Clustering Longitudinal Clinical Marker Trajectories from Electronic Health Data: Applications to Phenotyping and Endotype Discovery

Peter Schulam, Fredrick Wigley, Suchi Saria

Cait Harrigan

ML for health, October 2021

Table of Contents

Disease trajectory subtyping

Disease subtyping is an important clinical task; many diseases have distinct subtypes that may display differential treatment response, progression behaviours, cost of care, etc.

Disease trajectory subtyping

Disease subtyping is an important clinical task; many diseases have distinct subtypes that may display differential treatment response, progression behaviours, cost of care, etc.

Particularly useful for complex, systemic diseases

- Autism
- Cardiovascular disease
- Autoimmune disorders
- Scleroderma (this paper)

Scleroderma

Scleroderma results from an overproduction and accumulation of collagen in body tissues. [1]

[1] Mayo Clinic
[2] American College of Rheumatology

Scleroderma

Scleroderma results from an overproduction and accumulation of collagen in body tissues. [1]

- Rare - affects 75,000 to 100,000 people in the U.S., mostly women between the ages of 30 and 50. [2]

[1] Mayo Clinic

[2] American College of Rheumatology

Scleroderma

Scleroderma results from an overproduction and accumulation of collagen in body tissues. [1]

- Rare - affects 75,000 to 100,000 people in the U.S., mostly women between the ages of 30 and 50. [2]
- Results in hardening of skin, blood vessels

[1] Mayo Clinic

[2] American College of Rheumatology

Scleroderma

Scleroderma results from an overproduction and accumulation of collagen in body tissues. [1]

- Rare - affects 75,000 to 100,000 people in the U.S., mostly women between the ages of 30 and 50. [2]
- Results in hardening of skin, blood vessels
- Immune involvement - common to have autoimmune co-morbidities

[1] Mayo Clinic

[2] American College of Rheumatology

Scleroderma

Scleroderma results from an overproduction and accumulation of collagen in body tissues. [1]

- Rare - affects 75,000 to 100,000 people in the U.S., mostly women between the ages of 30 and 50. [2]
- Results in hardening of skin, blood vessels
- Immune involvement - common to have autoimmune co-morbidities
- Complications in circulation, blood pressure, fibrosis of organ tissue (especially lung, heart)

[1] Mayo Clinic

[2] American College of Rheumatology

Disease trajectory subtyping

EHR data often contains illness severity markers that are routinely collected over the course of patient care.

Disease trajectory subtyping

EHR data often contains illness severity markers that are routinely collected over the course of patient care.

For scleroderma, some examples are:

- Total Skin Score (TSS) - measures fibrosis
- Percent of predicted forced vital capacity (pFVC) - measures restricted lung capacity
- Percent of predicted diffusing capacity (pDLCO) - measures O_2 diffusion into blood
- Right ventricular systolic pressure (RVSP) - measures BP into arteries of the lung

These characterize a *disease activity trajectory*.

Disease trajectory subtyping

EHR data often contains illness severity markers that are routinely collected over the course of patient care.

For scleroderma, some examples are:

- Total Skin Score (TSS) - measures fibrosis
- Percent of predicted forced vital capacity (pFVC) - measures restricted lung capacity
- Percent of predicted diffusing capacity (pDLCO) - measures O_2 diffusion into blood
- Right ventricular systolic pressure (RVSP) - measures BP into arteries of the lung

These characterize a *disease activity trajectory*. Assuming patients that cluster share the same disease subtype, we can infer what the prototypical trajectory looks like for each subtype.

Challenges in clustering longitudinal data

Challenges:

Goals:

Challenges in clustering longitudinal data

Challenges:

- Data is rarely regularly sampled

Goals:

Challenges in clustering longitudinal data

Challenges:

- Data is rarely regularly sampled
- Disease severity is difficult to infer from ICD-9 codes

Goals:

Challenges in clustering longitudinal data

Challenges:

- Data is rarely regularly sampled
- Disease severity is difficult to infer from ICD-9 codes
- Data is collected over the course of years to decades

Goals:

Challenges in clustering longitudinal data

Challenges:

- Data is rarely regularly sampled
- Disease severity is difficult to infer from ICD-9 codes
- Data is collected over the course of years to decades

Goals:

- Make use of time-indexed observations

Challenges in clustering longitudinal data

Challenges:

- Data is rarely regularly sampled
- Disease severity is difficult to infer from ICD-9 codes
- Data is collected over the course of years to decades

Goals:

- Make use of time-indexed observations
- Make use of illness severity markers

Challenges in clustering longitudinal data

Challenges:

- Data is rarely regularly sampled
- Disease severity is difficult to infer from ICD-9 codes
- Data is collected over the course of years to decades

Goals:

- Make use of time-indexed observations
- Make use of illness severity markers
- Account for nuisance variability

Nuisance Variables

nuisance variability[3]*:* a random variable that is fundamental to the probabilistic model, but that is of no particular interest in itself

[3] Wikipedia

Nuisance Variables

nuisance variability[3]*:* a random variable that is fundamental to the probabilistic model, but that is of no particular interest in itself

- Covariate-dependent

[3] Wikipedia

Nuisance Variables

nuisance variability[3]*:* a random variable that is fundamental to the probabilistic model, but that is of no particular interest in itself

- Covariate-dependent
- Individual-specific long-term

[3] Wikipedia

Nuisance Variables

nuisance variability[3]*:* a random variable that is fundamental to the probabilistic model, but that is of no particular interest in itself

- Covariate-dependent
- Individual-specific long-term
- Individual-specific short-term

[3] Wikipedia

Table of Contents

Probabilistic modeling

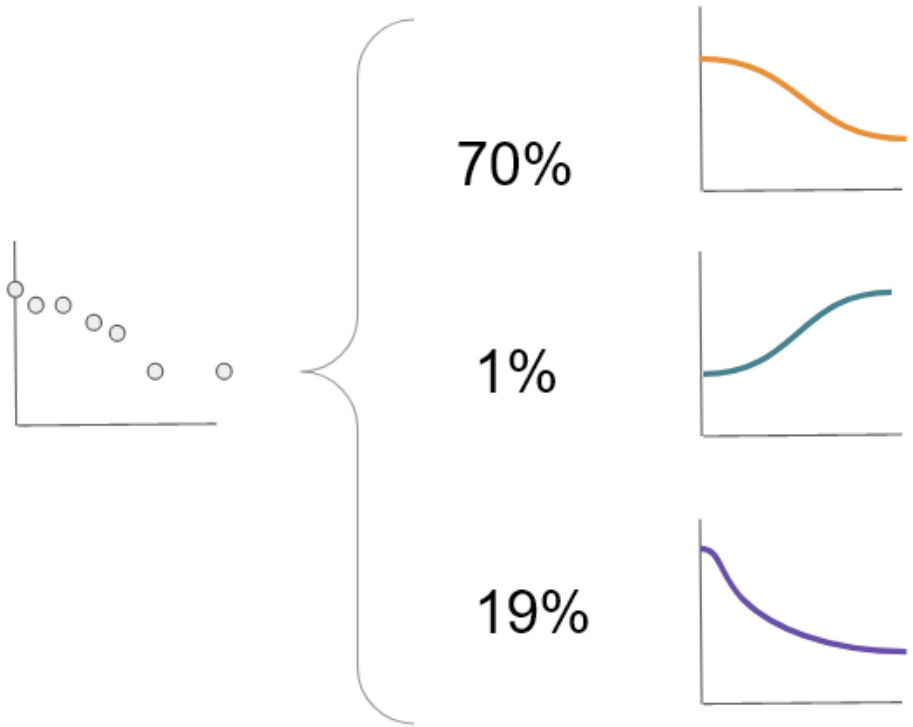
Generative modeling: set up the form of the model, then fit parameters to best explain the data.

Probabilistic modeling

Generative modeling: set up the form of the model, then fit parameters to best explain the data.

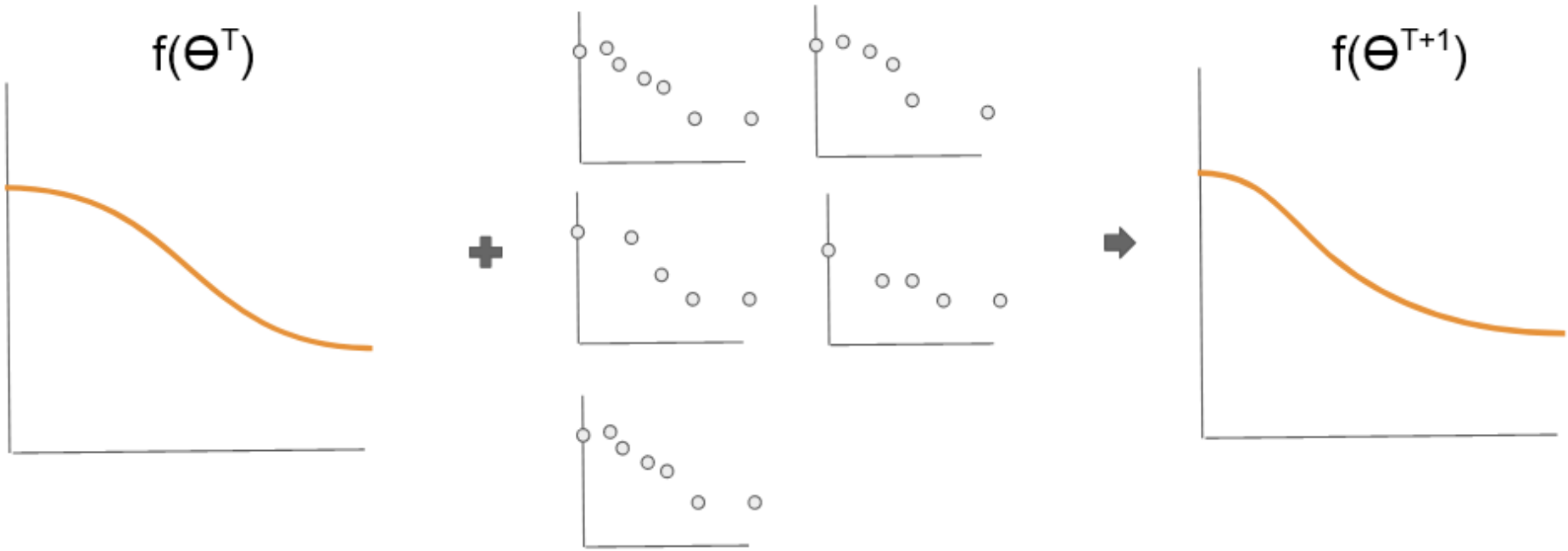
Schulam et al.: use EM to compute MAP estimates of all the parameters.

EM algorithm



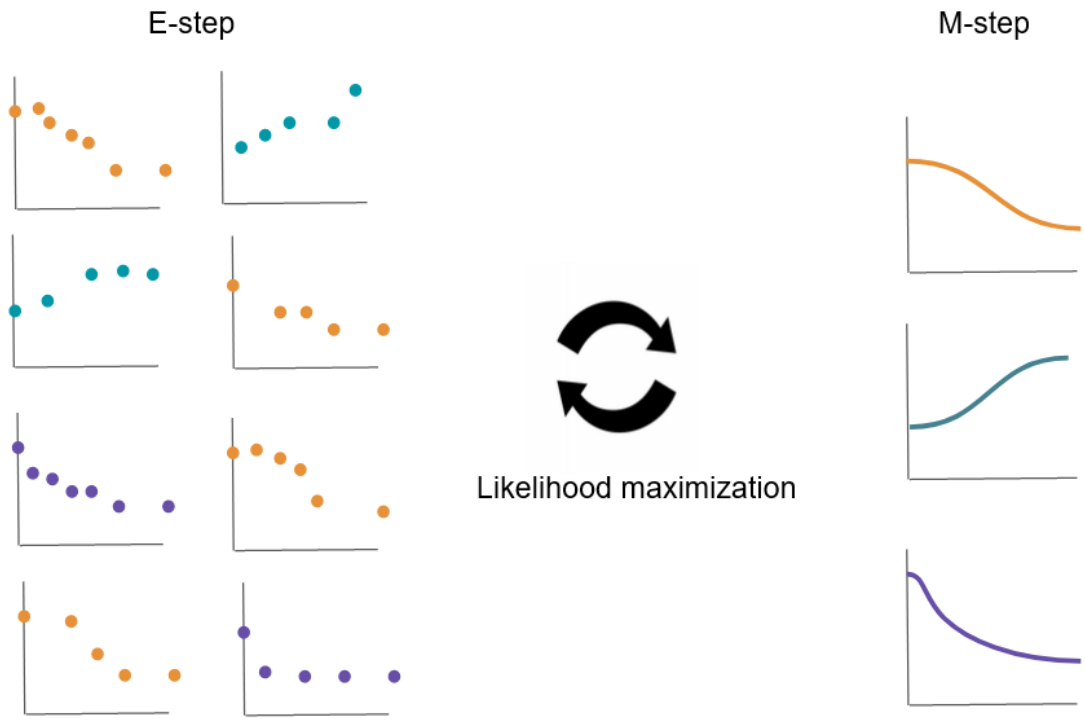
E-step: estimate posterior distribution over z_i for each individual.
$q_i(z_i) = p(z_i|y_i, \beta_{1:G}, \pi_{1:G}, B, t_i, x_i)$

EM algorithm



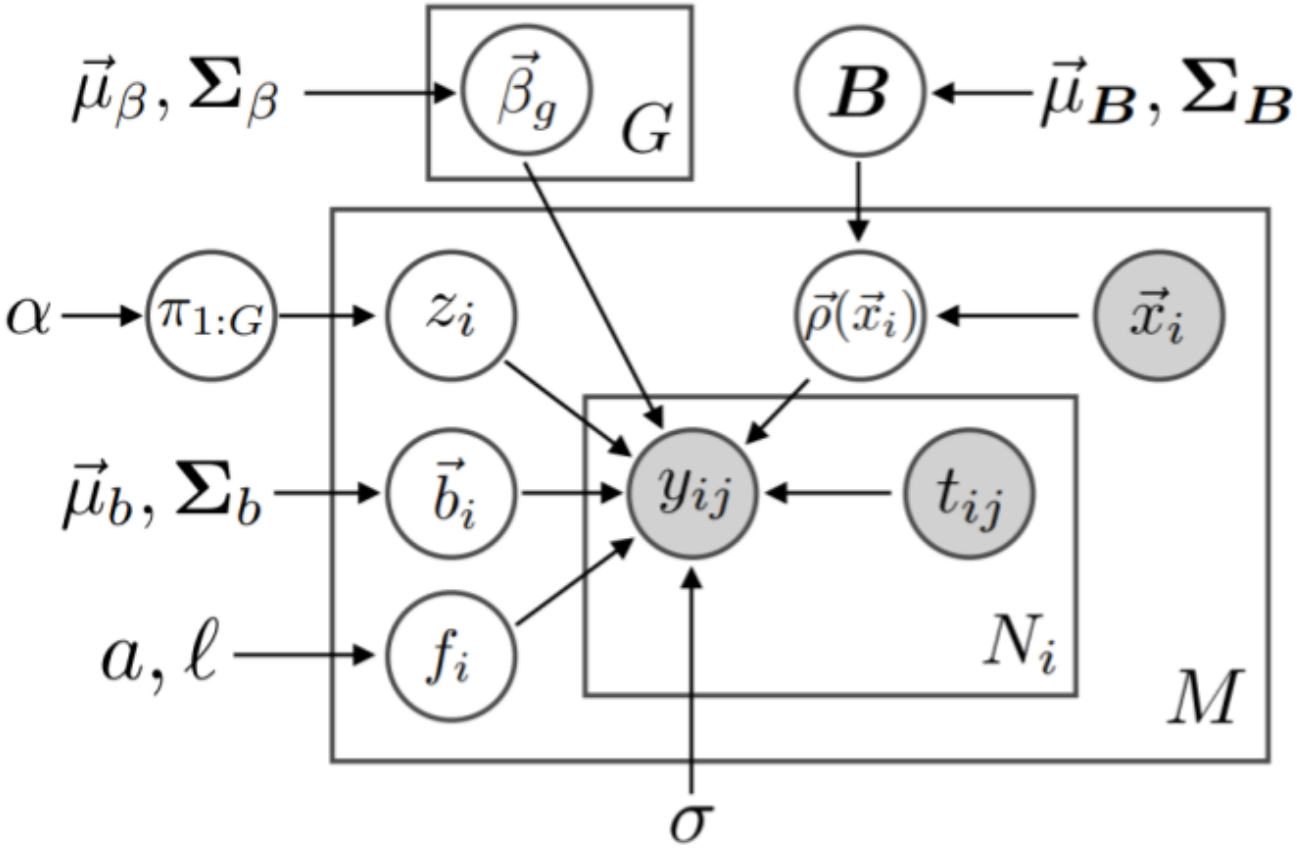
M-step: update parameters via maximum likelihood.
$L_i(\Theta^{\tau+1}|\Theta^{\tau}) = \mathbb{E}_{q_i}[logp(y_i|z_i, \beta_{1:G}^{\tau+1}, B^{\tau+1}, t_i, x_i)]$, where $\Theta = \{\beta_{1:G}, \pi_{1:G}, B\}$

Main Idea



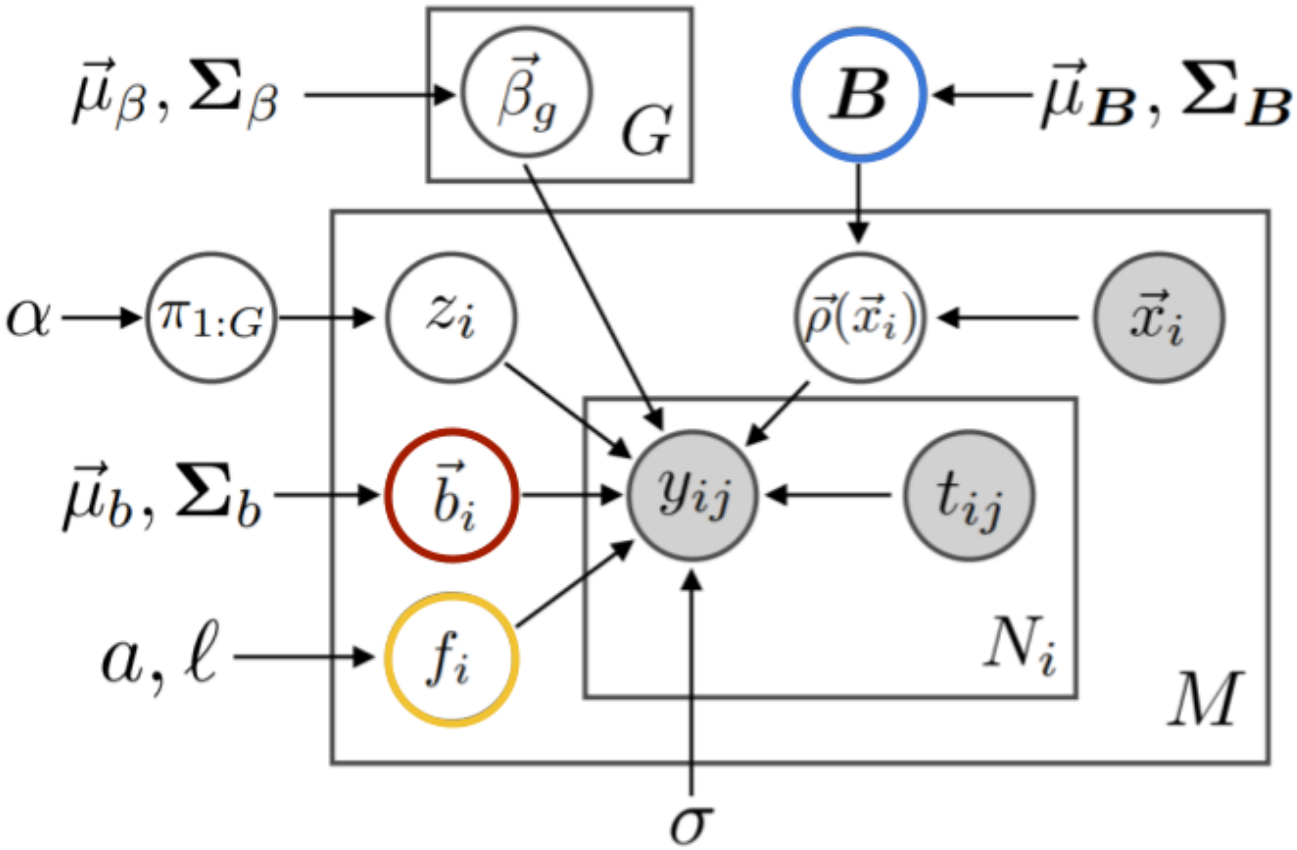
Compute parameter updates st. joint likelihood
$\prod_{i=1}^{M} p(y_i|\beta_{1:G}, \pi_{1:G}, B, t_i, x_i)p(\pi_{1:G}) \prod_{g=1}^{G} p(\beta_g)p(B)$ is maximized

Graphical model



Use BIC to select number of subtypes G

Graphical model

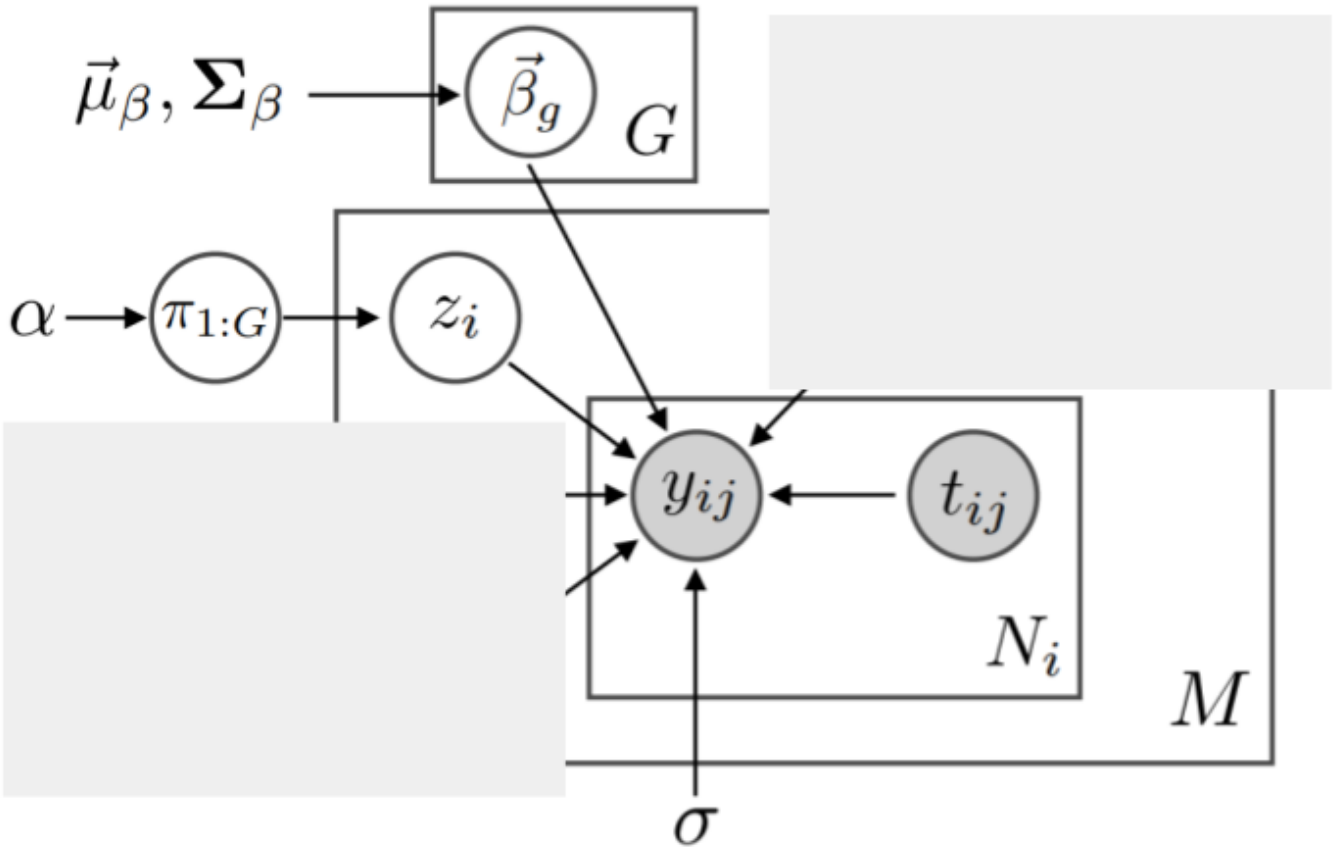


Covariate-dependent nuisance variability

Individual-specific long-term nuisance variability

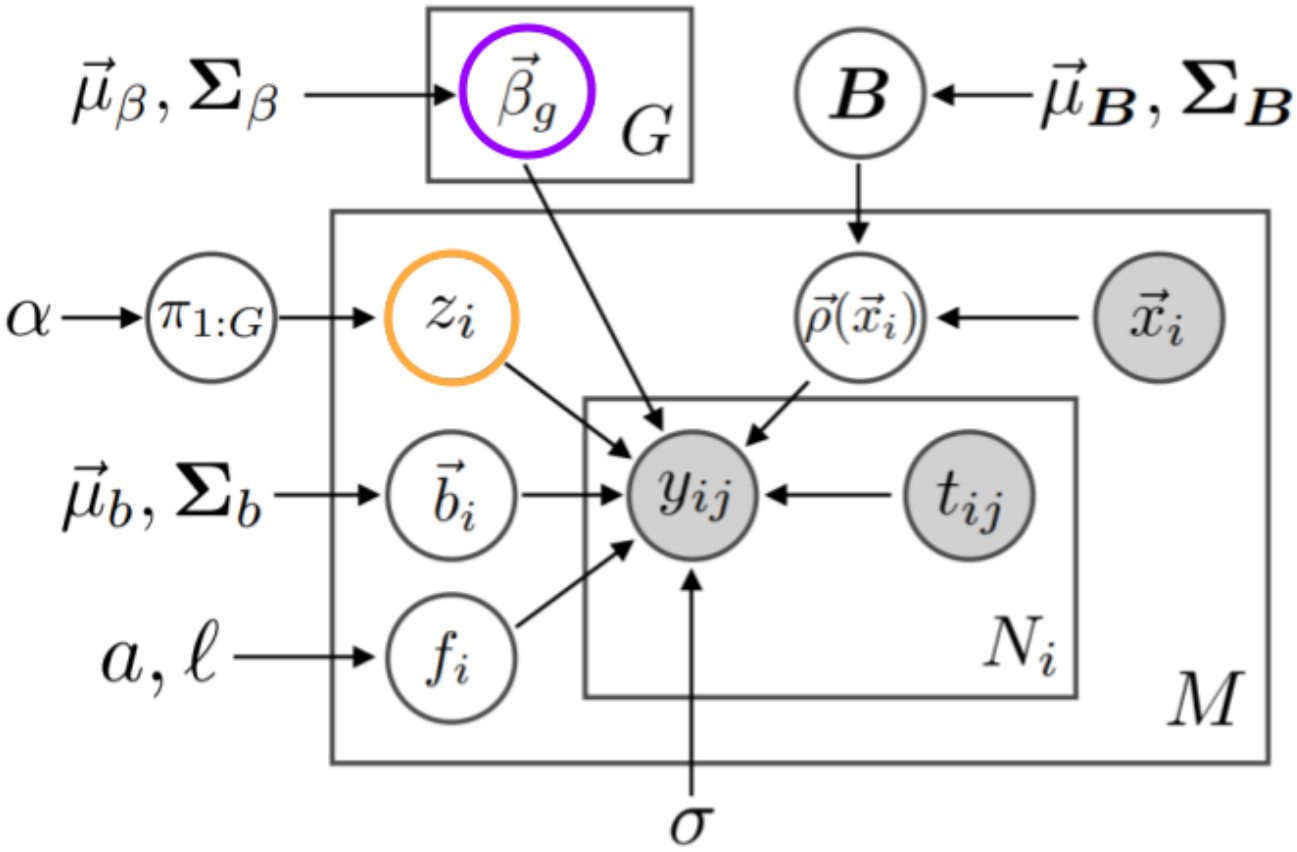
Individual-specific short-term nuisance variability

Graphical model



This looks kind of like a topic model.

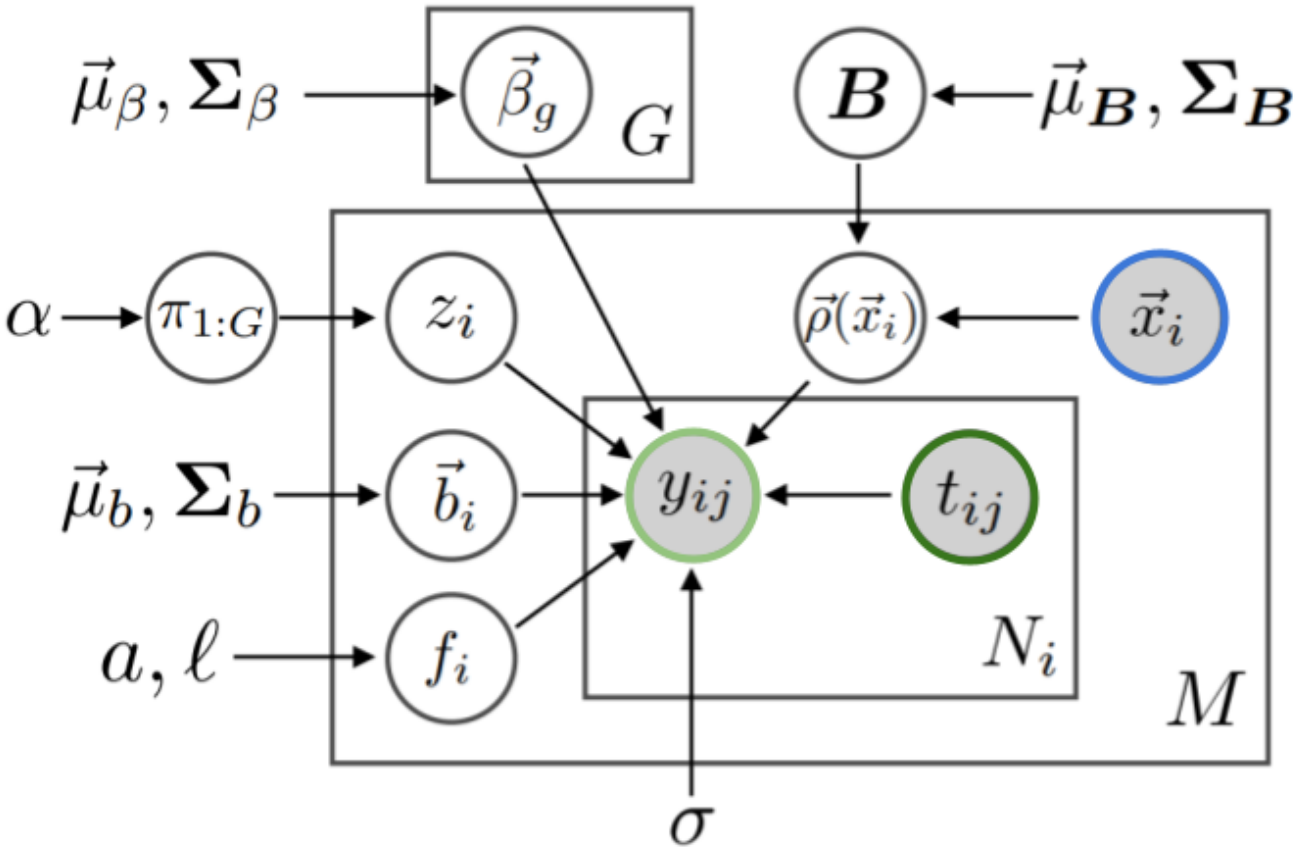
Graphical model



Subtype prototypical trajectory

Subtype membership

Graphical model



Covariates

Measurement time

Measurement value (severity marker)

Table of Contents

Results

- Authors show that it's useful to cluster using all 3 levels of nuisance variable by comparing with restricted subsets of models

Results

S-marker	C+G	C+G+L	PSM
TSS	5.32 ± 0.18	5.41 ± 0.07	$^{*}\mathbf{4.43} \pm 0.14$
pFVC	9.27 ± 0.49	9.34 ± 0.46	$^{*}\mathbf{7.69} \pm 0.39$
pDLCO	15.03 ± 1.82	15.13 ± 1.93	$\mathbf{14.08} \pm 1.77$
RVSP	12.21 ± 0.50	12.11 ± 0.44	$^{*}\mathbf{10.89} \pm 0.27$

Table 1: RMSE with standard errors for s-marker prediction. Bold shows best performance on s-marker; * shows statistical significance ($p \leq 0.05$).

C: covariates, G: group, L: individual long-term effects

Results

- Authors show that it's useful to cluster using all 3 levels of nuisance variable by comparing with restricted subsets of models
- Describe distinct subtypes for scleroderma - some that (A) have a steady, linear progression, (B) decline quickly within the first five years and then stabilize, and (C) are stable for the first five to ten years and then decline rapidly.

Results

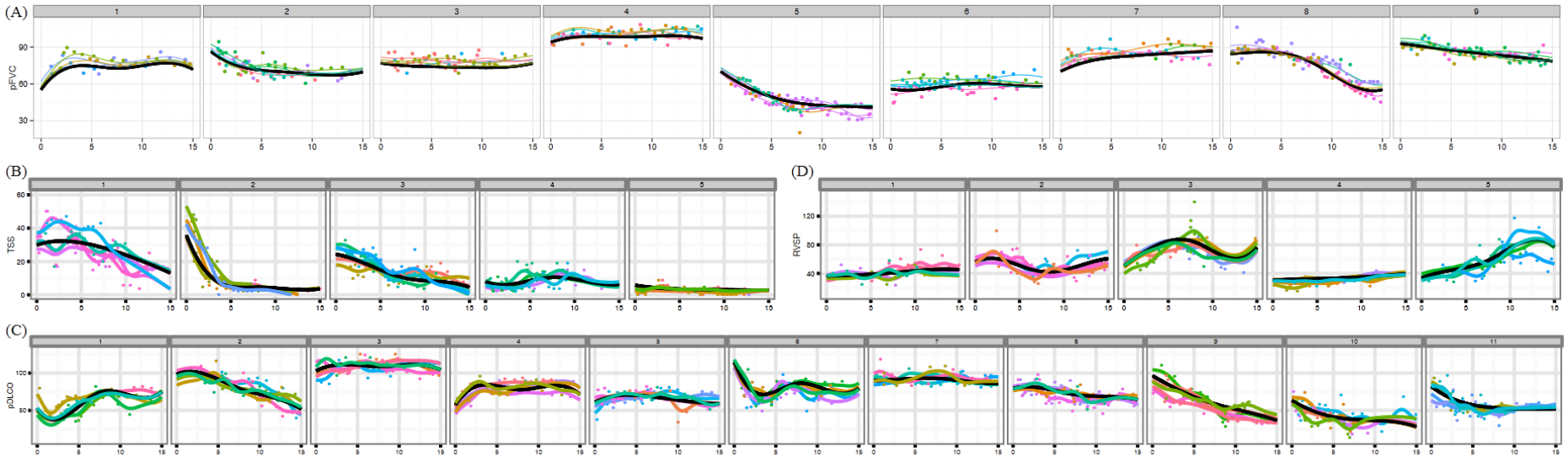


Figure 3: Discovered subtypes for all four s-markers. Panel (A) shows pFVC, panel (B) shows TSS, panel (C) shows pDLCO, and panel (D) shows RVSP. Prototypical s-marker trajectories are shown in black, and individuals sampled from the subtype are shown in color. Colored lines show the individualized s-marker trajectory, and colored points show the observed s-markers.

Implications

- Useful for clinical hypothesis generation
- Patient stratification, treatment decision making implications
- Explicit handling of nuisance variability
- Model is relatively interpretable - Baysean posteriors yield uncertainty estimates

Table of Contents

Limitations

- Little biological interpretation of discovered subtypes
  - You can always find clusters in a clustering task!

Limitations

- Little biological interpretation of discovered subtypes
  - You can always find clusters in a clustering task!
- Don't show that the "nuisance variability" parameters actually capture noise.

Limitations

- Little biological interpretation of discovered subtypes
  - You can always find clusters in a clustering task!
- Don't show that the "nuisance variability" parameters actually capture noise.
- Might have been valuable to interpret population covariates alongside subtyping.
  - Is it possible to map known predispositions to any particular trajectory subtypes?
  - Can you predict subtype membership for a patient early-on?

Conclusions

Interesting to see a topic modelling-eqsue approach for longitudinal data

Conclusions

Interesting to see a topic modelling-eqsue approach for longitudinal data

This was a relatively early paper on doing clustering with EHR timeseries. There is no strict concept of distances between patient trajectories, which has been developed a lot since.