Computational Statistics
Latent-class regression

The file `program_effectiveness.txt` contains a data set of 32 observations collected to study whether a new method of teaching economics, the Personalized System of Instruction (PSI), significantly influenced performance in later economics courses (Greene, 2008). Variables in the data set include

- `GPA`: the student's grade point average,
- `PSI`: dummy variable for whether the individual participated in the PSI,
- `TUCE`: the student's score on a pretest in economics.

Let us denote these three variables by W, u and v, respectively. We assume that the students are taken from two groups, and we denote by Z a binary latent variable indicating the group membership. We further assume that the relation between W and u depends on Z, and that v contains some information on Z. More precisely, we postulate the following model:

$$g(w_i|z_i = k, u_i) = \phi(w_i; \beta_{1k} + \beta_{2k}u_i, \sigma_k), \quad k = 0, 1,$$

where $\phi(\cdot; \mu, \sigma)$ denotes the normal density with mean μ and standard deviation σ, and

$$P(Z_i = 1|v_i) = \frac{\exp(\alpha_1 + \alpha_2 v_i)}{1 + \exp(\alpha_1 + \alpha_2 v_i)}.$$

The vector of parameters is $\theta = (\beta_{11}, \beta_{21}, \beta_{10}, \beta_{20}, \alpha_1, \alpha_2, \sigma_1, \sigma_0)$.

1. Write a function that computes the observed-data log-likelihood for this model, and maximize it using the function `optim`.
2. Write an EM algorithm for this problem. Compare the solution obtained with this algorithm to that of the direct approach investigated in the previous question, for different random initial values of the parameters.