Mixture of regressions with automatic variable selection

Lukasz T. Gatarek

Contents

1 Mixture of regression **2**

1 Mixture of regression

Any cloud of data can be modeled by finite mixture of regressions. Despite the fact that the regressions are linear, their composition allows for representing any nonlinear behavior, which is linear in selected part of the domain, or for selecteed subsets of observations.

The model setup follows a mixture of G Normal components

$$y_i|x_i, \{\beta_g, \sigma_g^2, \pi_g\}_{g=1}^G \sim \sum_{g=1}^{G} \pi_g N(x_i\beta_g, \sigma_g^2), \quad \sum_{g=1}^{G} \pi_g = 1, \quad i = 1, \ldots, n, \tag{1}$$

where M explanatory variables are considered

$$x_i = [x_{i,1}\ x_{i,2}\ \ldots x_{i,M}] \tag{2}$$

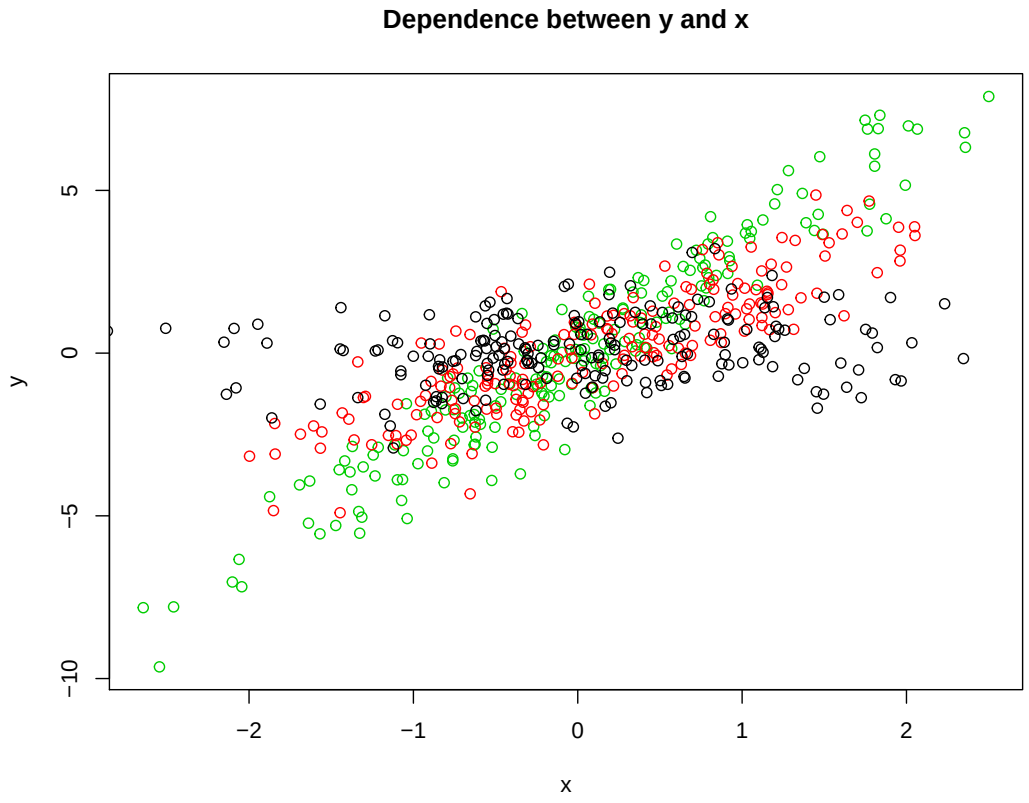
together with the corresponding vector of parameters

$$\beta_g = [\beta_{g,1}\ \beta_{g,2}\ \ldots \beta_{g,M}]. \tag{3}$$

Parameters β can vary freely across different explanatory variables $j = 1, \ldots, M$ and components $g = 1, \ldots, G$.

The model allows each Normal component to possess its own variance σ_g and mean, controlled by the regression parameters β_g.

Figure 1: Data generated from mixture of regressions with 3 components, each with 1 explanaotry variable



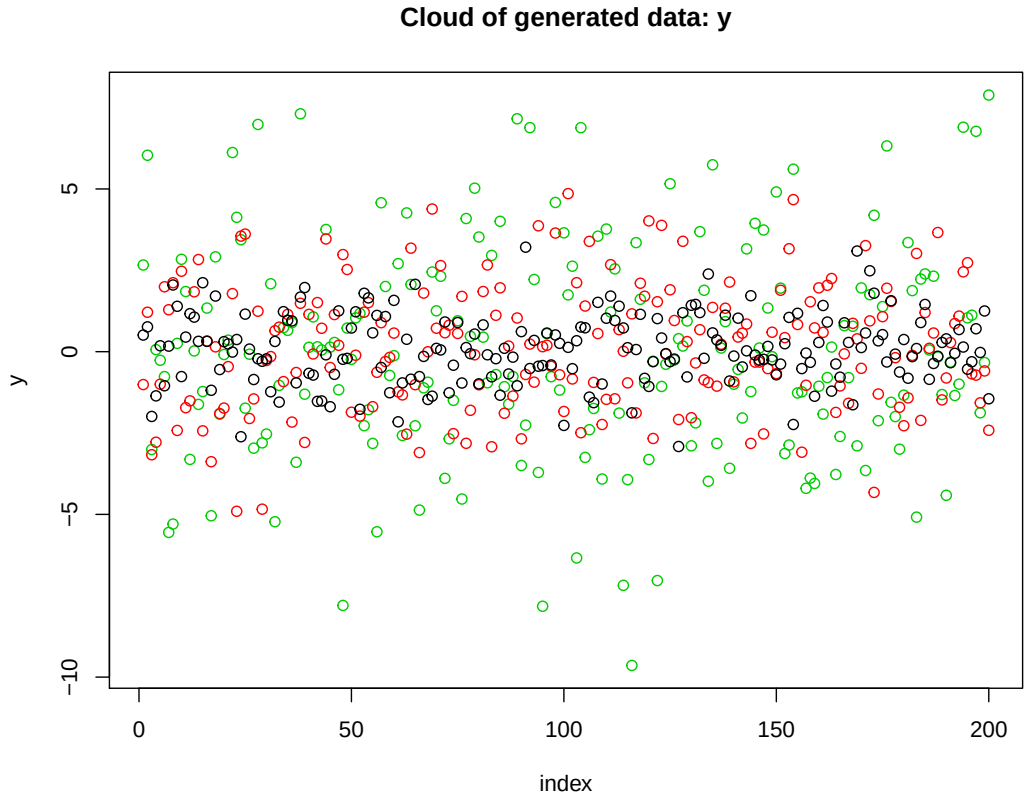
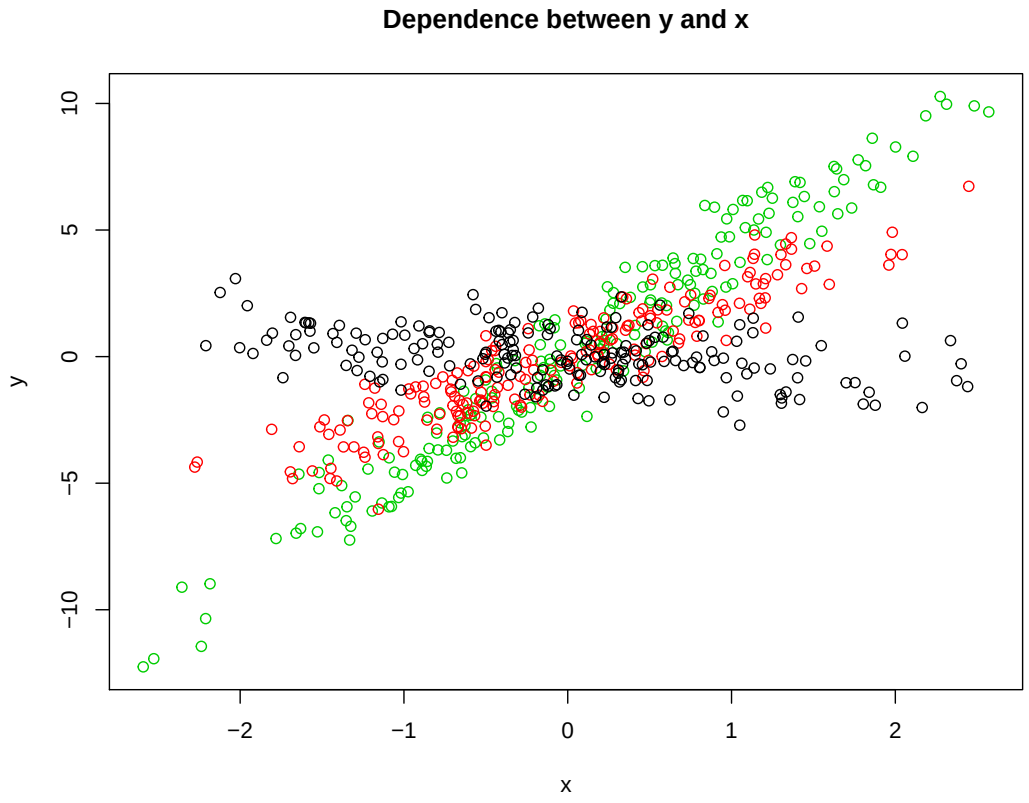


Figure 2: Data generated from mixture of regressions with 3 components, each with 1 explanaotry variable



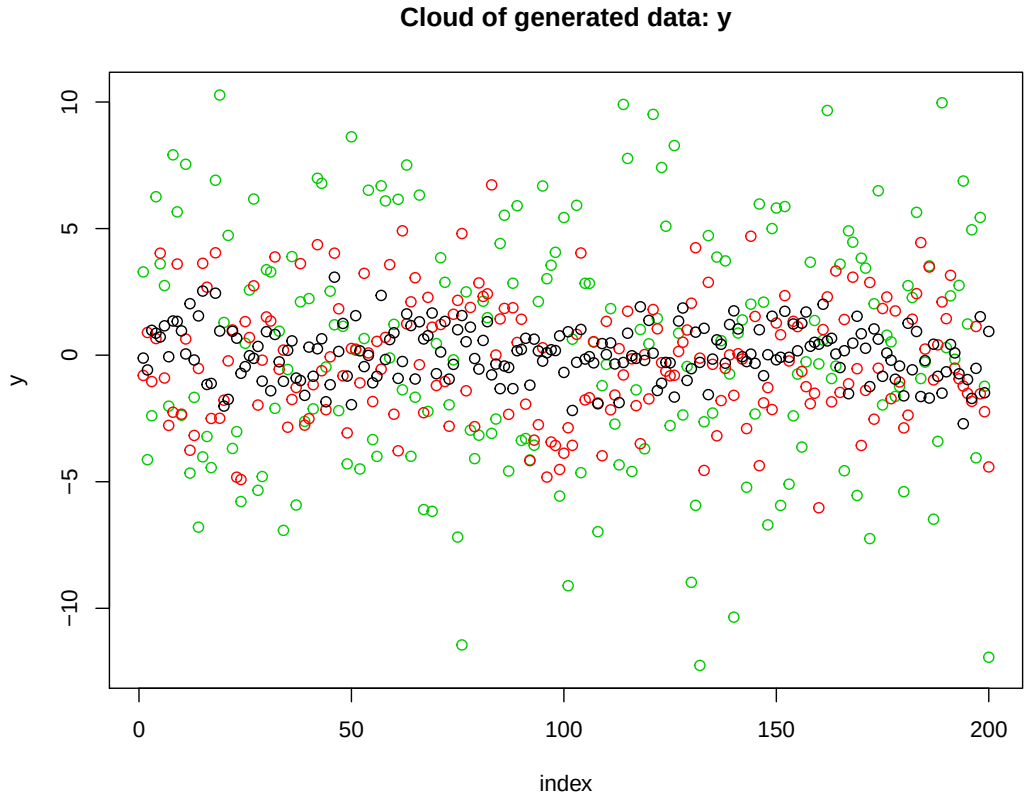
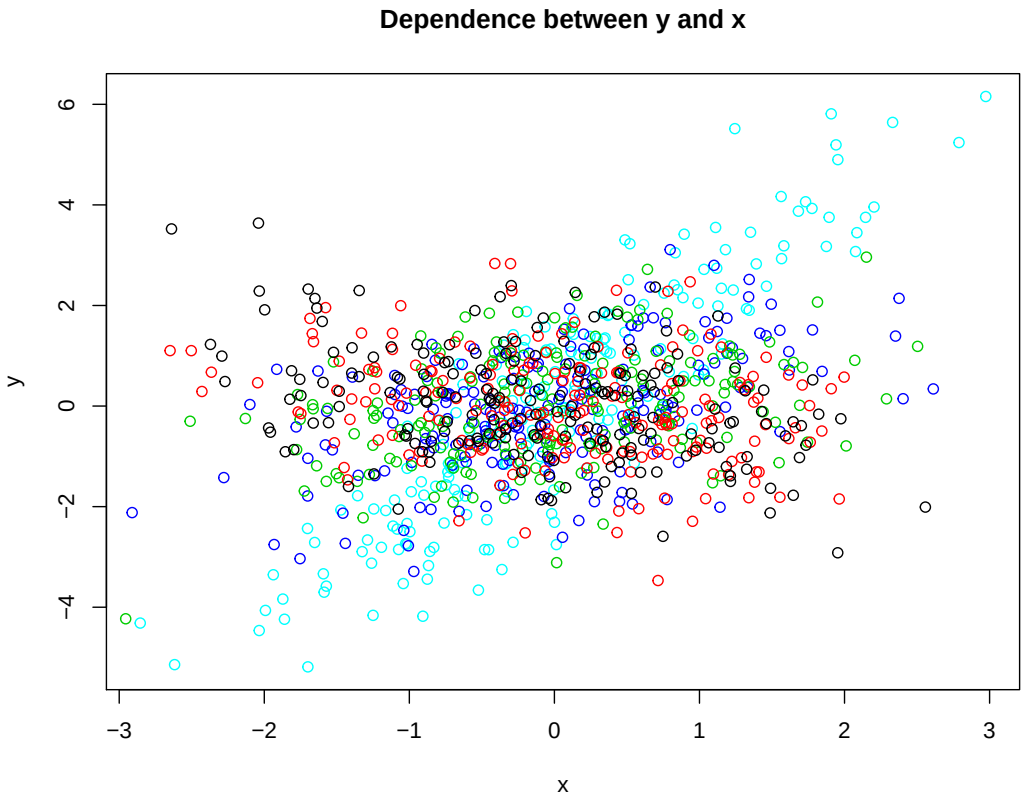


Figure 3: Data generated from mixture of regressions with 5 components, each with 1 explanaotry variable



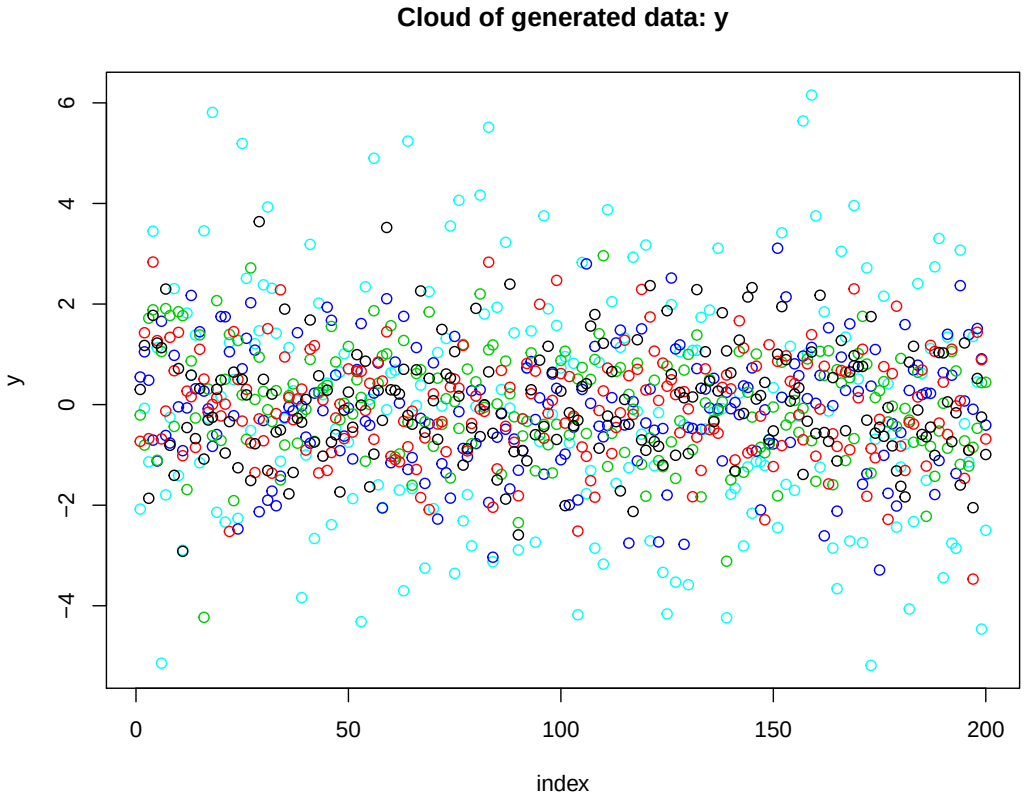
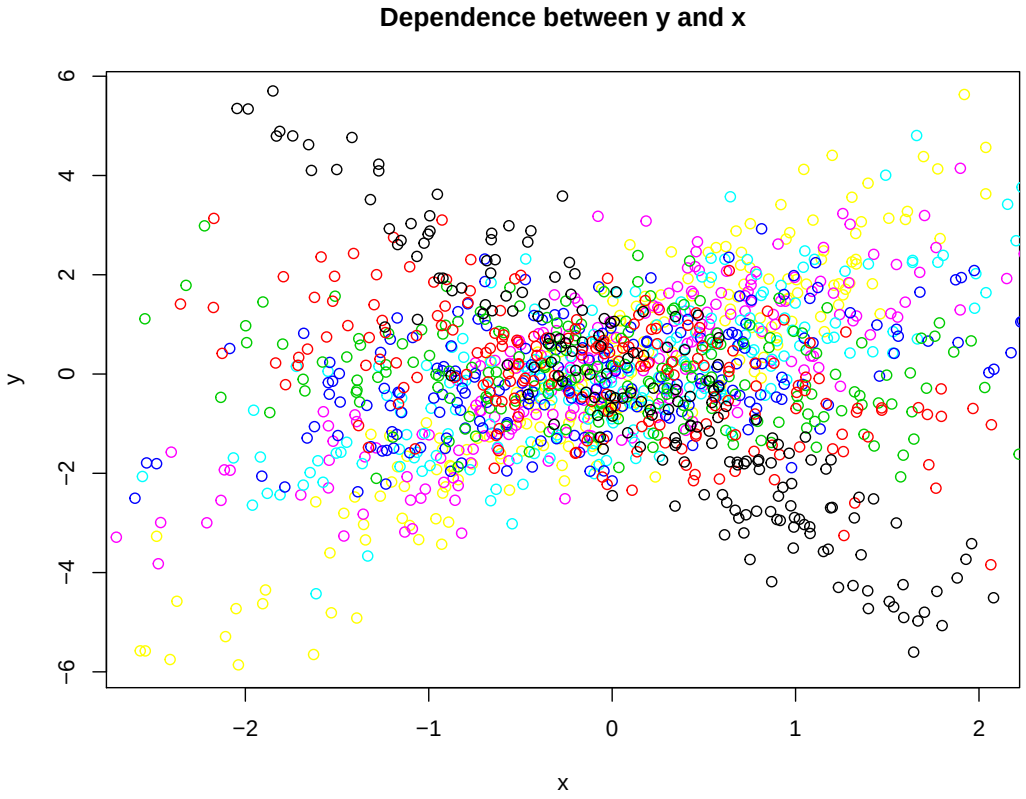
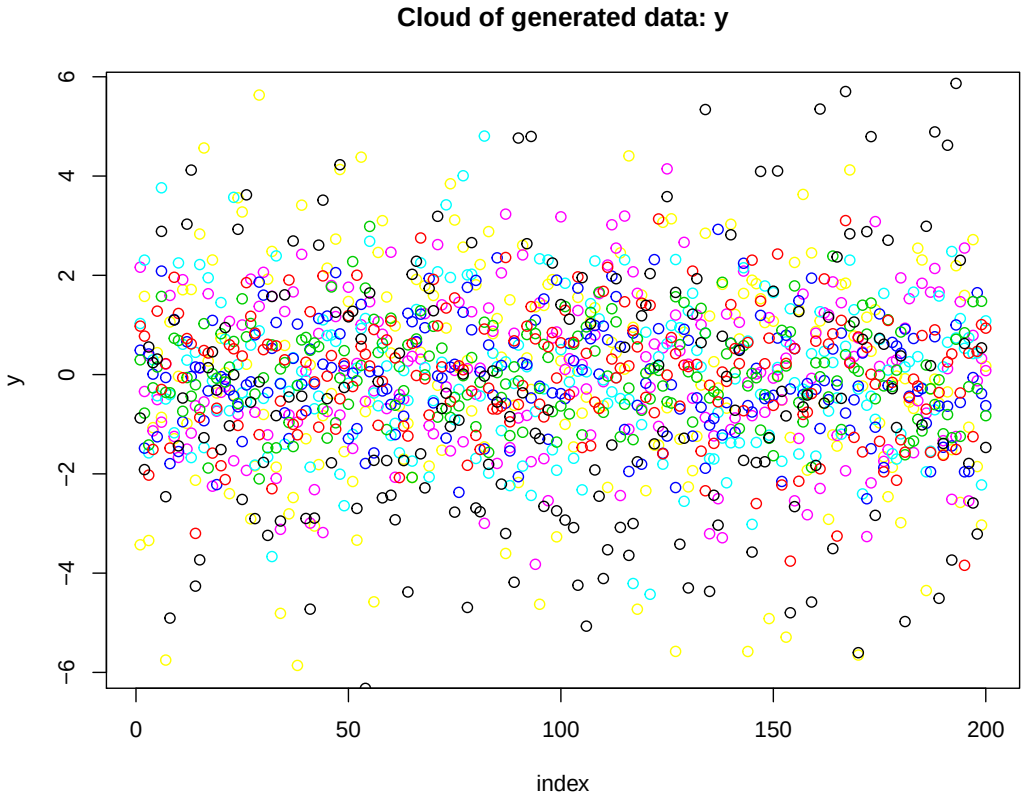


Figure 4: Data generated from mixture of regressions with 5 components, each with 1 explanaotry variable





Such models are estimated by augmenting the model with a set of components label vectors $\{z_i\}_{i=1}^n$, where

$$z_i = [z_{1i}\ z_{2i}\ \ldots z_{Gi}], \tag{4}$$

and $z_{gi} = 1$ implies that the i-th individual is drawn from the g-th component of the mixture, and $\sum_{g=1}^{G} z_{gi} = 1$, for a given i, so that an individual can not belong to two mixtures at the same time.

The component label vector z_i depends on a vector of component probabilities π. The likelihood function for this model is based on the Normal probability and the augmented label vector (let θ denote all the parameters and component indicator variables in the model)

$$L(\theta) = \Pi_{i=1}^{n}[\phi(y_i; x_i\beta_1, \sigma_1^2)]^{z_1 i}[\phi(y_i; x_i\beta_2, \sigma_2^2)]^{z_2 i} \ldots [\phi(y_i; x_i\beta_G, \sigma_G^2)]^{z_G i}. \tag{5}$$

Thus, depending on which z is equal to 1, the respective component of the likelihood is active.

Apart from likelihood, the priors need to be stated for component indicators z and component probabilities $\pi = [\pi_1, \pi_2, \ldots, \pi_G]$.

$$z_i|\pi \sim M(1, \pi), \quad p(z_i|\pi) = \Pi_{g=1}^{G}\pi_g^{z_{g_i}}, \tag{6}$$

$$\pi \sim D(\alpha_1, \alpha_2, \ldots, \alpha_G), \quad p(\pi) \propto \pi_1^{\alpha_1 - 1}\pi_2^{\alpha_2 - 1} \ldots \pi_G^{\alpha_G - 1}, \tag{7}$$

where M and D denote the multinomial and Dirichlet distributions, respectively.

Furthermore the priors for the parameters governing the Normal components are necessary. Traditionally for the mean and variance, the normal and the inverted gamma priors are imposed, respectively

$$\beta_g \sim N(\mu_{\beta_g}, V_{\beta_g}) \tag{8}$$

$$\sigma_g^2 \sim IG(a_g, b_g), \tag{9}$$

where $\mu_{\beta_g}, V_{\beta_g}, a_g, b_g$ denote the hyperparameters.

Given this setup a complete set of conditional distributions can be obtained. For clarity of notation let us denote by θ_{-x} a set of all parameters apart from x.

$$\beta_g|\theta_{-\beta_g}, y \sim N(D_{\beta_g}d_{\beta_g}, D_{\beta_g}), \quad g = 1, 2, \ldots, G, \tag{10}$$

where

$$D_{\beta_g} = \left[\left(\sum_i z_{gi} x_i x_i' \right) / \sigma_g^2 + V_{\beta_g}^{-1} \right]^{-1} \tag{11}$$

and

$$d_{\beta_g} = \left(\sum_i z_{gi} x_i' y_i \right) / \sigma_g^2 + V_{\beta_g}^{-1} \mu_{\beta_g}. \tag{12}$$

Then for the variance of the regression error term the conditional distribution is given in inverted gamma form

$$\sigma_g^2 | \theta_{-\sigma_g^2}, y \sim IG \left(\frac{n_g}{2} + a_g, \left[b_g^{-1} + \frac{1}{2} \sum_i z_{gi} (y_i - x_i \beta_g)^2 \right]^{-1} \right), \tag{13}$$

where $n_g = \sum_i z_{gi}$ denotes the number of observations in the g-th component of the mixture.

For the indicators the conditional distribution is given

$$z_i | \theta_{-z_i}, y \sim M \left(1, \left[\frac{\pi_1 \phi(y_i; x_i \beta_1, \sigma_1^2)}{\sum_{g=1}^G \pi_g \phi(y_i; x_i \beta_g, \sigma_g^2)}, \frac{\pi_2 \phi(y_i; x_i \beta_2, \sigma_2^2)}{\sum_{g=1}^G \pi_g \phi(y_i; x_i \beta_g, \sigma_g^2)}, \ldots, \frac{\pi_G \phi(y_i; x_i \beta_G, \sigma_G^2)}{\sum_{g=1}^G \pi_g \phi(y_i; x_i \beta_g, \sigma_g^2)} \right] \right). \tag{14}$$

Finally, for the component probability vector π, the posterior distribution is in Dirichlet form similarly to the prior

$$\pi | \theta_{-\pi} \sim D(n_1 + \alpha_1, n_2 + \alpha_2, \ldots, n_G + \alpha_G) \tag{15}$$

To sample the joint posterior distribution, the Gibbs sampler can be implemented to sample the individual condititonal distributions in the following order (10, 13, 14, 15). The initial values can be drawn from the respective priors accordingly.

Now let us assume that the model in (1) is a general form with fixed number of variables in x_i, but our objective is to select a functional form which is optimal in terms of fit. To that end we need a variable selection algorithm. One of the method to introduce it comes with stochastic search variable selection ideas. The model specification above would require minor alteration, which can have deep consequences for the model flexibility.

This alteration comes for the prior for the regression coefficients and is useful for the case where a larage number of explanatory variables exist, but the researcher does not know which are likely to be important in the model. To capture that the prior for each regression coefficient is specified as a mixture of two normals, both with mean zero, but with completely different variance. One of the term in the mixture has very small variance, what implies that the coefficient is virtually zero and thus the variable can be effectively excluded from the model. The other coefficient has a large variance, what means that it is most likely different from

0 and thus the variable should be retained in the model. Formally, for each coefficient $\beta_{g,j}$ the prior is given by

$$\beta_{g,j}|\gamma_{g,j} \sim (1-\gamma_{g,j})N(0,\tau^2) + \gamma_{g,j}N(0,c^2\tau^2) \tag{16}$$

Then $V_{\beta_g}(\gamma_g)$ is the diagonal matrix with the (j,j)-th element given by $(1-\gamma_{g,j})\tau^2 + \gamma_{g,j}c^2\tau^2$. It enters in the sampling algorithm via (11). At the same time elements $\gamma_{g,j}$ need to be sampled appropriately. They are driven by the binomial distribution controlled by the probability derived from (16)

$$\gamma_{g,j}|\theta_{-\gamma_{g,j}}, \beta_{g_j} \sim B(1, \frac{p\phi(\beta_{g,j};0,c^2\tau^2)}{p\phi(\beta_{g,j};0,c^2\tau^2) + (1-p)\phi(\beta_{g,j};0,\tau^2)}), \quad g=1,\ldots,G, \quad j=1,\ldots,M. \tag{17}$$

where parameter p denotes the prior probability for the variable to be included in the model.