LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Finite Mixtures of Generalized Linear Regression Models

Bettina Grün[1] **& Friedrich Leisch**[2]

[1] Department of Statistics, WU Wien
[2] Department of Statistics, LMU München

ISM, Tokyo, 6.12.2007

Overview

- Motivation and definition of model class
- Identification and estimation
- Varying and fixed effect models
- Dealing with label switching and genuine multimodality

Finite mixture models

$$H(y \mid x, w, \Theta) = \sum_{k=1}^{K} \pi_k(w, \alpha) F(y \mid x, \theta_k)$$

$$\forall\, w, \alpha: \qquad \sum_{k=1}^{k} \pi_k(w, \alpha) = 1, \qquad \pi_k(w, \alpha) \geq 0$$

H ... mixture distribution (density h)
y ... response
x ... regressors
Θ ... vector of all parameters
K ... number of components (classes)
π_k ... prior class probabilities
w ... concomitant variables
α ... concomitant parameters
F ... component distribution function (density f)
θ_k ... component specific parameters

Special Cases

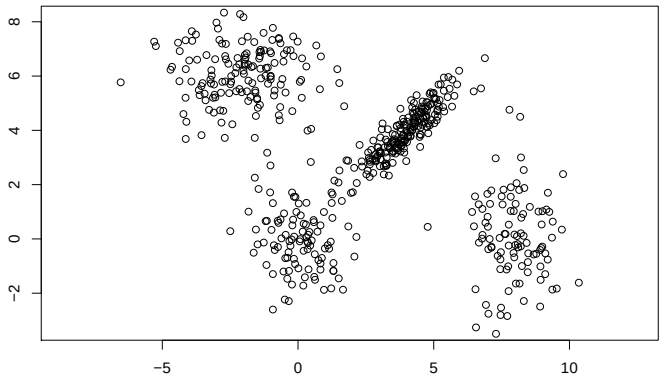
- Constant prior probabilities:

$$H(y \mid x, \Theta) = \sum_{k=1}^{K} \pi_k F(y \mid x, \theta_k)$$

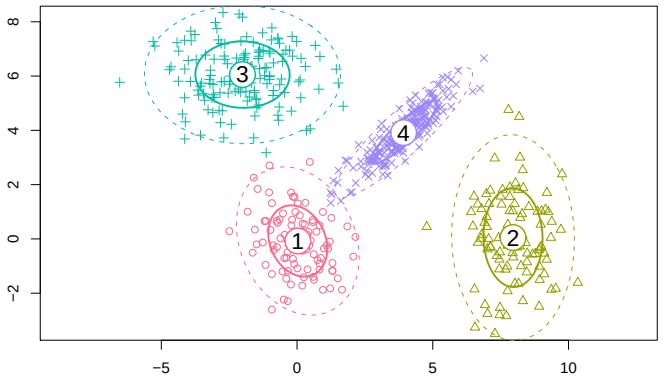
- Constant prior probabilities and no regression part (model-based clustering):

$$H(y, \Theta) = \sum_{k=1}^{K} \pi_k F(y, \theta_k)$$

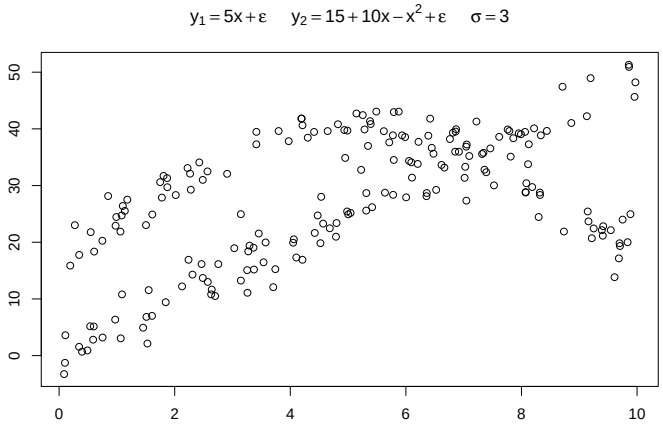
Simple Artificial Examples



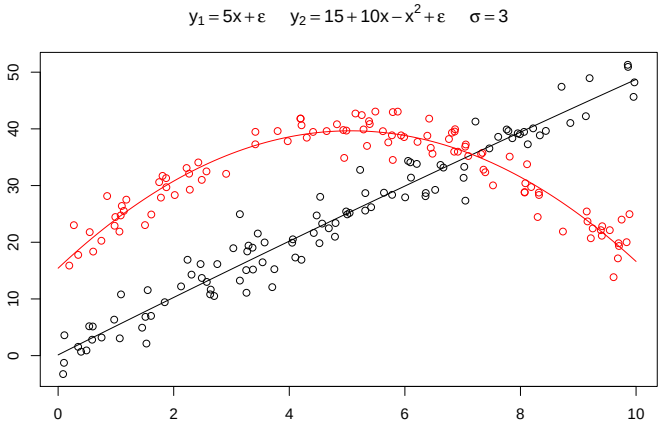
Simple Artificial Examples



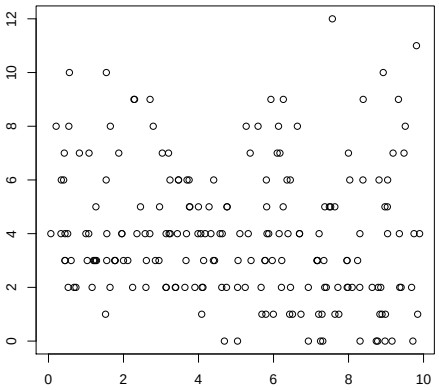
Simple Artificial Examples



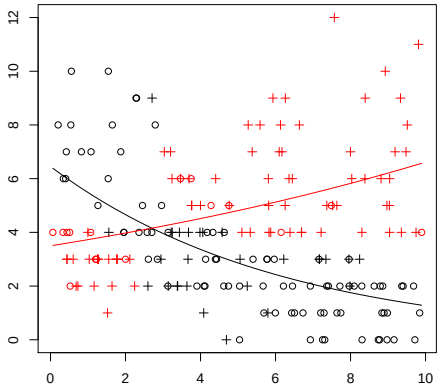
Simple Artificial Examples



Simple Artificial Examples



Simple Artificial Examples



Estimation

Bayesian: MCMC, Gibbs-Sampling

Maximum Likelihood:
- Direct optimization of likelihood (mostly in simpler cases)
- EM-Algorithm for more complicated models
- ...

Both Bayesian and EM-Estimation augment each observation (x_n, w_n, y_n) with an unobserved multinomial variable $z_n = (z_{n1}, \ldots, z_{nK})$, where $z_{nk} = 1$ if (x_n, w_n, y_n) belongs to class k and $z_{nk} = 0$ otherwise.

Estimation with EM: E-Step

Log-Likelihood for augmented (complete) sample is linear in latent variable z_n:

$$l(\Theta) = \log L(y_1, \ldots y_N \,|\, x_1, \ldots x_N, w_1, \ldots w_N, \Theta, z_1, \ldots, z_N) =$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\log \pi_k(w_n, \alpha) + \log F(y_n \,|\, x_n, \theta_k)\right)$$

E-step takes expectation of expression above:

$$Q(\Theta|\Theta^{(j)}) = \mathbb{E}(l(\Theta)|\Theta^{(j)})$$

$\rightarrow$ replace the missing data z_{nk} by their expectation, the estimated a-posteriori probabilities

$$\hat{p}_{nk} = \frac{\pi_k^{(j)}(w_n, \alpha) f(y_n|x_n, \theta_k^{(j)})}{\sum_{l=1}^{K} \pi_l^{(j)}(w_n, \alpha) f(y_n|x_n, \theta_l^{(j)})}.$$

Estimation with EM: M-Step

Given the estimates for the a-posteriori probabilities $\hat{p}_{nk}$ (which are functions of $\Theta^{(j)}$), obtain new estimates $\Theta^{(j+1)}$ of the parameters by solving

$$\begin{aligned}\Theta^{(j+1)} &= \arg\max_{\Theta} Q(\Theta|\Theta^{(j)}) \\ &= \arg\max_{\Theta} \left(Q_1(\theta^{(j+1)}|\Theta^{(j)}) + Q_2(\pi^{(j+1)}|\Theta^{(j)})\right)\end{aligned}$$

where

$$Q_1(\theta|\Theta^{(j)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \hat{p}_{nk} \log(f(y_n|x_n, \theta_k))$$

and

$$Q_2(\alpha|\Theta^{(j)}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \hat{p}_{nk} \log(\pi_k(w_n, \alpha)).$$

Q_1 and Q_2 can be maximized separately, both are standard weighted ML-problems.

Estimation with EM: M-Step

- Q_1 is maximized using weighted ML estimation of the component models, e.g., GLMs.

- The most popular choice of concomitant model $\pi_k(w, \alpha)$ is multimomial logit or probit, this can again be estimated by weighted ML.

  If the π_k are constant, the solution is simple:

$$\hat{\pi}_k = \frac{1}{N}\sum_{n=1}^{N} \hat{p}_{nk} \quad \forall k = 1, \ldots, K.$$

Estimation with EM: M-Step

Variations: If maximization of weighted loglikelihood in M-step is infeasible: assign each observation to one class and maximize likelihood only on those:

$$\max_{\theta_k} \sum_{n:z_{nk}=1} \log F(y_n|x_n, \vartheta_k) \tag{1}$$

This corresponds to allow only 0 and 1 as weights.

Possibilities:

- hard assignment to class with maximum posterior probability p_{nk}.
  $\rightarrow$ classification likelihood (Fraley & Raftery, 2000), similar to K-means.

- random assignment to classes with probabilities p_{nk},
  $\rightarrow$ "Bayesian E-step".

Identifiability

Can the true model be estimated from a sample of infinite size?

label switching: model is invariant against permutation of class numbers $\rightarrow$ sorting necessary.

intra-component label switching: For categorical regressors, parts of the components can switch their labels.

redundant paramters: components with $\pi_k = 0$ or $\theta_i = \theta_k$. Mathematically exluded in our definition, but what about "numeric equality"?

generic problems: caused by family of component distributions F_k.

Mixtures without regression

identifiable: (multivariate) normal, gamma, exponential, Cauchy and Poisson distribution

not identifiable: discrete and continous uniform distribution

conditionally identifiable: mixtures of binomial and multinomial distributions are identifiable if

$$M \geq 2K - 1$$

where M is the number of repeated measurements ber individual.

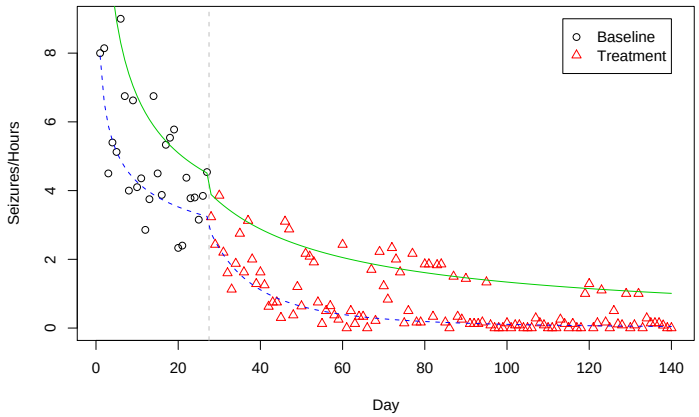
Mixtures without regression

linear models: Mixtures of linear regression models with Gaussian noise are identifiable, if the number of components K is smaller than the minimal number of hyperplanes necessary to cover all regressors without the intercept (Hennig, 2000).

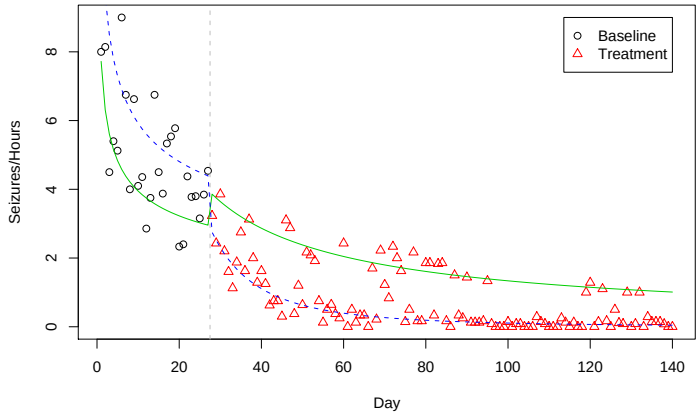
generalized linear models: Analogous condition for linear predictor, additional conditions depending on distribution of response (Grün 2006, Grün & Leisch, 200x).

Note: Sufficient condition that all models from a certain class are identifiable, not necessary for all possible models of a class. Hard to check in practice.

Example: Epilepsy



Example: Epilepsy



Varying and fixed effects

For notational simplicity, assume we have a mixture of standard linear regression models, i.e., for observation (x_n, y_n) and component k

$$y_n \sim N(x_n^t \beta_k, \sigma_k)$$

The following effects for β_k and σ_k can be distinguished:

fixed effects: The parameters are fixed over all components.
varying effects: The parameters vary for all components.
nested varying effects: There are groups of parameters with the same coefficients.

Simple Example

Finite mixture of linear regression models with 3 components given by

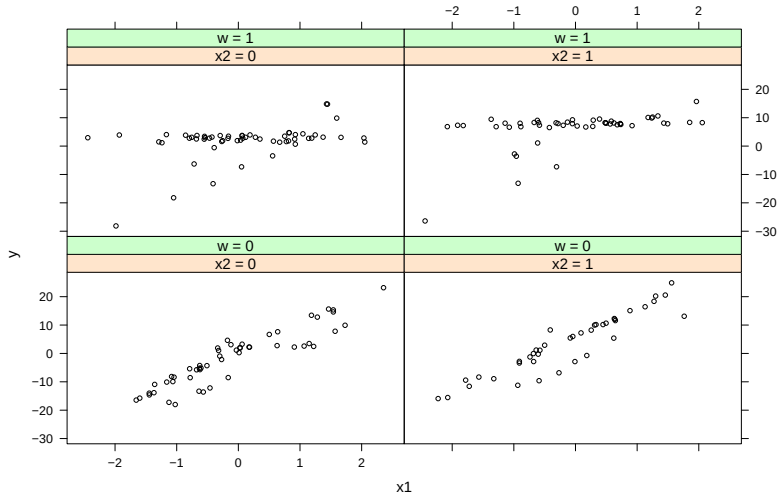
$$\begin{aligned} \text{Class 1:} \quad & y = -8 + 10x_1 + 5x_2 + \epsilon \\ \text{Class 2:} \quad & y = 1 + 10x_1 + 5x_2 + \epsilon \\ \text{Class 3:} \quad & y = 3 + + 5x_2 + \epsilon \end{aligned}$$

with $x_1, \epsilon \sim N(0,1)$ and $x_2 \in \{0,1\}$.

The component weights depend on the variable $w \in \{0,1\}$ and are determined by

$$\begin{aligned} \text{Class 2:} \quad & \text{logit}(\pi_2(w,\alpha)) = 2 - 2w \\ \text{Class 3:} \quad & \text{logit}(\pi_3(w,\alpha)) = 2w \end{aligned}$$

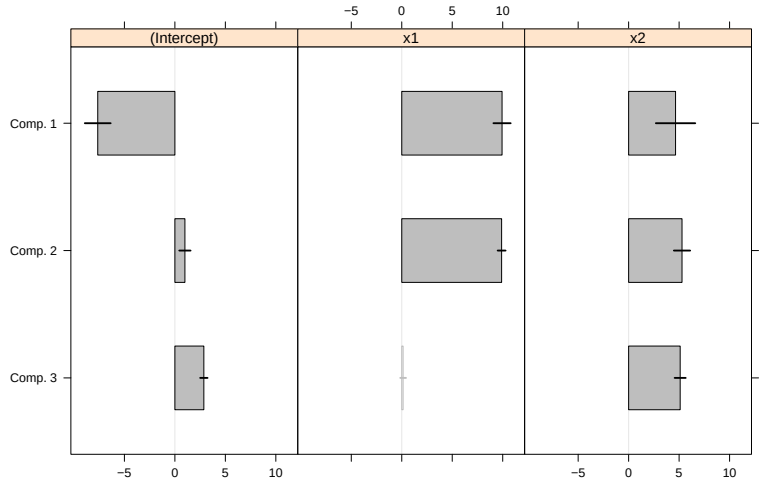
Simple Example



Simple Example

```
Call:
summary(object = refit(Fitted.1))

Number of components: 3

$Comp.1
            Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -7.64105    0.62570 -12.2120 < 2.2e-16
x1           9.93517    0.41755  23.7937 < 2.2e-16
x2           4.64805    0.95338   4.8753 1.086e-06

$Comp.2
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99396    0.28129  3.5336  0.00041
x1           9.89248    0.19719 50.1661  < 2e-16
x2           5.28808    0.40523 13.0496  < 2e-16

$Comp.3
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.86916    0.18516 15.4952   <2e-16
x1           0.13515    0.14474  0.9337   0.3505
x2           5.10620    0.27630 18.4805   <2e-16
BIC: 883.59
```

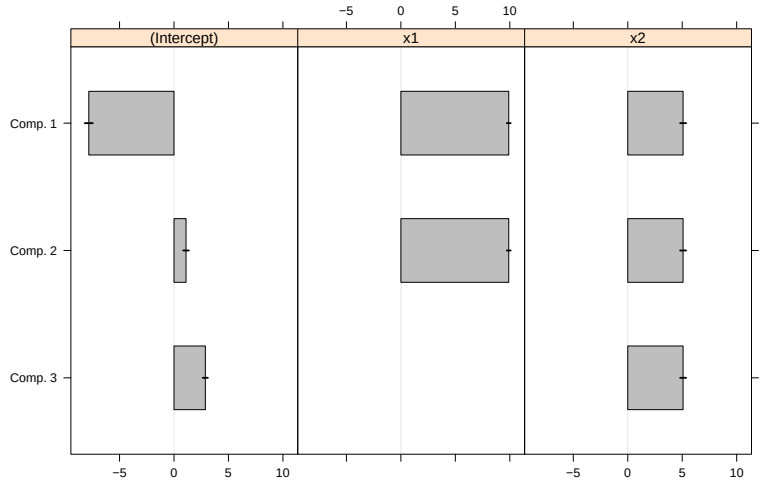
Simple Example



Simple Example

```
Call:
summary(object = refit(Fitted.2))

Number of components: 3

$Comp.1
             Estimate Std. Error t value  Pr(>|t|)
x2           5.080554   0.140110  36.261 < 2.2e-16
x1           9.903456   0.090602 109.307 < 2.2e-16
(Intercept) -7.822345   0.186496 -41.944 < 2.2e-16

$Comp.2
            Estimate Std. Error  t value  Pr(>|t|)
x2          5.080554   0.140110  36.2613 < 2.2e-16
x1          9.903456   0.090602 109.3068 < 2.2e-16
(Intercept) 1.106160   0.133236   8.3023 1.680e-14

$Comp.3
            Estimate Std. Error t value  Pr(>|t|)
x2           5.08055    0.14011  36.261 < 2.2e-16
(Intercept)  2.87856    0.11937  24.115 < 2.2e-16
BIC: 865.85
```


Simple Example



Parameters of Mixtures

Ω is the space of admissible parameter vectors Θ for mixtures with K components where

- $0 < \pi_k < 1 \ \forall k$
- $\sum_{k=1}^{K} \pi_k = 1$
- $\theta_k \neq \theta_l \ \forall l \neq k$

$\mathcal{A}_K = \mathcal{A}_K(F, \Omega)$ denotes the set of all finite mixture models with K components and mixture distributions of form $H(\cdot|\cdot, \Theta)$, $\Theta \in \Omega$.

$\mathcal{X}_N$ is a given sample of N i.i.d. observations from the data generating process. $a(\mathcal{X}_N) \in \mathcal{A}_K$ denotes the model with the maximum likelihood for the given dataset $\mathcal{X}_N$.

Parameters of Mixtures

As $\mathcal{X}_N$ is a random variable, $a(\mathcal{X}_N)$ is also a random variable which follows a distribution $\mathcal{A}_K$.

As each model $a(\mathcal{X}_N)$ can be represented by a parameterization $\Theta \in \Omega$, these parameterizations follow a distribution $\mathcal{O}$.

Multimodality

The mixture likelihood is or might be multimodal due to

- label switching
- identifiability problems
- local modes

$\mathscr{A}_K$ induces a system of equivalence classes Ξ on Ω:

$$\Theta_1, \Theta_2 \in \Xi \Leftrightarrow \exists \nu \in \text{Perm}(K) : \Theta_1 = \nu(\Theta_2)$$

Let $\tilde{\Omega} = \text{ident}(\Omega) \subset \Omega$ be the subset of parameterizations which contain only one permutation of each possible set of component parameters.

Definition: The distribution $\mathcal{O}$ of the parameters $\Theta \in \Omega$ is called genuinely multimodal if it holds for the set of modes $\mathcal{M}$ of $\mathcal{O}$ that

$$\exists \Theta_1, \Theta_2 \in \mathcal{M} : \Theta_1 \neq \nu(\Theta_2) \qquad \forall \nu \in \text{Perm}(K)$$

Multimodality

Determine $\tilde{\Omega} \subset \Omega$ by

- imposing an ordering constraint on one parameter
- fixing the membership of some observations (Chung et al. 2004)
- using relabelling algorithms based on
  - label-invariant loss functions (Stephens 2000)
  - arbitrary loss functions and constrained clustering (Grün & Leisch 2006)

Approximate $\mathcal{O}$

1. Determine $\hat{a}(\mathcal{X}_N) \in \mathscr{A}_K$ and a corresponding parameterization $\hat{\Theta} \in \Omega$, e.g. with the EM algorithm using the best solution of several random initializations.

2. Sample B bootstrap samples $\mathcal{X}_N^b$ independently for $b = 1, \ldots, B$ with the parametric bootstrap, i.e. $\mathcal{X}_N^b \sim \hat{a}(\mathcal{X}_N)$.

3. Fit models to the bootstrap samples, i.e. deterime $\hat{a}(\mathcal{X}_N^b) \in \mathscr{A}_K$ using the EM algorithm with several random initializations.

4. Analyze the parameterizations $\hat{\Theta}_b$ of the bootstrap models $\hat{a}(\mathcal{X}_N^b)$ which imply an approximation of the distribution $\mathcal{O}$.

Detect multimodality

Given the approximation of $\mathcal{O}$:

- Test for unimodality of the component specific parameters of $\Theta \in \tilde{\Omega}$.
- Compare the variation of the bootstrap results to confidence bands determined using standard asymptotic theory.

Constrained clustering

- Alternative method to relabelling algorithm.
- Use the estimates of each component separately in the clustering algorithm.
- In addition impose a must-not link constraint in order to ensure that components from the same draw/ bootstrap sample are in different clusters.

Constrained clustering

For N MCMC draws or bootstrap samples with component specific estimates $x_{k,n}$, $k = 1, \ldots, K$, $n = 1, \ldots, N$:

1. Start with a random set of initial centroids $C_S := \{c_1, \ldots, c_S\}$ with $S \geq K$.
2. Assign each point $x_{k,n} \in X_{K,N}$ to the cluster of the closest centroid.
3. If the constraint is violated, find the best assignment under the constraint.
4. Update the set of centroids holding the cluster memberships fixed.
5. Repeat from 2 until convergence.

Step 3 can be performed by solving a linear sum assignment problem.

Constrained clustering

Convergence: of the resulting clustering algorithm is guaranteed, because each step cannot increase the objective function and there is only a finite number of possible assignments of N groups to S clusters.

Initialization: Convergence is only to a local optimum. Different random starts are advisable.

Distance measures: e.g. Kullback-Leibler divergence between rescaled posteriors

Simple Example

The mixture regression is given by

$$H(y|\mathbf{x},\Theta) \;=\; \sum_{s=1}^{3} \frac{1}{3} N(\mu_s(\boldsymbol{x}), 0.01)$$

where $\mu_s(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_s$ and $N(\mu, \sigma^2)$ is the Gaussian distribution.

The covariates are:

- intercept
- $x_1 \in [0, 1]$
- interaction between x_1 and $x_2 \in \{0, 1\}$

Component means:

- $\boldsymbol{\mu}(x_1 = 0, x_2 = 0) = (4, 4, 2)$
- $\boldsymbol{\mu}(x_1 = 1, x_2 = 0) = (4, 2, 2)$
- $\boldsymbol{\mu}(x_1 = 1, x_2 = 1) = (2, 0, 2)$

Simple Example

This mixture is not identifiable. There are two observational equivalent parameterizations:

Solution 1:

$$\begin{aligned}\boldsymbol{\beta}_1^{(1)} &= (4, -2, \;\; 0)' \\ \boldsymbol{\beta}_2^{(1)} &= (4, \;\; 0, -4)' \\ \boldsymbol{\beta}_3^{(1)} &= (2, \;\; 0, \;\; 0)'\end{aligned}$$

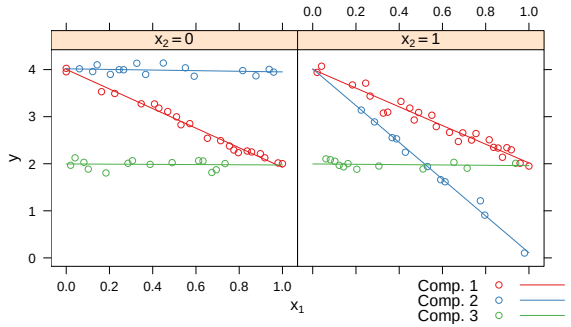
Solution 2:

$$\begin{aligned}\boldsymbol{\beta}_1^{(2)} &= (4, -2, -2)' \\ \boldsymbol{\beta}_2^{(2)} &= (4, \;\; 0, -2)' \\ \boldsymbol{\beta}_3^{(2)} &= (2, \;\; 0, \;\; 0)'\end{aligned}$$

Simple Example

Sample with 100 observations with equidistant x_1 values for each of the two x_2 values

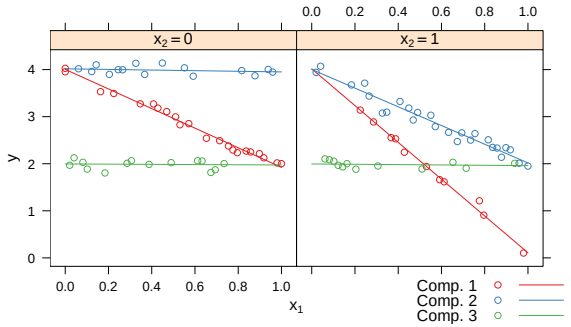
Solution 1:



Simple Example

Sample with 100 observations with equidistant x_1 values for each of the two x_2 values

Solution 2:

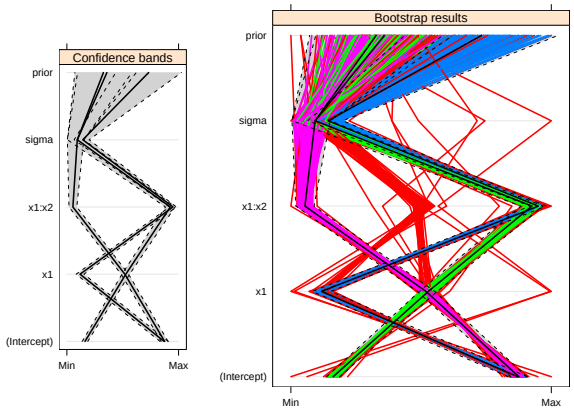


Simple Example

- Fit finite mixture with 3 components using the EM-algorithm
- Investigate presence of "genuine" multimodality:
  - Draw 200 parametric bootstrap samples
  - Fit a mixture with 3 components to each bootstrap sample with the EM-algorithm using random initialization

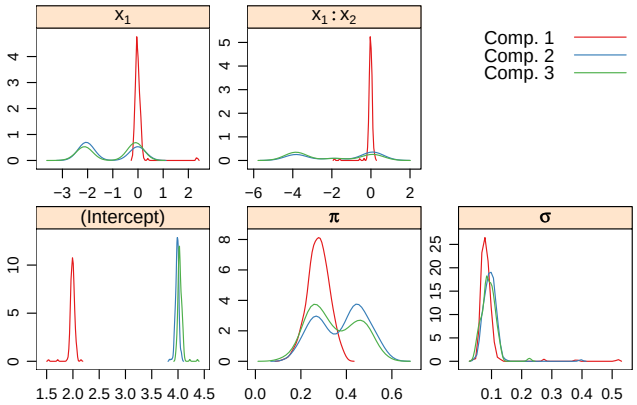
Simple Example

Comparison of bootstrap samples and asymptotic confidence bands:



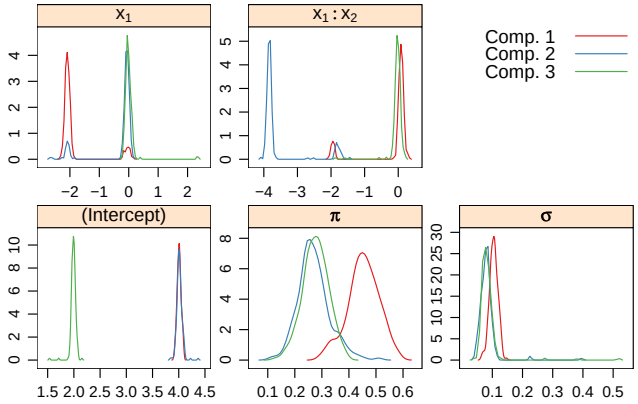
Simple Example

Ordering constraint on the intercept:



Simple Example

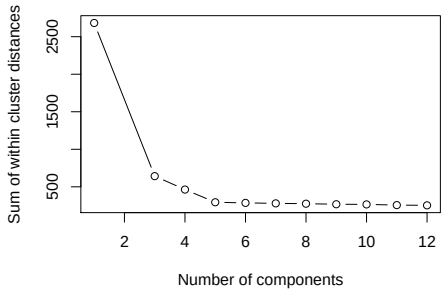
Stephens-relabelling with respect to KL-divergence between the posteriors:



Simple Example

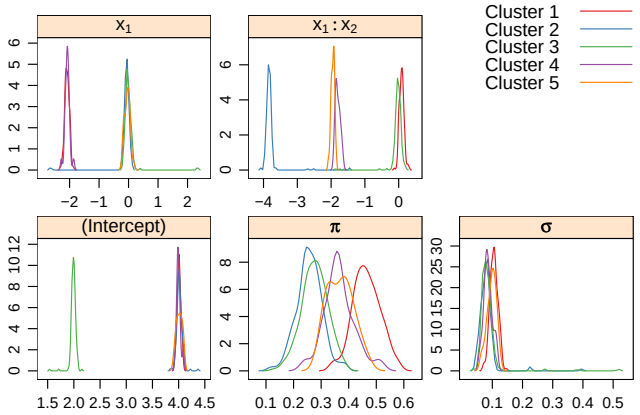
Constrained clustering using the rescaled posteriors with the KL divergence with different number of components:

→ Sum of within cluster distances indicates 5 components.



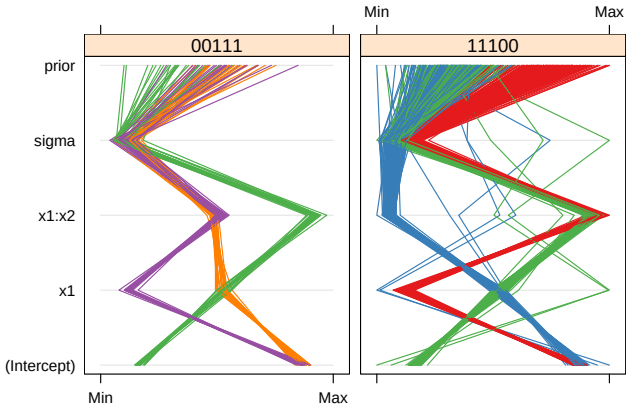
Simple Example

Constrained clustering using the rescaled posteriors with the KL divergence with different number of components:



Simple Example

Combining the cluster assignments with the group constraints allows to determine the different "genuine" modes



Software & Papers

- Software is available as package `flexclust` from CRAN (talk tomorrow).
- Lists of papers (with PDFs where possible) are available at

http://www.statistik.lmu.de/~leisch/
http://www.ci.tuwien.ac.at/research/mixtures/index.html