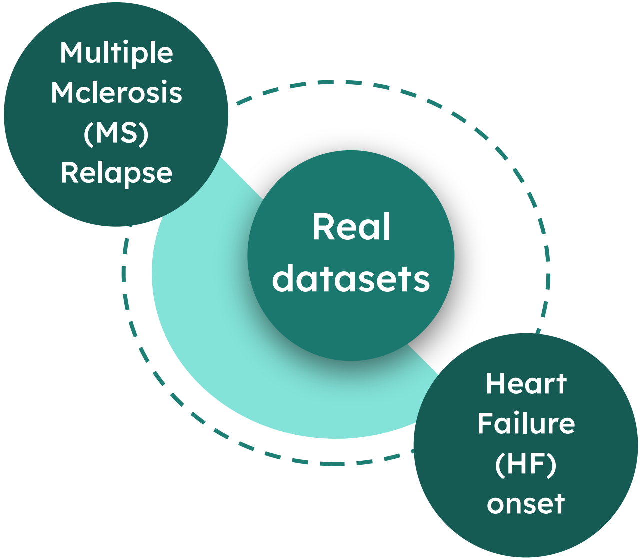
➢ **Simulation experiment**

➥ To assess the robustness of SAMGEP to violations of model assumptions

followed by

➢ **Analyses of real-world datasets**

➥ To compare the predictive accuracies between SAMGEP and previous methods



Results: Key findings

➢ **Simulation experiment**

⏩150 count features

⏩1000, 5000, and 20000 unlabelled patients

⏩100 labelled patients

`simulations were run with the number of standardized labels varying from 5 to 100`

➥ **Robust to model misspecification**

➥ **Optimal performance achieved when n varies from 50 to 100 labels**

100 labelled patients

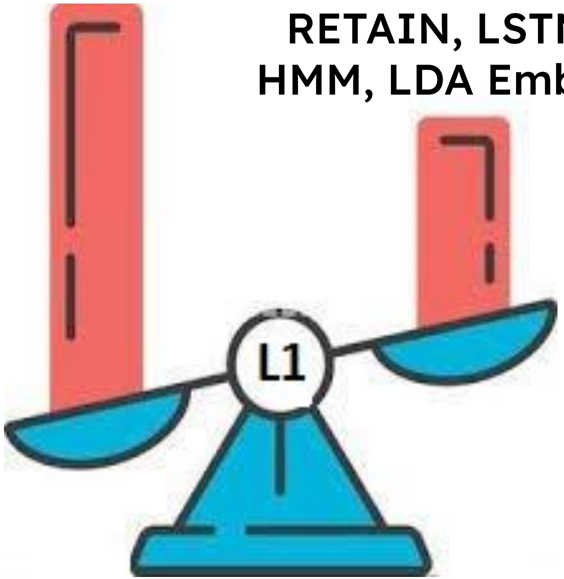
Results: Key findings

➢ **Simulation experiment**

🔼Y|T and X|Y are correctly specified

SAMGEP

RETAIN, LSTM, HMM, LDA Embed



🔼Y is independent of T

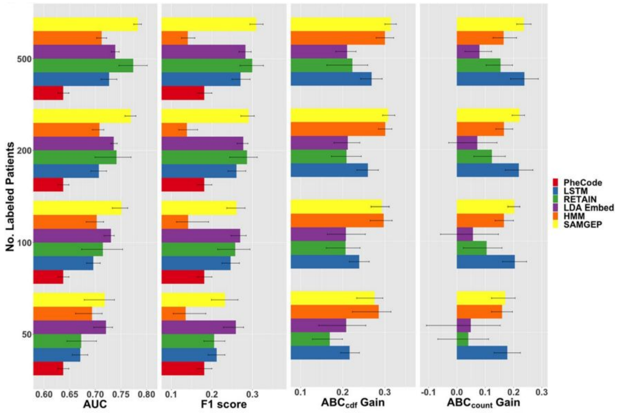
🔼 If Y is over-parametrized

Results: Key findings

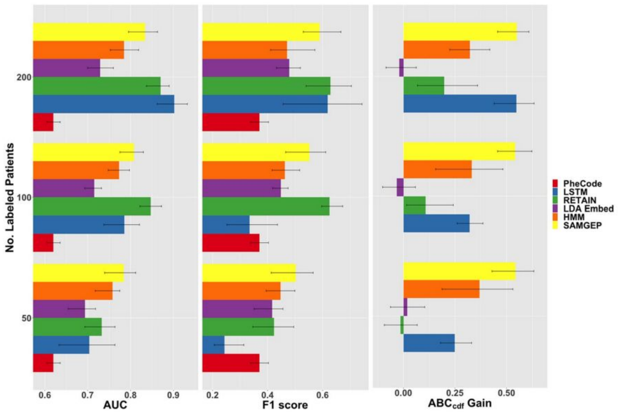
➢ **Real data analyses**

➥ outperformed or worked relatively as well as previous methods

esp. with a small number of labelled phenotype features



MS relapse



HF onset

➥ successfully predicted phenotype events as a process even with $n > 100$

Results: Diagnostics

➤ **Real data analyses**

🔼 Y|H is a stochastic Markov process

`real EHRs data align with this assumption`

🔼X|T follows a Gaussian process

`using tests for normality on a finite collection of patients`

Recap

➢ **No semisupervised methods simultaneously leveraging**

  🔼 longitudinal data

  🔼 *some* gold-standard labels

➢ **SAMGEP uses the few gold-standard labels to**

  🔼 obtain Weight matrix

  🔼 initialize parameters for EM algorithm via supervised learning

➢ **Results show SAMGEP outperforms alternatives for**

  🔼 low n (# of gold-standard labels)

  🔼 correct model specification

Thanks for listening!