A Bayesian latent class linear mixed model for monotonic processes subject to measurement error

Lizbeth Naranjo Albarrán Ruth Fuentes García

Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico

Osvaldo Espin-Garcia

Department of Public Health Sciences & Department of Statistical Sciences, University of Toronto; Department of Biostatistics, University Health Network, Canada

Biometrics in the Bay of Islands 2023

Abstract

- A latent class linear mixed model is considered with the assumptions:
  - The true process for the disease progression is continuous.
  - The 'true' process is the monotonic process, since the disease progressively worsens, and it is accounted for via truncated normal distributions.
  - The observed responses are subject to measurement errors, since they have both decreasing and increasing patterns.
  - The main purpose is to classify the response trajectories through the latent classes to better describe the disease progression within homogeneous subpopulations.
- Bayesian methods.

Osteoarthritis Initiative (OAI)

- OAI is a cohort study, started in 2002 across multiple centers in the United States, with the objective of pinpointing risk factors and biomarkers for knee OA.
- A total of 4,796 participants were recruited into one of three subcohorts, according to distinct inclusion criteria and followed up for almost 10 years:
  - Progression subcohort (≈29 %): participants with symptomatic radiographic knee OA.
  - Incidence subcohort (≈68 %): participants with a higher risk of developing symptomatic radiographic knee OA based on clinicodemographic factors.
  - Control subcohort (≈3 %): a limited number of participants with no risk factors nor symptomatic radiographic knee OA.
- Rich data collection including biospecimen (e.g., urine, serum, plasma), imaging (x-rays and MRIs), and self-administered questionnaire information performed over time (every year from years 1-4 and every two years thereafter).

Main purpose in radiographic diagnosis of osteoarthritis

- The main interest lies in identifying trajectory groups based on the reported medial minimum joint space width (MCMJSW) measured in millimeters.
- It is subject to measurement error and follows a non-increasing process due to the chronic and progressive nature of OA.
- We focus on a subset of participants in the progression subcohort with complete information across visit times ($n = 505$).
- The response of interest is the total, i.e., the sum, MCMJSW across both left and right knees as a measure of overall knee structural state.
- Covariates of interest: age, biological sex, and time-varying features such as body mass index (BMI) and the maximum across knees of the Western Ontario and McMaster Universities Arthritis Index (WOMAC) total score.

Total MCMJSW

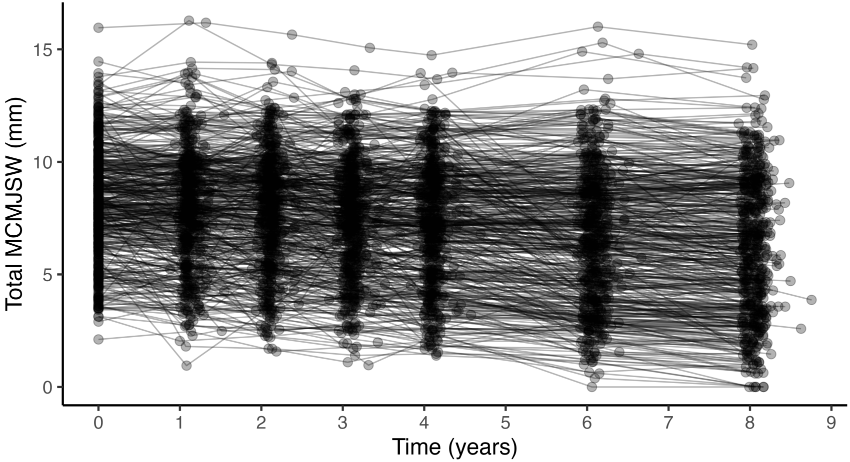


Figura: Spaghetti plots for the subjects' trajectories ($n = 505$) in total (i.e., across both right and left knees) medial minimum joint space width (MCMJSW) measured in millimeters (mm).

WOMAC questionnaire

- The WOMAC questionnaire consists of 24 items divided into 3 subscales capturing pain, stiffness, and physical function.

Tabla: Summary of the response and covariate features across 505 subjects at baseline.

Feature n=505	Mean (sd)	Median (Min,Max)
Total MCMJSW	8.2 (2.3)	8.2 (2.1,16.0)
Age	59.7 (8.8)	59 (45,79)
Body Mass Index (BMI)	29.7 (4.7)	29.4 (18.2,44.2)
WOMAC Total Score (max)	23.5 (16.9)	20.9 (0.0,85.0)
Biological Sex	Female	Male
	270 (53 %)	235 (47 %)

Latent class linear mixed models (LCLMM)

- A latent class linear mixture model (LCLMM) is a mixture of K linear mixed models (LMMs), where each one of the K LMMs corresponds to one latent class.
- The model assumes that the population is heterogeneous and composed of K groups of subjects characterized by K mean trajectory profiles.
- Each subject belongs to only one latent class.

LCLMM Class Membership

- Let c_{ik} be an indicator variable denoting whether subject i belongs to class k,

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ is a member of class } k, \\ 0 & \text{if subject } i \text{ is not a member of class } k, \end{cases}$$

$$\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iK})' \sim \mathrm{Multinomial}\left(1, \boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})'\right)$$

for $i = 1, \ldots, N$ and $k = 1, \ldots, K$.

- The probabilities π_{ik} for each latent class are given by the multinomial logistic model:

$$\pi_{ik} = \mathbb{P}(c_{ik} = 1 | \boldsymbol{v}_i) = \frac{\exp(\boldsymbol{v}_i' \boldsymbol{\alpha}_k)}{\sum_{j=1}^{K} \exp(\boldsymbol{v}_i' \boldsymbol{\alpha}_j)},$$

$\boldsymbol{v}_i$ covariates determining class membership for subject i, $\boldsymbol{\alpha}_k$ is a vector of regression parameters for class k.

LCLMM Linear Predictor

- Let W_{it} be the true, i.e., error-free, response score for the ith subject at time t, which is related to three sets of exogenous covariates throughout the linear predictor η_{it}:

$$\eta_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_{it}\boldsymbol{\gamma}_i + \boldsymbol{u}'_{it}\boldsymbol{\lambda}_{k_i},$$

- $\boldsymbol{x}_{it}$ covariates for the overall fixed effects,
- $\boldsymbol{z}_{it}$ covariates associated with the random effects,
- $\boldsymbol{u}_{it}$ covariates for the class-specific fixed effects.
- β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_i$, and $\boldsymbol{\lambda}_{k_i}$ are regression coefficient vectors.

Monotonic continuous process

- Assume that the response scores follow a monotonically non-increasing continuous process, i.e.,

$$W_{i1} \geq W_{i2} \geq \cdots \geq W_{i,T-1} \geq W_{iT},$$

  W_{it} represents the true gradual process, which could be difficult to score quantitatively and is thus unobservable.
- The response variable W_{it} is treated as a latent variable:

$$\begin{aligned} W_{i1} &\sim \mathrm{N}\left(\eta_{i1}, \tau^2\right), \quad t = 1, \\ W_{it} \middle| W_{i,t-1} = w_{i,t-1} &\sim \mathrm{N}\left(\eta_{it}, \tau^2\right) \mathrm{I}\left[W_{it} \leq w_{i,t-1}\right], \end{aligned}$$

  for $t = 2, \ldots, T$, using truncated normal distributions.
- The continuous process satisfies a first-order Markov chain property.

Measurement error

- Let Y_{it}^{*} be the (observed) continuous response for the ith subject at time t measured with error, which may result in non-monotonic patterns.
- We assume that

$$Y_{it}^{*}|W_{it} = w_{it} \quad \sim \quad \mathrm{N}\left(w_{it}, \sigma_k^2\right). \tag{1}$$

  It denotes a classical additive measurement error model, and ε_{it} is independent of W_{it}.
- The variance σ_k^2 for the latent class k is related to the measurement error, if the data in latent class k does not show measurement error $\sigma_k^2 = 0$. The greater the measurement error for class k, the greater variance σ_k^2 will be.

Prior elicitation

- Normal prior distributions are used for the regression coefficients $\beta_0 \sim \mathrm{N}(0, 100)$, $\boldsymbol{\beta} \sim \mathrm{N}_{M_1}(\mathbf{0}, 100\boldsymbol{I})$, $\gamma_i \sim \mathrm{N}_{M_2}(\mathbf{0}, \boldsymbol{\Gamma})$, $\boldsymbol{\lambda}_k \sim \mathrm{N}_{M_3}(\mathbf{0}, 100\boldsymbol{I})$, $\boldsymbol{\alpha}_k \sim \mathrm{N}_Q(\boldsymbol{a}, \boldsymbol{A})$.
- In order to ensure the identifiability of the parameters in the covariance matrix of the random effects' regression coefficients, $\boldsymbol{\Gamma}$, at least one variance must be set to a constant.
- Inverse Gamma prior distributions are used for the measurement error's variance parameters; that is, $\sigma_k^2 \sim \mathrm{IG}(0.01, 0.01)$.
- In finite mixture models, informative prior distributions should be considered to ensure that observations are assigned to each mixture component, $\frac{1}{\tau^2} \sim \mathrm{Gamma}(2, \kappa_\tau)$, with $\kappa_\tau \sim \mathrm{Gamma}(0.5, 10/R_y)$, R_y is the range.

Nonidentifiability problems and label switching

- In a finite mixture model there are three types of nonidentifiability problems that could arise:
  - (i) invariance of the likelihood under the relabelling of the components, a phenomenon called label switching;
  - (ii) potential overfitting introduced when either one component is empty or two components are equal, which means that there are more components defined than actually needed;
  - (iii) another generic property, e.g. when different parameters describe the same density.
- We use R package *label.switching*, which includes eight relabelling methods. These algorithms are used as a post process to relabel the latent classes in the fitted model.

Exploring posterior distributions

- The likelihood function for the post process:

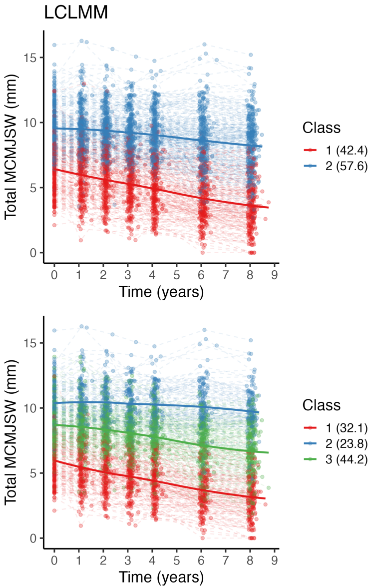
$$
\begin{aligned}
&\mathcal{L}\left(\boldsymbol{w}, \beta_0, \boldsymbol{\beta}, \gamma, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \tau^2, \boldsymbol{\sigma}^2 \middle| \boldsymbol{y}^*, \boldsymbol{c}\right) \\
&= \prod_{i=1}^{n} \left\{ \left[\prod_{t=1}^{T} \mathrm{p}\left(y_{it}^* | w_{it}, \sigma_{k_i}^2\right) \right] \mathrm{p}\left(w_{i1} | \boldsymbol{\lambda}_{k_i}, \omega\right) \right. \\
&\times \left. \left[\prod_{t=2}^{T} \mathrm{p}\left(w_{it} | w_{i,t-1}, \omega, \boldsymbol{\lambda}_{k_i}\right) \right] \right\},
\end{aligned}
$$

$\mathrm{p}(\xi)$ denotes the pdf of the distribution corresponding to ξ.

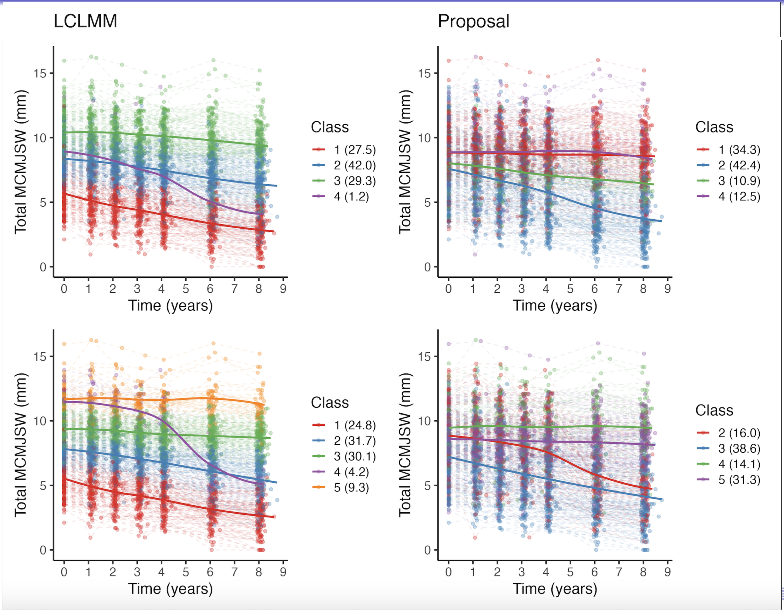
- The joint posterior distribution:

$$
\begin{aligned}
&\mathrm{p}\left(\boldsymbol{w}, \beta_0, \boldsymbol{\beta}, \gamma, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \tau^2, \boldsymbol{\sigma}^2 \middle| \boldsymbol{y}^*, \boldsymbol{c}\right) \\
&\propto \mathcal{L}\left(\boldsymbol{w}, \beta_0, \boldsymbol{\beta}, \gamma, \boldsymbol{\Gamma}, \boldsymbol{\lambda}, \tau^2, \boldsymbol{\sigma}^2 \middle| \boldsymbol{y}^*, \boldsymbol{c}\right) \\
&\times \mathrm{p}(\beta_0)\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\gamma)\mathrm{p}(\boldsymbol{\Gamma})\mathrm{p}(\boldsymbol{\lambda})\mathrm{p}(\tau^2)\mathrm{p}(\boldsymbol{\sigma}^2).
\end{aligned}
$$

Results: Spaghetti plots for the subjects' trajectories



Results: Spaghetti plots for the subjects' trajectories



Results

- The WAIC, and LOO criteria were computed for each model using the *loo* package in R.

Tabla: WAIC, and LOO values for LCLMM (without monotonic constraint) and Proposed (with monotonic constraint) models when the number of classes (K) ranges between 2 and 5 under the ECR-1 method.

	WAIC		**LOO**	
K	LCLMM	Proposal	LCLMM	Proposal
2	9920.87	8199.82	10411.07	8614.59
3	9735.32	6801.50	10222.50	7328.03
4	9770.05	6885.16	10219.89	7345.95
5	9724.57	6672.77	10192.98	7117.77

Conclusions

- Advantages:

  - The monotonic constraint can be used whenever there is a good rationale for sustained increasing/decreasing values in disease outcomes.

  - Consider the measurement error in the answer.

  - It takes into account variability *within* subjects.

  - Initial information about measurement errors is not necessary.

- Disadvantages:

  - Normality (Gaussian) assumptions.

  - Higher computational cost.

References

- Buonaccorsi JP. Measurement Error: Models, Methods and Applications. Boca Raton, Florida: Chapman and Hall/CRC, 2010.
- Carroll RJ, Ruppert D, Stefanski LA et al. Measurement Error in Nonlinear Models: A Modern Perspective. Second ed. Boca Raton, Florida: Chapman and Hall/CRC, 2006.
- Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Am. Statist. Ass.* **96**, 194–209.
- Lester, Gayle. (2012). The Osteoarthritis Initiative: A NIH Public-Private Partnership. *HSS Journal* **8**(1), 62–63. PMID: 23372535.
- McConnell, S, Kolopack, P and Davis, A M. (2001, 10). The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. *Arthritis and rheumatism* **45**, 453–61.
- Papastamoulis, Panagiotis. (2016). label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets* **69**(1), 1–24.

Universidad Nacional Autónoma de México