Finite Mixture of Linear Regression Models[1]

We observe a variable y_i, where $i = 1, ..., N$ and a vector of covariates, $\mathbf{x}_i$. Let there be also a vector of unknown coefficients $\boldsymbol{\beta}_k$, where $k = 1, ..., K$, that we like to estimate jointly with some variance σ_k^2. If $K = 1$, this model is an ordinary linear regression model. If $K > 1$, then there are K discrete groups of observations which differ in their parameters. The tricky part is that the allocations in these groups is unknown and has to be estimated from the data. The canonical treatment of this Finite Mixture of Linear Regression Model[2] appears e.g. in Frühwirth-Schnatter (2006).

The model can be written as:

$$\begin{aligned} y_i &= \mathbf{x}_i \boldsymbol{\beta}_{k[S_i]} + e_{ik[S_i]} \\ e_{ik} &\sim \mathrm{N}(0, \sigma^2_{k[S_i]}). \end{aligned}$$

I am borrowing the Gelman-Hill notation (Gelman and Hill, 2006) and let an indicator S_i select the corresponding indices k depending on the i^{th} observation.

We can write the joined density of the (y_i, S_i) conditional on the parameters and the data:

$$p(y_i, S_i | \mathbf{x}_i, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \sigma_1^2, ..., \sigma_K^2).$$

We marginalize over S_i to arrive at:

$$\sum_{k=1}^{K} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2) p(S_i = k | \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2).$$

Assuming independence of S_i from the data $(\mathbf{x}_i)$ and the parameters $\boldsymbol{\beta}_k, \sigma_k^2$ we have:

$$\sum_{k=1}^{K} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2) p(S_i = k).$$

Assuming independence across observations and defining $\eta_k = p(S_i = k)$ the likelihood function of this model is as follow:

[1] Notes by Moritz Marbach. Comments welcome.

[2] Other common names: switching regression models in economics, latent class regression models in marketing and mixture-of-expert models in machine-learning literature and as mixed models in biology Frühwirth-Schnatter (2006, p. 246)

$$\mathcal{L} = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \Phi(y_i | \mathbf{x}_i \boldsymbol{\beta}_k, \sigma_k^2) \eta_k \right),$$

where η_k are usually called the 'weights' of the mixture distribution. The larger η_k is the higher the weight, the higher the contribution of the component's density to the likelihood.

The prior densities are commonly chosen to be:

$$\begin{aligned} \boldsymbol{\beta}_k &\sim \mathbf{N}(\mathbf{b}_0, \mathbf{B}_0) \\ \sigma_k^2 &\sim \text{invGamma}(e_0/2, f_0/2) \\ \boldsymbol{\eta} &\sim \mathbf{Dirichlet}(n_0, ..., n_0) \end{aligned}$$

Finite mixture of multivariate mixtures of normals are (generic) identifiable (Frühwirth-Schnatter, 2006, 21 and references therein). However, identifiability does not follow directly for mixtures of regression models. For Finite mixture of linear regression models are identified if there are at least $K(B-1)+1$ distinct rows in the design matrix (where B is the number of covariates that vary across K components) (Frühwirth-Schnatter, 2006, 246).

A Gibbs Sampler can be found for example in (Frühwirth-Schnatter, 2006, p. 75, 253), which I adapted in notation.

At each iteration m do:

1. Draw η:

$$\boldsymbol{\eta}^{(m)} \sim \mathbf{Dirichlet}(n_1, ..., n_k, ..., n_K)$$
$$n_k = \sum_{i=1}^{N} I(S_i^{(m-1)} = k) + n_0,$$

where $I(.)$ is an indicator function taking the value 1 if the condition inside it holds, and zero otherwise.

2. For each $k = 1, ..., K$ draw $\beta_{\mathbf{k}}$:

$$\boldsymbol{\beta}_k^{(m)} \sim \mathbf{N}(\mathbf{b}_{1k}, \mathbf{B}_{1k})$$
$$\mathbf{b}_{1k} = \mathbf{B}_{1k}\left((1/\sigma_k^{2(m-1)})\mathbf{X}_k^{'(m-1)}\mathbf{y}_k^{(m-1)} + \mathbf{B}_0^{-1}\mathbf{b}_0\right)$$
$$\mathbf{B}_{1k} = \left((1/\sigma_k^{2(m-1)})\mathbf{X}_k^{'(m-1)}\mathbf{X}_k^{(m-1)} + \mathbf{B}_0^{-1}\right)^{-1}$$

3. For each $k = 1, ..., K$ draw σ_k:

$$\sigma_k^{2(m)} \sim \text{invGamma}(e_1/2, f_1/2)$$
$$e_1 = e_0 + \sum_{i=1}^{N} I(S_i^{(m-1)} = k)$$
$$f_1 = f_0 + (\mathbf{y}_k^{(m-1)} - \mathbf{X}_k^{(m-1)}\boldsymbol{\beta}_k^{(m)})'(\mathbf{y}_k^{(m-1)} - \mathbf{X}_k^{(m-1)}\boldsymbol{\beta}_k^{(m)}),$$

4. For each $i = 1, ..., N$, draw S_i:

$$S_i^{(m)} \sim \mathbf{Multinomial}(p_0, ..., p_k, ..., p_k)$$
$$p_k = \eta_k^{(m-1)}\Phi(y_i|\mathbf{x}_i\boldsymbol{\beta}_k^{(m)}, \sigma_k^{2(m)})$$

5. For each $k = 1, ..., K$ and using $\mathbf{S}^{(m)}$ construct $\mathbf{X}_k$ and $\mathbf{y}_k$ (the design matrix and dependent variable for all observation where $S_i^{(m)} = k$).

Repeat M times until convergence.

Extensions of this model, might model the weights as function of additional data $p(S_i = k|\mathbf{z}_i)$ (Frühwirth-Schnatter, 2006, p. 274).

Suppose, that some of the coefficients are assumed to be the same across mixture components, such as $\beta_{lk} = \beta_{l1} = ... = \beta_{lK}$ for at least some l. We define the model as in Frühwirth-Schnatter (2006, p. 257):

$$
\begin{aligned}
y_i &= \mathbf{x}_i^f \boldsymbol{\delta} + \mathbf{x}_i^r \boldsymbol{\beta}_{k[S_i]} + e_{ik[S_i]} \\
e_{ik} &\sim \mathrm{N}(0, \sigma^2_{k[S_i]}).
\end{aligned}
$$

To derive the Gibbs sampler define a vector that collects all regression coefficients:

$$\boldsymbol{\alpha} = (\boldsymbol{\delta}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K),$$

and dummy variable D_{ik} that takes a 1 if $S_i = k$ and 0 otherwise. Finally, define

$$\mathbf{z}_i = (\mathbf{x}_i^f, \mathbf{x}_i^r D_{i1}, ..., \mathbf{x}_i^r D_{iK}).$$

Note, that $\mathbf{z}_i$ is a vector that includes the variables with a common effect, and K times the variables with a differential effect over the K groups. Since only one D_{ik} can take the value 1, all other terms in this vector are going to be zero, expect for the first term and one of the K terms. The collection of these vectors, is the matrix $\mathbf{Z}$.

The priors are similar as above, but:

$$
\begin{aligned}
\boldsymbol{\alpha} &\sim \mathbf{N}(\mathbf{a}_0, \mathbf{A}_0) \\
\sigma_k^2 &\sim \mathrm{invGamma}(e_0/2, f_0/2) \\
\boldsymbol{\eta} &\sim \mathbf{Dirichlet}(n_0, ..., n_0)
\end{aligned}
$$

At each iteration m do:

1. Draw η:

$$\boldsymbol{\eta}^{(m)} \sim \mathbf{Dirichlet}(n_1, ..., n_k, ..., n_K)$$
$$n_k = \sum_{i=1}^{N} I(S_i^{(m-1)} = k) + n_0,$$

where $I(.)$ is an indicator function taking the value 1 if the condition inside it holds, and zero otherwise.

2. Draw α:

$$\boldsymbol{\alpha}^{(m)} \sim \mathbf{N}(\mathbf{a}_1, \mathbf{A}_1)$$
$$\mathbf{a}_1 = \mathbf{A}_1 \left(\sum_{i=1}^{N} \left((1/\sigma_{k[S_i]}^{2(m-1)}) \mathbf{z}_i^{'(m-1)} y_i \right) + \mathbf{A}_0^{-1} \mathbf{a}_0 \right)$$
$$\mathbf{A}_1 = \left(\sum_{i=1}^{N} \left((1/\sigma_{k[S_i]}^{2(m-1)}) \mathbf{z}_i^{'(m-1)} \mathbf{z}_i^{(m-1)} \right) + \mathbf{A}_0^{-1} \right)^{-1}$$

3. For each $k = 1, ..., K$ draw σ_k:

$$\sigma_k^{2(m)} \sim \text{invGamma}(e_1/2, f_1/2)$$
$$e_1 = e_0 + \sum_{i=1}^{N} I(S_i^{(m-1)} = k)$$
$$f_1 = f_0 + \mathbf{w}'\mathbf{w}$$
$$\mathbf{w} = \mathbf{y}_k^{(m-1)} - \mathbf{X}_k^{f(m-1)} \boldsymbol{\delta}^{(m)} - \mathbf{X}_k^{r(m-1)} \boldsymbol{\beta}_k^{(m)},$$

4. For each $i = 1, ..., N$, draw S_i:

$$S_i^{(m)} \sim \mathbf{Multinomial}(p_0, ..., p_k, ..., p_k)$$
$$p_k = \eta_k^{(m-1)} \Phi(y_i | \mathbf{x}_k^{f(m-1)} \boldsymbol{\delta}^{(m)} + \mathbf{x}_k^{r(m-1)} \boldsymbol{\beta}_k^{(m)}, \sigma_k^{2(m)})$$

5. For each $k = 1, ..., K$ and using $\mathbf{S}^{(m)}$ construct $\mathbf{X}_k^f$, $\mathbf{X}_k^r$, $\mathbf{Z}$ and $\mathbf{y}_k$ (the design matrix and dependent variable for all observation where $S_i^{(m)} = k$).

Repeat M times until convergence.

Frühwirth-Schnatter (2006) notes, that for large K and many regressors, the algorithm can be time consuming. See Frühwirth-Schnatter (2006, p. 259) for alternatives.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Heidelberg: Springer.

Gelman, A. and J. Hill (2006). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.