Machine Learning Techniques for Longitudinal Data

Introduction to Mixed Models

M2 Data Science & Artificial Intelligence

Juliette Chevallier

January 21, 2021

Introduction

1.1 Times Series vs Longitudinal Data Analysis
1.2 Some Reminders About Time Series Analysis
1.3 Linear Regression Models

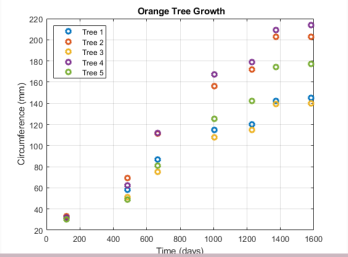
Times Series vs Longitudinal Data Analysis

- Repeated observations of the same variables over time
  $\rightarrow$ $\boldsymbol{d}$ features observed $\boldsymbol{k}$ times

Times Series vs Longitudinal Data Analysis

- Repeated observations of the same variables over time
  $\rightarrow$ $\boldsymbol{d}$ features observed $\boldsymbol{k}$ times

Times Series Analysis	Longidudinal Data Analysis
High k, Low d	Low k, High d

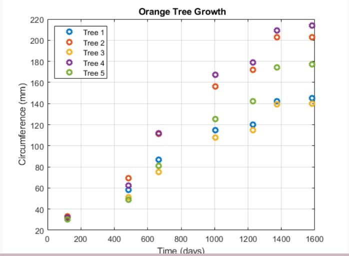


Times Series vs Longitudinal Data Analysis

- Repeated observations of the same variables over time
  $\rightarrow$ $\boldsymbol{d}$ features observed $\boldsymbol{k}$ times

Times Series Analysis	**Longidudinal Data Analysis**
High k, Low d	Low k, High d

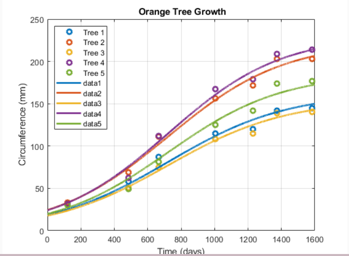
- *Forecasting* future time points;
- Modeling various *cyclical* and *trend* processes;
- Describing temporal dynamics in great *detail*;
- *Specific* interest: unemployment rate, stock market indices, *etc.*



Times Series vs Longitudinal Data Analysis

- Repeated observations of the same variables over time
  $\rightarrow$ $\boldsymbol{d}$ features observed $\boldsymbol{k}$ times

Times Series Analysis	**Longidudinal Data Analysis**
High k, Low d	Low k, High d
• *Forecasting* future time points; • Modeling various *cyclical* and *trend* processes; • Describing temporal dynamics in great *detail*; • *Specific* interest: unemployment rate, stock market indices, *etc.*	• Make inferences about the *population*; • Fairly *general* temporal processes: growth, disease monitoring, *etc.*; • *Variation* in change processes: (early) detection for Alzheimer's disease.



Introduction

1.1 Times Series vs Longitudinal Data Analysis
1.2 **Some Reminders About Time Series Analysis**
1.3 Linear Regression Models

Some Reminders About Time Series Analysis

- **Assumption**: the observed data is a realization of a stochastic process
  $\longrightarrow$ Properties of stochastic processes: Stationarity, ergodicity (Hidden Markov Model, HMM), *etc.*;

Some Reminders About Time Series Analysis

- **Assumption**: the observed data is a realization of a stochastic process
  $\longrightarrow$ Properties of stochastic processes: Stationarity, ergodicity (Hidden Markov Model, HMM), *etc.*;

- **Decomposition**: trend m_t, seasonality s_t and (stochastic) reminder Z_t;

$$Y_t = m_t + s_t + Z_t \quad , \quad \text{where } t \in T \subset \mathbb{Z} \text{ or } \mathbb{N}$$

Some Reminders About Time Series Analysis

- **Assumption**: the observed data is a realization of a stochastic process
  $\longrightarrow$ Properties of stochastic processes: Stationarity, ergodicity (Hidden Markov Model, HMM), *etc.*;

- **Decomposition**: trend m_t, seasonality s_t and (stochastic) reminder Z_t;

$$Y_t = m_t + s_t + Z_t \quad , \quad \text{where } t \in T \subset \mathbb{Z} \text{ or } \mathbb{N}$$

- **Trend**: Long-term variations, Most often *polynomial*:
  - Differentiation to determine the degree,
  - Linear regression for the coefficients;

  Otherwise, more complicated *estimation* procedure;

- **Detrending**: Moving average, exponential smoothing, *Holt-Winters* smoothing, *etc.*;

Some Reminders About Time Series Analysis

- **Assumption**: the observed data is a realization of a stochastic process
  $\longrightarrow$ Properties of stochastic processes: Stationarity, ergodicity (Hidden Markov Model, HMM), *etc.*;

- **Decomposition**: trend m_t, seasonality s_t and (stochastic) reminder Z_t;

$$Y_t = m_t + s_t + Z_t \quad , \quad \text{where } t \in T \subset \mathbb{Z} \text{ or } \mathbb{N}$$

- **Trend**: Long-term variations, Most often *polynomial*:
  - Differentiation to determine the degree,
  - Linear regression for the coefficients;

  Otherwise, more complicated *estimation* procedure;

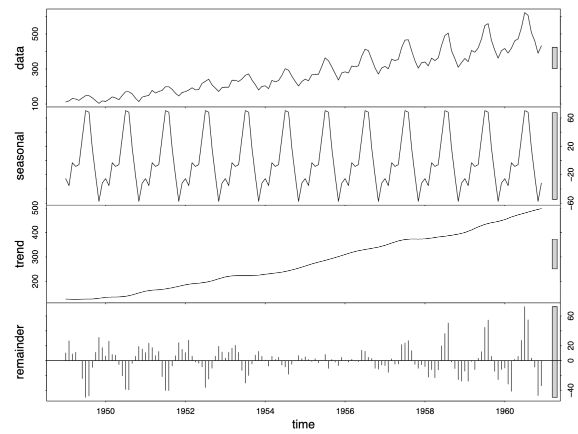
- **Detrending**: Moving average, exponential smoothing, *Holt-Winters* smoothing, *etc.*;

- **Seasonality**: Periodic deterministic function,
  Combination of *sinusoidal* functions, *Indicator* functions;

Some Reminders About Time Series Analysis

$$Y_t = m_t + s_t + Z_t \quad , \quad \text{where } t \in T \subset \mathbb{Z} \text{ or } \mathbb{N}$$

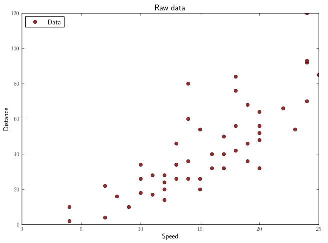
- **Reminder**: Stationary process (Dickey Fuller or KPSS tests), Auto Regressive Moving Average (ARMA) models.



Introduction

1.1 Times Series vs Longitudinal Data Analysis
1.2 Some Reminders About Time Series Analysis
1.3 Linear Regression Models

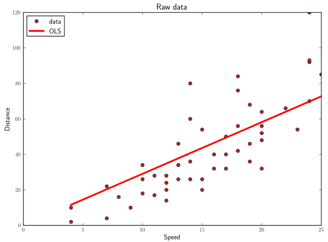
One-Dimensional Least Squares



Braking distance of a car according to its speed

Observations: (t_j, y_j), where $j \in [\![1, k]\!]$;

One-Dimensional Least Squares

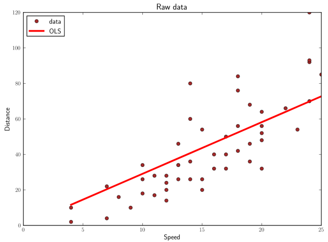


Braking distance of a car according to its speed

Observations: (t_j, y_j), where $j \in [\![1, k]\!]$;

Idea: $y_j \simeq \theta_0^* + \theta_1^* t_j$;

One-Dimensional Least Squares



Braking distance of a car according to its speed

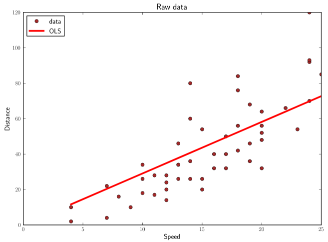
Observations: (t_j, y_j), where $j \in [\![1, k]\!]$;

Idea: $y_j \simeq \theta_0^* + \theta_1^* t_j$;

Probabilistic formulation:

$$y_j = \theta_0^* + \theta_1^* t_j + \varepsilon_j, \quad \text{where} \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2);$$

One-Dimensional Least Squares



Braking distance of a car according to its speed

Observations: (t_j, y_j), where $j \in [\![1, k]\!]$;

Idea: $y_j \simeq \theta_0^* + \theta_1^* t_j$;

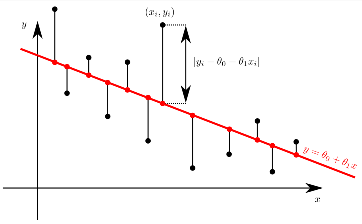
Probabilistic formulation:

$$y_j | \theta_0, \theta_1, \sigma \sim \mathcal{N}(\theta_0^* + \theta_1^* t_j, \sigma^2)$$

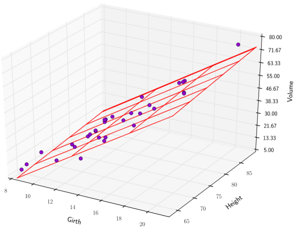
Maximum likelihood estimator:

$$(\hat{\theta}_0, \hat{\theta}_1) \in \underset{(\theta_0, \theta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{j=1}^{k} |y_j - \theta_0 - \theta_1 t_j|^2$$

$\longrightarrow$ Closed form if $(t_j)_j$ non-constant.



Multidimensional Least Squares



Volume of trees according to their height/circumference

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_d \end{pmatrix} \quad A = \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \dots & t_n^d \end{pmatrix}$$

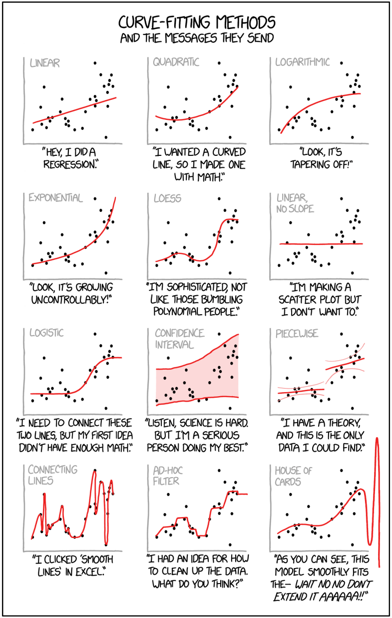
$$\boxed{y \sim \mathcal{N}(A\theta^*, \sigma^2 I_d)}$$

Maximum likelihood estimator: $\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^{d+1}} \|y - A\theta\|_2^2$.

$\longrightarrow$ Closed form if ${}^t AA$ is invertible : $\boxed{\theta^* = ({}^t AA)^{-1}\, {}^t Ay}$

Remark: Go check `http://mfviz.com/hierarchical-models/` for a visual explanation of hierarchical modeling

Linear Regression Assumptions



- **Linearity**...
- **Normality**, especially for confidence intervals and significance tests and small sample size (*cf.* central limit theorem otherwise);
- **Homogeneity of variance** (Homoscedasticity), as above;
- **Independence**: errors in the model is not related to each other/

Remark: Generalized linear model,

$$y \sim q(\theta(t))$$

for the parameter θ and some distribution q

Mixed-Effect Models

2.1 Linear Mixed-Effect Models
2.2 Nonlinear Mixed-Effect Models
2.3 Statistical Inference for Mixed Effects Models

Extend Traditional Linear Models

Real world data:

- complex and messy,
- highly structured,
- may have different grouping factors: populations, species, sites, gender, *etc.*

Basic idea: Two different types of effects:

- *fixed effects* shared by all of the individuals in the population,
- *random effects* specific to each individual.

Observation = Fixed Effect + Random Effect + Error

Linear Mixed Effect Model (LME)

Dataset: Repeated observations of a phenomenon $(t_i, y_i) \in \mathbb{R}^{k_i} \times \mathbb{R}^{k_i}$,
$t_i = (t_{i,j})_{j \in [\![1,k_i]\!]}$, $y_i = (y_{i,j})_{j \in [\![1,k_i]\!]}$, $i \in [\![1,n]\!]$.

Laird and Ware (1982) : $y_i = H_i^\alpha \alpha + H_i^\beta \beta_i + \varepsilon_i$

- $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$, $\Sigma \in \mathcal{S}_{k_i(\mathbb{R})}$,
- For each $i \in [\![1,n]\!]$, $H_i^\alpha \in \mathcal{M}_{k_i,p_\alpha}(\mathbb{R})$ and $H_i^\beta \in \mathcal{M}_{k_i,p_\beta}(\mathbb{R})$,
- Equivalent writing: $y_i \sim \mathcal{N}(H_i^\alpha \alpha + H_i^\beta \beta_i \, , \Sigma)$

The Rats Example

Observations: 30 young rats i, weights $y_{i,j}$ measured weekly for five weeks j.

Individual vs. population **growth**:

Three possible analysis

1. Each rat has its own line, no population-level study

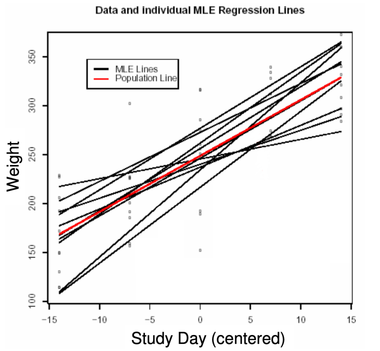
$$y_{i,j} \sim \mathcal{N}\left(a_i t_{i,j} + b_i\,, \sigma^2\right)$$

2. All rats follow the same line, no consideration of individuals

$$y_{i,j} \sim \mathcal{N}\left(\bar{a} t_{i,j} + \bar{b}\,, \sigma^2\right)$$

3. Compromise: Each rat has its own line, but they come from a joint distribution.

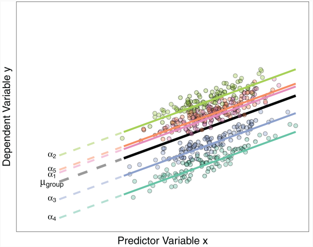
$\longrightarrow$ *Random Intercept and Random Slope Model*



Random Intercept and Random Slope Model

Random Intercept

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left(\bar{a} t_{i,j} + (\bar{b} + b_i), \sigma^2\right) \\ b_i \sim \mathcal{N}(0, \tau^2), \quad \tau \in \mathbb{R}^+ \end{cases}$$



Credit: DOI: 10.7717/peerj.4794/fig-1

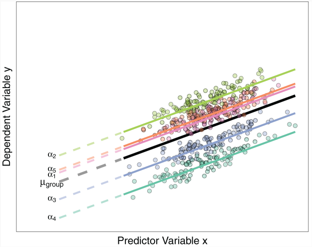
Random Intercept and Random Slope Model

Random Intercept

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left(\bar{a}t_{i,j} + (\bar{b} + b_i), \sigma^2\right) \\ b_i \sim \mathcal{N}(0, \tau^2), \quad \tau \in \mathbb{R}^+ \end{cases}$$

A *hierarchical* model:

- Observation: y,
- Latent variable: b_i,
- Parameters: $\theta = (\bar{a}, \bar{b}, \tau^2, \sigma^2)$,



Credit: DOI: 10.7717/peerj.4794/fig-1

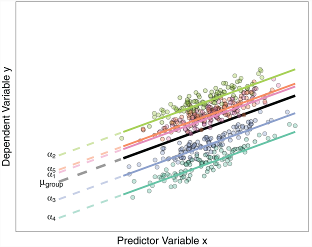
Random Intercept and Random Slope Model

Random Intercept

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left(\bar{a}t_{i,j} + (\bar{b} + b_i)\,, \sigma^2\right) \\ b_i \sim \mathcal{N}(0, \tau^2)\,, \quad \tau \in \mathbb{R}^+ \end{cases}$$

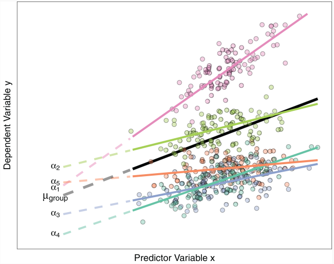
A *hierarchical* model:

- Observation: y,
- Latent variable: b_i,
- Parameters: $\theta = (\bar{a}, \bar{b}, \tau^2, \sigma^2)$,



Random Intercept and Slope

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left((\bar{a} + a_i)t_{i,j} + (\bar{b} + b_i)\,, \sigma^2\right) \\ (a_i, b_i) \sim \mathcal{N}(0, \Sigma)\,, \quad \Sigma \in \mathcal{S}_2(\mathbb{R}) \end{cases}$$



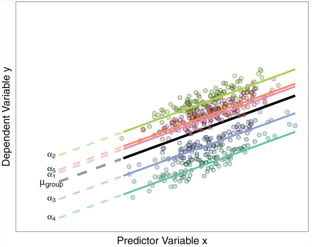
Credit: DOI: 10.7717/peerj.4794/fig-1

Random Intercept and Random Slope Model

Random Intercept	Random Intercept and Slope
$\begin{cases} y_{i,j} \sim \mathcal{N}\left(\bar{a}t_{i,j} + (\bar{b} + b_i)\,, \sigma^2\right) \\ b_i \sim \mathcal{N}(0, \tau^2)\,, \quad \tau \in \mathbb{R}^+ \end{cases}$	$\begin{cases} y_{i,j} \sim \mathcal{N}\left((\bar{a} + a_i)t_{i,j} + (\bar{b} + b_i)\,, \sigma^2\right) \\ (a_i, b_i) \sim \mathcal{N}(0, \Sigma)\,, \quad \Sigma \in \mathcal{S}_2(\mathbb{R}) \end{cases}$

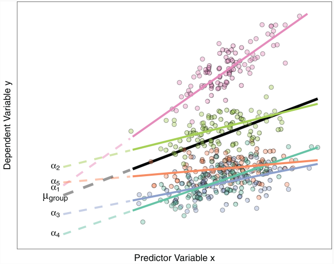
A *hierarchical* model:

- Observation: y,
- Latent variable: b_i,
- Parameters: $\theta = (\bar{a}, \bar{b}, \tau^2, \sigma^2)$,



A *hierarchical* model:

- Observation: y,
- Latent variables: (a_i, b_i),
- Parameters: $\theta = (\bar{a}, \bar{b}, \Sigma, \sigma^2)$,



Credit: DOI: 10.7717/peerj.4794/fig-1

Random Intercept and Random Slope Model

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left((\bar{a} + a_i)t_{i,j} + (\bar{b} + b_i)\,, \sigma^2\right) \\ (a_i, b_i) \sim \mathcal{N}(0, \Sigma) \end{cases}, \qquad \Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{pmatrix} \in \mathcal{S}_2(\mathbb{R})$$

Remark:

$$\begin{cases} \mathcal{V}ar(y_{i,j}) = \Sigma_1^2 + 2\Sigma_{12}\, t_{i,j} + \Sigma_2^2\, t_{i,j}^2 + \sigma^2\,; \\ \mathcal{C}ov(y_{i,j}, y_{i,k}) = \Sigma_1^2 + \Sigma_{12}\,(t_{i,j} + t_{i,k}) + \Sigma_2^2\, t_{i,j}\, t_{i,k}\,; \\ \mathcal{C}ov(y_{i,j}, y_{\ell,k}) = 0\,. \end{cases}$$

- Within person, samples are correlated,
- Between persons samples are uncorrelated,
- Constant correlations within person for random intercept model,
  Complex correlations possible with random slope (*e.g. distant in time*)

Mixed-Effect Models

2.1 Linear Mixed-Effect Models
2.2 Nonlinear Mixed-Effect Models
2.3 Statistical Inference for Mixed Effects Models

Nonlinear Mixed-Effect Models (NLME)

Dataset: Repeated observations of a phenomenon $(t_i, y_i) \in \mathbb{R}^{k_i} \times \mathbb{R}^{k_i}$,
$t_i = (t_{i,j})_{j \in [\![1,k_i]\!]}$, $y_i = (y_{i,j})_{j \in [\![1,k_i]\!]}$, $i \in [\![1,n]\!]$.

Sheiner and Beal (1980), Bates and Watts (1988): $\forall i \in [\![1,n]\!]$, $\forall j \in [\![1,k_i]\!]$,

$$\begin{cases} y_{i,j} &= f\left(z_i; t_{i,j}\right) + \varepsilon_{i,j} \\ z_i &= H_i^\alpha \alpha + H_i^\beta \beta_i \end{cases}$$

- $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$, $\sigma \in \mathbb{R}^+$, $z \in \mathbb{R}^{p_z}$,
- For each $i \in [\![1,n]\!]$, $H_i^\alpha \in \mathcal{M}_{k_i, p_\alpha}(\mathbb{R})$ and $H_i^\beta \in \mathcal{M}_{k_i, p_\beta}(\mathbb{R})$,
- f nonlinear function,
- Equivalent writing: $y_{i,j} \sim \mathcal{N}\left(f(H_i^\alpha \alpha + H_i^\beta \beta_i, t_{i,j}), \sigma^2\right)$.

$\rightsquigarrow$ A multitude of (N)LME models: As many as there are situations to study.

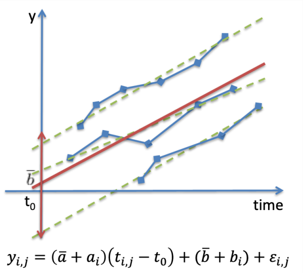
Time alignment in dimension 1

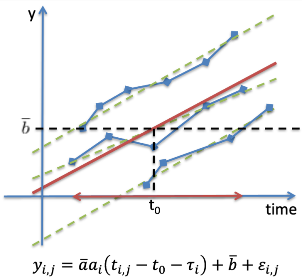
Random Intercept and Random Slope Model:

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left((\bar{a} + a_i)(t_{i,j} - t_0) + (\bar{b} + b_i)\,, \sigma^2\right), & \text{where } t_0 \in \mathbb{R} \text{ reference time,} \\ \theta = (\bar{a}, \bar{b}, \Sigma, \sigma^2) \end{cases}$$

Without obvious reference time: Estimate t_0 as a parameter of the model

$$\begin{cases} y_{i,j} \sim \mathcal{N}\left(\bar{a} a_i(t_{i,j} - t_0 - \tau_i) + \bar{b}\,, \sigma^2\right), & \text{where } t_0 \in \mathbb{R} \text{ reference time,} \\ \theta = (\bar{a}, \bar{b}, t_0, \Sigma, \sigma^2) \end{cases}$$

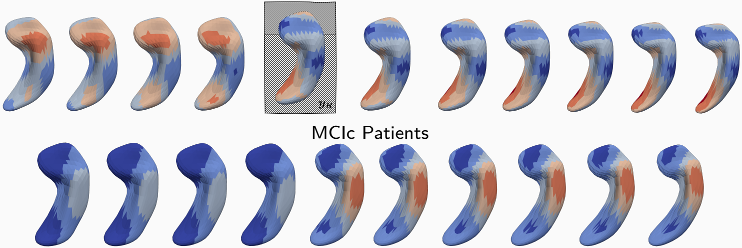




Schiratti et al. (2015)

A Multitude of (N)LME Models

- Model for processing non-scalar data: matrices, anatomical shapes, *etc.*
- Bayesian framework → Prediction, new subject



CN patients

The ADNI data set. Representative shape evolution

Mixed-Effect Models

2.1 Linear Mixed-Effect Models
2.2 Nonlinear Mixed-Effect Models
2.3 Statistical Inference for Mixed Effects Models

The Expectation-Maximization Algorithm

The Expectation-Maximization algorithm

Let $\mathcal{Y} \subset \mathbb{R}^{n_y}$, $\mathcal{Z} \subset \mathbb{R}^{n_z}$ and $\Theta \subset \mathbb{R}^{n_\theta}$.

MLE: Given $y_1^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$,

$$\widehat{\theta}_n^{MLE} \in \operatorname*{argmax}_{\theta \in \Theta} q\big(y_1^n; \theta\big)$$

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta_k) = \int_{\mathcal{Z}} \log q(y, z; \theta)\, q(z|y; \theta_k)\ \mathrm{d}\mu(z)\,;$$

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in Θ:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta \in \Theta} Q(\theta|\theta_k)\,.$$

The Expectation-Maximization Algorithm

The Expectation-Maximization algorithm

Let $\mathcal{Y} \subset \mathbb{R}^{n_y}$, $\mathcal{Z} \subset \mathbb{R}^{n_z}$ and $\Theta \subset \mathbb{R}^{n_\theta}$.

MLE: Given $y_1^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$,

$$\widehat{\theta}_n^{MLE} \in \operatorname*{argmax}_{\theta \in \Theta} q\left(y_1^n; \theta\right)$$

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta_k) = \int_{\mathcal{Z}} \log q(y, z; \theta)\, q(z|y; \theta_k)\; \mathrm{d}\mu(z)\,;$$

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in Θ:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta \in \Theta} Q(\theta|\theta_k)\,.$$

Convergence for curved exponential families

(M1) $\exists S : \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_s}$ Borel function $\mathcal{C}onv(S) \subset \mathcal{S}$, $\int_{\mathcal{Z}} \|S(y,z)\|\, q(z|y;\theta)\, \mathrm{d}\mu(z) \; < \; +\infty$

$$q(y, z; \theta) \;=\; \exp\left(-\psi(\theta) + \langle\; S(y,z) \;|\; \phi(\theta) \;\rangle\right)$$

The Expectation-Maximization Algorithm

The Expectation-Maximization algorithm

Let $\mathcal{Y} \subset \mathbb{R}^{n_y}$, $\mathcal{Z} \subset \mathbb{R}^{n_z}$ and $\Theta \subset \mathbb{R}^{n_\theta}$.

MLE: Given $y_1^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$,

$$\widehat{\theta}_n^{MLE} \in \underset{\theta \in \Theta}{\text{argmax}}\; q\big(y_1^n; \theta\big)$$

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta_k) = \int_{\mathcal{Z}} \log q(y, z; \theta)\, q(z|y; \theta_k)\; \mathrm{d}\mu(z)\,;$$

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in Θ:

$$\theta_{k+1} \in \underset{\theta \in \Theta}{\text{argmax}}\; Q(\theta|\theta_k)\,.$$

Convergence for curved exponential families

(M1) $\exists\, S : \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_s}$ Borel function $\mathcal{C}onv(S) \subset \mathcal{S}$, $\int_{\mathcal{Z}} \|S(y,z)\|\, q(z|y;\theta)\, \mathrm{d}\mu(z) < +\infty$

$$q(y, z; \theta) = \exp\left(-\psi(\theta) + \langle\, S(y,z) \mid \phi(\theta)\, \rangle\right)$$

(M2) $\psi \in \mathcal{C}^2(\Theta, \mathbb{R})$ and $\phi \in \mathcal{C}^2(\Theta, \mathcal{S})$;

(M3) $\theta \mapsto \int_{\mathcal{Z}} S(y,z) q(z|y;\theta)\, \mathrm{d}\mu(z) \in \mathcal{C}^1(\Theta, \mathcal{S})$;

(M4) $\ell \colon \theta \mapsto \int_{\mathcal{Z}} q(y,z;\theta)\, \mathrm{d}\mu(z) \in \mathcal{C}^1(\Theta, \mathbb{R})$ and

$$\partial_\theta \int_{\mathcal{Z}} q(y,z;\theta)\, \mathrm{d}\mu(z) = \int_{\mathcal{Z}} \partial_\theta\, q(y,z;\theta)\, \mathrm{d}\mu(z)\,;$$

(M5) $\exists\, \hat{\theta} \in \mathcal{C}^1(\theta, \mathcal{S})$ s.t.

$$\psi(\hat{\theta}(s)) + \langle s | \phi(\hat{\theta}(s))\rangle \geqslant \psi(\theta) + \langle s | \phi(\theta)\rangle\,.$$

The Expectation-Maximization Algorithm

Convergence EM – Delyon, Lavielle, Moulines (1999)

Assume (M1-5) and that $(\theta_k)_{k\in\mathbb{N}}$ remains in a compact subset. Then, for any initial point,

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0\,,$$

where $\mathcal{L} = \{\, \theta \in \Theta \,|\, \partial_\theta \ell(\theta) = 0 \,\}$.

E-step: Conditional expected log-likelihood

$$Q(\theta|\theta_k) \;=\; \int_{\mathcal{Z}} \log q(y,z;\theta)\, q(z|y;\theta_k)\;\mathrm{d}\mu(z)\,;$$

M-step: Maximize $Q(\,\cdot\,|\theta_k)$ in Θ:

$$\theta_{k+1} \;\in\; \operatorname*{argmax}_{\theta\in\Theta}\, Q(\theta|\theta_k)\,.$$

Convergence for curved exponential families

(M1) $\exists\, S : \mathbb{R}^{n_y}\times\mathbb{R}^{n_z} \to \mathcal{S} \subset \mathbb{R}^{n_s}$ Borel function $\mathcal{C}onv(S) \subset \mathcal{S}$, $\int_{\mathcal{Z}} \|S(y,z)\|\, q(z|y;\theta)\,\mathrm{d}\mu(z) \;<\; +\infty$

$$q(y,z;\theta) \;=\; \exp\left(-\psi(\theta) + \langle\; S(y,z) \;|\; \phi(\theta) \;\rangle\right)$$

(M2) $\psi \in \mathcal{C}^2(\Theta,\mathbb{R})$ and $\phi \in \mathcal{C}^2(\Theta,\mathcal{S})$;

(M3) $\theta \;\mapsto\; \int_{\mathcal{Z}} S(y,z) q(z|y;\theta)\,\mathrm{d}\mu(z) \in \mathcal{C}^1(\Theta,\mathcal{S})$;

(M4) $\ell\colon \theta \;\mapsto\; \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z) \in \mathcal{C}^1(\Theta,\mathbb{R})$ and

$$\partial_\theta \int_{\mathcal{Z}} q(y,z;\theta)\,\mathrm{d}\mu(z) \;=\; \int_{\mathcal{Z}} \partial_\theta\, q(y,z;\theta)\,\mathrm{d}\mu(z)\,;$$

(M5) $\exists\,\hat\theta \in \mathcal{C}^1(\theta,\mathcal{S})$ s.t.
$\psi(\hat\theta(s)) + \langle s|\phi(\hat\theta(s))\rangle \;\geqslant\; \psi(\theta) + \langle s|\phi(\theta)\rangle\,.$

The Expectation-Maximization Algorithm

Convergence EM – Delyon, Lavielle, Moulines (1999)

Assume (M1-5) and that $(\theta_k)_{k\in\mathbb{N}}$ remains in a compact subset. Then, for any initial point,

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0\,,$$

where $\mathcal{L} = \{\,\theta \in \Theta \,|\, \partial_\theta \ell(\theta) = 0\,\}$.

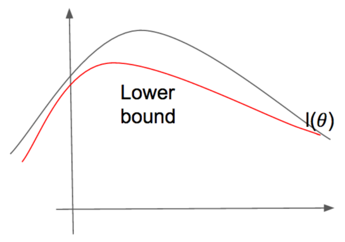
E-step: Conditional expected log-likelihood

$$Q(\theta|\theta_k) = \int_{\mathcal{Z}} \log q(y,z;\theta)\, q(z|y;\theta_k)\, \mathrm{d}\mu(z)\,;$$

M-step: Maximize $Q(\cdot\,|\theta_k)$ in Θ:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q(\theta|\theta_k)\,.$$

Convergence for curved exponential families



Intuition: Jensen inequality

+ Maximize a lower bound at each step

Variants of the EM Algorithm

Speeding-up ← EM

Variants of the EM Algorithm

Speeding-up ← EM

M-step

GEM – Generalized EM

E-step: Compute

$$Q(\theta|\theta_k) = \mathbb{E}\left[\log q(Z|y,\theta_k)\right] \; ;$$

M-step: Find $\theta_{k+1} \in \Theta$ s.t.

$$Q(\theta_{k+1}|\theta_k) \geqslant Q(\theta_k|\theta_k),$$

(Delyon et al., 1999)

See also *Gradient EM* *(Lange, 1995)*

Variants of the EM Algorithm

Speeding-up ← EM → (E-step) *SAEM*

EM → (M-step) GEM; EM → (E-step) SEM; EM → (E-step) MCEM

GEM – Generalized EM

E-step: Compute

$$Q(\theta|\theta_k) = \mathbb{E}\left[\log q(Z|y, \theta_k)\right] ;$$

M-step: Find $\theta_{k+1} \in \Theta$ s.t.

$$Q(\theta_{k+1}|\theta_k) \geqslant Q(\theta_k|\theta_k),$$

(Delyon et al., 1999)

See also *Gradient EM* *(Lange, 1995)*

SEM – Stochastic EM

S-step: Draw an unobserved sample z_k from $q(\,\cdot\,|y;\theta_k)$;

M-step: Maximize Q_{k+1}:

$$\theta_{k+1} \in \underset{\theta\in\Theta}{\text{argmax}}\; Q_{k+1}(\theta) \,.$$

(Celeux and Diebolt, 1985)

MCEM – Monte-Carlo EM

S-step: Draw m samples $z_k^j \sim q(\,\cdot\,|y;\theta_k)$;

E-step: Monte-Carlo estim.

$$Q_k(\theta) = \frac{1}{m}\sum_{j=1}^{m} \log q(y, z_k^j; \theta) ;$$

M-step: Maximize Q_{k+1}.

(Wei and Tanner, 1990)

The Stochastic Approximation EM Algorithm

The SAEM algorithm

- *Idea:* Replace the E-step by a *stochastic approximation*,
- Sequence of positive step-size $(\gamma_k)_{k\in\mathbb{N}}$.

S-step: Draw $z_k \sim q(\,\cdot\,|y;\theta_k)\,;$

SA-step: Update $Q_k(\theta)$ as

$$\begin{aligned} Q_{k+1}(\theta) \;&=\; Q_k(\theta) \\ &\quad + \gamma_k\big(\,\log q(y,z_k;\theta) - Q_k(\theta)\,\big)\,; \end{aligned}$$

M-step: Maximize Q_{k+1} in Θ:

$$\theta_{k+1} \in \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)\,.$$

The Stochastic Approximation EM Algorithm

The SAEM algorithm

- *Idea:* Replace the E-step by a *stochastic approximation*,
- Sequence of positive step-size $(\gamma_k)_{k\in\mathbb{N}}$.

S-step: Draw $z_k \sim q(\,\cdot\,|y;\theta_k)\,;$

SA-step: Update $Q_k(\theta)$ as

$$Q_{k+1}(\theta) \;=\; Q_k(\theta) + \gamma_k\big(\,\log q(y,z_k;\theta) - Q_k(\theta)\,\big)\,;$$

M-step: Maximize Q_{k+1} in Θ:

$$\theta_{k+1} \;\in\; \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)\,.$$

Convergence for curved exponential families

(SAEM1) $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty}\gamma_k = \infty$ and $\sum_{k=1}^{\infty}\gamma_k^2 < \infty\,;$

(SAEM2) $\psi \in \mathcal{C}^{n_s}(\Theta,\mathbb{R})$ and $\phi \in \mathcal{C}^{n_s}(\Theta,\mathcal{S})\,;$

(SAEM3) $\mathbb{E}\big[\phi(Z_{k+1})\big|\mathcal{F}_k\big] = \int_{\mathcal{Z}} \phi(z)q(z|y;\theta_k)\,\mathrm{d}\mu(z)\,;$

(SAEM4) $\int_{\mathcal{Z}} \|S(y,z)\|^2\, q(y,z;\theta)\,\mathrm{d}\mu(z) \;<\; +\infty\,.$

The Stochastic Approximation EM Algorithm

Cvgce SAEM – *Delyon et al. (1999)*

Assume (M1-5), (SAEM1-4) and that $(s_k)_{k\in\mathbb{N}}$ remains in a compact subset. Then, for any initial point,

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0\,,$$

where $\mathcal{L} = \{\, \theta \in \Theta \,|\, \partial_\theta \ell(\theta) = 0 \,\}$.

S-step: Draw $z_k \sim q(\,\cdot\,|y;\theta_k)$;

SA-step: Update $s_k(\theta)$ as

$$s_{k+1}(\theta) \;=\; s_k(\theta) + \gamma_k\big(S(y,z_k) - s_k(\theta)\big)\,;$$

M-step: Maximize Q_{k+1} in Θ:

$$\theta_{k+1} \;\in\; \operatorname*{argmax}_{\theta\in\Theta} Q_{k+1}(\theta)\,.$$

Convergence for curved exponential families

(SAEM1) $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty}\gamma_k = \infty$ and $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$;

(SAEM2) $\psi \in \mathcal{C}^{n_s}(\Theta,\mathbb{R})$ and $\phi \in \mathcal{C}^{n_s}(\Theta,\mathcal{S})$;

(SAEM3) $\mathbb{E}\big[\phi(Z_{k+1})\big|\mathcal{F}_k\big] = \int_{\mathcal{Z}} \phi(z) q(z|y;\theta_k)\,\mathrm{d}\mu(z)$;

(SAEM4) $\int_{\mathcal{Z}} \|S(y,z)\|^2\, q(y,z;\theta)\,\mathrm{d}\mu(z) \;<\; +\infty\,.$

MCMC-SAEM: Monte-Carlo Markov chain procedure in the S-step

(Kuhn and Lavielle, 2004)

(Allassonnière et al., 2010)

Keep Home Message

- Low sample size, many features;
- Highly structured data, grouping factors such as species, gender, *etc.*;
- Two different types of effects: Fixed *vs* random effects;
- Bayesian frameworks allows prediction;
- Estimation performed through the EM algorithm (or its variants).