Computational Statistics

Homework 2

Joaquin Cavieres

Problem 1

EM algorithm

Consider a linear regression model where $\mathbf{y} = (y_i, \ldots, y_n)^T$ and y_i is a mixture of normal distribution such that $y_i \sim \mathcal{N}(x_i^t\beta, \sigma^2)$ and $\mathcal{U}(-a, b)$. Here, x_i is a vector of covariates and β are the parameters associated with these covariates. The pdf for the mixture distribution of y_i is:

$$f(y_i; \boldsymbol{\theta}) = \pi\phi(y_i; x_i^t\beta, \sigma) + (1 - \pi)c,$$

where $\boldsymbol{\theta} = (\beta^T, \sigma, \pi)^T$ and $c = \frac{1}{2a}$

1) Simulate a numeric covariate $x_i \sim \mathcal{U}(-5, 5)$ from a sample of size $n = 200$. Consider initial values for $\beta = (0, 1.5)$, $\sigma = 1.0$, $a = 20$ and $\pi = 0.45$.
2) Calculate the ordinary least squares (OLS) estimates for the regression coefficients, and plot the resulting regression line.
3) Propose an EM algorithm for this problem (*hint*: In the M-step, you will need to solve a weighted least-squares problem, therefore you can use the `lm()` function for this part with the argument `weights`).

Problem 2

Statistical modelling using `glmmTMB` and `R-INLA`

Consider the following information to simulate data and fit a GLMM model:

- n_obs = Number of observations per group
- n_clus = Number of groups
- $\boldsymbol{\theta} = (\beta_0, \beta_1)$
- A numeric covariate x from a Normal distribution with $\mu = 0$ and $\sigma = 1$.
- sigma_clus = standard deviation used to simulate the random effects assumed as $i.i.d$ and cluster $\sim \mathcal{N}(0, \text{sigma\_clus})$.
- λ is the mean of each observation.
- The response variable is assumed $y \sim \text{Poisson}(\lambda)$.

1) Using the library "glmmTMB" fit a GLMM model (including the fixed and random effects). Important: You can use the "glmmTMB" library as commonly you can use a library to fit a GLMM model. It is not necessary to use the TMB template to compile the model!).
2) Using the library "R-INLA" fit a GLMM model (including the fixed and random effects).
3) Compare the results, focusing on the estimated values for the fixed and random effect parameters obtained from each library.
4) Create a comparative plot of the parameter estimates from both methodologies to visualize the differences between the models. For the model fitted in R-INLA show the posterior distribution of the parameters.