Modeling Disease Progression: A Tale of Two Approaches

Sarah Urbut

2024

Outline

1. Problem Setting
2. Data Structure
3. Model Approaches

Problem Setting

Goal

Model individual disease progression over time considering:

- Multiple diseases
- Individual genetic factors
- Time-varying disease risks
- No disease recurrence

Key Variables

- $i \in \{1, \ldots, N\}$: Individuals
- $d \in \{1, \ldots, D\}$: Diseases
- $t \in \{1, \ldots, T\}$: Time points
- $k \in \{1, \ldots, K\}$: Disease topics/patterns

Data Structure

Input Data

- Binary tensor $Y \in \{0,1\}^{N \times D \times T}$
  - $y_{idt} = 1$ if individual i develops disease d at time t
- Genetic covariates $G \in \mathbb{R}^{N \times P}$
  - P genetic variants for each individual

Challenges

- Sparse observations
- Right-censoring
- No disease recurrence constraint
- High dimensionality

Model 1: Tensornoulli
Core Idea

Tensor decomposition with genetic effects
Model Structure

$$\Theta = (U_1(G) \otimes_3 U_2) \odot (W \otimes_3 U_3)$$

where:

- $U_1(G) \in \mathbb{R}^{N \times K \times R_1}$: Individual loadings
- $U_2 \in \mathbb{R}^{T \times R_1}$: Temporal basis for individuals
- $W \in \mathbb{R}^{D \times K \times R_2}$: Disease weights
- $U_3 \in \mathbb{R}^{T \times R_2}$: Temporal basis for diseases

Tensornoulli: Genetic Integration

Genetic Effects

$$U_1(G) = f(GB + C)$$

where:

- $B \in \mathbb{R}^{P \times K \times R_1}$: Genetic effects
- $C \in \mathbb{R}^{N \times K \times R_1}$: Non-genetic effects
- f is an activation function (e.g., ReLU)

Element-wise Probability

$$\pi_{idt} = \text{logistic}(\theta_{idt})$$

Gaussian Process-based topic model
Latent Variables

$$\lambda_{ik}(t) \sim GP(\Gamma_k^\top g_i, \Sigma_k)$$
$$\phi_{kd}(t) \sim GP(\mu_d, \Omega_k)$$

where:

- $\lambda_{ik}(t)$: Topic k's influence on individual i at time t
- $\phi_{kd}(t)$: Relationship between topic k and disease d at time t

Aladynoulli: Probability Structure

Topic Proportions

$$\theta_{ik}(t) = \text{softmax}(\lambda_{ik}(t)) = \frac{\exp(\lambda_{ik}(t))}{\sum_{j=1}^{K} \exp(\lambda_{ij}(t))}$$

Disease Probability

$$\pi_{id}(t) = \sum_{k=1}^{K} \theta_{ik}(t) \cdot \text{sigmoid}(\phi_{kd}(t))$$

$$L_{id} = \left[\prod_{t=1}^{T_{id}-1} (1 - \pi_{idt}) \right] \cdot [\pi_{idT_{id}}]^{I(T_{id} < T)}$$

Aladynoulli For event at time t:

$$\ell_{id} = -\sum_{s=0}^{t-1} \ln(1 - \pi_{id}(s)) - \ln(\pi_{id}(t))$$

For censoring at time t_c:

$$\ell_{id} = -\sum_{s=0}^{t_c} \ln(1 - \pi_{id}(s))$$

Inference Approach

Tensornoulli:

- Gradient descent optimization
- Project updates onto covariate space
- Handle time bases through tensor contractions

Aladynoulli:

- Stochastic Gradient Langevin Dynamics (SGLD)
- Gaussian Process kernel computations
- Explicit handling of temporal dependencies

Model Specifications - Common Structure

Data Components

- Binary tensor $Y \in \{0, 1\}^{N \times D \times T}$
- Genetic covariates $G \in \mathbb{R}^{N \times P}$
- No-recurrence constraint: once $y_{idt} = 1$, future values fixed

Event Times

$$T_{id} = \min\{t : y_{idt} = 1\} \text{ or } T \text{ if event never occurs}$$

Tensornoulli - Full Specification

Core Decomposition

$$\Theta = (U_1(G) \otimes_3 U_2) \odot (W \otimes_3 U_3)$$

Element-wise Formulation

$$\theta_{idt} = \sum_{k=1}^{K} \left[\sum_{r_1=1}^{R_1} U_{1,ik}(G_i) \cdot U_{2,tr_1} \right] \cdot \left[\sum_{r_2=1}^{R_2} W_{dkr_2} \cdot U_{3,tr_2} \right]$$

where

$$U_1(G) = \sum_p G_{ip} \cdot B_{pkr1} + C_{ikr1}$$

Probability

$$\pi_{idt} = \text{logistic}(\theta_{idt})$$

ALAdynoulli - Full Specification

Latent Variables

$$\lambda_{ik}(t) \sim \mathrm{GP}(\Gamma_k^\top g_i, \Sigma_k)$$
$$\phi_{kd}(t) \sim \mathrm{GP}(\mathrm{logit}(\mu_d), \Omega_k)$$

Topic and Disease Probabilities

$$\theta_{ik}(t) = \mathrm{softmax}(\lambda_{ik}(t))$$
$$\pi_{idt} = \sum_{k=1}^{K} \theta_{ik}(t) \cdot \mathrm{sigmoid}(\phi_{kd}(t))$$

Likelihood Functions

Tensornoulli Survival Likelihood

$$L_{id} = \left[\prod_{t=1}^{T_{id}-1} (1 - \pi_{idt}) \right] \cdot [\pi_{idT_{id}}]^{I(T_{id} < T)}$$

ALAdynoulli Survival Likelihood For event at time t:

$$\ell_{id} = -\sum_{s=0}^{t-1} \ln(1 - \pi_{id}(s)) - \ln(\pi_{id}(t))$$

For censoring at time t_c:

$$\ell_{id} = -\sum_{s=0}^{t_c} \ln(1 - \pi_{id}(s))$$

Implementation Comparison

Tensornoulli Implementation

- Fixed rank tensor decomposition
- Individual trajectories:

$$\lambda_{ikt} = \sum_r U1_{ikr}(G) \cdot U2_{tr}$$

- Disease trajectories:

$$\phi_{dkt} = \sum_r W_{dkr} \cdot U3_{tr}$$

- Genetic effects:

$$U1_{ikr}(G) = \sum_p G_{ip} \cdot B_{pkr} + C_{ikr}$$

ALAdynoulli Implementation

- GP-based topic model
- Individual trajectories:

$$\lambda_{ik} \sim \mathrm{GP}(\Gamma_k^\top g_i, K_k)$$

- Disease trajectories:

$$\phi_{kd} \sim \mathrm{GP}(\mu_d, \Omega_k)$$

- Combining:

$$\pi_{id} = \sum_k \theta_{ik} \cdot \mathrm{sigmoid}(\phi_{kd})$$

Technical Trade-offs

Tensornoulli

- Fixed rank representation (R_1, R_2)
- Explicit genetic effects via B_{pkr}
- Shared basis functions (U_2, U_3)
- Direct modeling of log-odds (θ)
- Lower computational complexity

ALAdynoulli

- GP kernels with learned parameters
- Genetic effects via GP mean
- Topic-specific covariance matrices
- Flexible time evolution
- Natural uncertainty quantification

Implementation Comparison

Tensornoulli Implementation

- Fixed rank tensor decomposition
- Individual trajectories:

$$\lambda_{ikt} = \sum_r U1_{ikr}(G) \cdot U2_{tr}$$

- Disease trajectories:

$$\phi_{dkt} = \sum_r W_{dkr} \cdot U3_{tr}$$

- Genetic effects:

$$U1_{ikr}(G) = \sum_p G_{ip} \cdot B_{pkr} + C_{ikr}$$

ALAdynoulli Implementation

- GP-based topic model
- Individual trajectories:

$$\lambda_{ik} \sim \mathrm{GP}(\Gamma_k^\top g_i, K_k)$$

- Disease trajectories:

$$\phi_{kd} \sim \mathrm{GP}(\mu_d, \Omega_k)$$

- Combining:

$$\pi_{id} = \sum_k \theta_{ik} \cdot \mathrm{sigmoid}(\phi_{kd})$$

Technical Trade-offs

Tensornoulli

- Fixed rank representation (R_1, R_2)
- Explicit genetic effects via B_{pkr}
- Shared basis functions (U_2, U_3)
- Direct modeling of log-odds (θ)
- Lower computational complexity

ALAdynoulli

- GP kernels with learned parameters
- Genetic effects via GP mean
- Topic-specific covariance matrices
- Flexible time evolution
- Natural uncertainty quantification

Tensornoulli Solution: TensorClass

Core Components

- Polynomial basis generation for temporal structure:

$$\text{Phi} = \{\text{legendre}_r(2t-1)\}_{r=1}^{R}$$

- Mode-dot products for efficient computation:

$$A2_T = B2 \otimes_2 \text{Phi}_{A2}, \quad A1_T = B1 \otimes_2 \text{Phi}_{A1}$$

- Final theta computation:

$$\theta_{\text{fit}} = \langle A1_T, A2_T \rangle = \text{einsum}('irt, jrt- > ijt', A1_T, A2_T)$$

Tensornoulli: Optimization Strategy

Gradient-based Updates

For each iteration:

1. Update $B2$ with line search:

$$B2_{\text{new}} = B2 - \text{stepsize}_{B2} \cdot \nabla B2$$

   where $\nabla B2 = \text{einsum}('ijt, jrt- > irt', L_{\text{nabla}}, A1_T)$

2. Update $B1$ with covariate projection:

$$B1_{\text{new}} = B1 - \text{stepsize}_{B1} \cdot \nabla B1$$

$$B1_{\text{new}} = B1_{\text{new}} \otimes_0 (X(X^T X)^{-1} X^T)$$

Adaptive Step Sizes

- Backtracking line search with parameters α, β
- Step size reduction when loss doesn't improve

ALAdynoulli Solution: GP-Softmax

Neural Network Architecture

- PyTorch Module with GP priors
- Automatic differentiation for gradients
- Learnable GP kernel parameters:

$$\text{length\_scales}_k \in [T/20, T/2]$$

$$\text{amplitude}_k = \exp(\text{log\_amplitude}_k), \text{log\_amplitude}_k \in [-2, 1]$$

Kernel Construction

$$K_k = \text{amplitude}_k^2 \exp\left(-\frac{1}{2}\frac{(t_i - t_j)^2}{\text{length\_scale}_k^2}\right) + \text{jitter} \cdot I$$

with adaptive jitter for numerical stability

Numerical Considerations

Tensornoulli

- QR decomposition for basis stability
- Line search for convergence
- Explicit tensor contractions
- Covariate space projection
- Fixed rank control

ALAdynoulli

- Adaptive kernel jitter
- Constrained parameter ranges
- Cholesky stability checks
- Gradient clipping
- Condition number monitoring

Next Steps

- Results comparison
  - Prediction accuracy
  - Computational efficiency
  - Model interpretability
- Model advantages/disadvantages
- Computational considerations
- Future directions