# Prediction with a flexible finite mixture-of-regressions

Ilmari Ahonen[a,b,1,*], Jaakko Nevalainen[c,a], Denis Larocque[d]

[a] *Department of Mathematics and Statistics, University of Turku, Finland*
[b] *Institute of Biomedicine, University of Turku, Finland*
[c] *School of Health Sciences, University of Tampere, Finland*
[d] *Department of Decision Sciences, HEC Montreal, Canada*

**Abstract**

Finite mixture regression (FMR) is widely used for modeling data that originate from heterogeneous populations. In these settings, FMR can offer increased predictive power compared to more traditional one-class models. However, existing FMR methods rely heavily on mixtures of linear models, where the linear predictor must be given as an input. A flexible FMR model is presented using a combination of the random forest learner and a penalized linear FMR. The performance of the new method is assessed by predictive log-likelihood in extensive simulation studies. The method is shown to achieve equal performance with the existing FMR methods when the true regression functions are in fact linear and superior performance in cases where at least one of the regression functions is nonlinear. The method can handle a large number of covariates, and its predictive ability is not greatly affected by surplus variables.

*Keywords:* finite mixture regression, random forest, prediction intervals, bootstrap, penalization

## 1. Introduction

Finite mixture models provide a tool for analyzing data that are suspected to arise from a heterogeneous population (Quandt, 1972; De Veaux, 1989; Khalili and Chen, 2007; McLachlan and Peel, 2004). Specifically, a finite mixture-of-regressions (FMR) is used for modelling a continuous response as

*Corresponding author, email: ilmari.ahonen@utu.fi, address: Department of Mathematics and Statistics, Vesilinnantie 5, Quantum, 20014 Turun yliopisto, Finland

a function of covariates. Consider a continuous outcome variable $Y$ and a $p$-dimensional covariate vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$. The data is assumed to originate from a population consisting of multiple latent, unobserved classes. Depending on the class, a separate regression model between the outcome and the covariate applies. Assuming a Gaussian linear model for each component, the density function of $Y|\boldsymbol{x}$ is

$$f(y|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \phi(y; \boldsymbol{x'}\boldsymbol{\beta}_k, \sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1, \tag{1}$$

where $K$ is the number of classes, $\phi(y; \mu, \sigma)$ is the density function of the normal distribution with mean $\mu$ and variance $\sigma^2$, $\boldsymbol{\beta}_k$ is the vector of regression coefficients for class $k$, $\sigma_k$ is the standard deviation of the error term in class $k$ and $\pi_k > 0$ is the probability of membership in class $k$. Models of this type are applied in a variety of fields including ecology, genetics, economics and marketing where such unobserved latent classes are often appropriate. The parameters of this model are typically estimated using the expectation maximization (EM) -algorithm (Dempster et al., 1977; Hastie et al., 2009). Penalized versions of the FMR have also been proposed (Khalili and Chen, 2007; Galimberti et al., 2009; Khalili et al., 2010; Städler et al., 2010).

In this paper, we aim to achieve greater predictive power than linear FMR methods such as model (1). This is realised by considering a more flexible structure for the linear predictor $\boldsymbol{x'}\boldsymbol{\beta}_k$ to better model possible nonlinear dependencies and interactions between the covariates. However, instead of trying to accurately capture the underlying mean functions, we aim to estimate the full mixture density $f(y|\boldsymbol{x})$, which can then be used for prediction purposes. Thus, our emphasis is on identifying $f$ as a whole, not on the more conventional aim of recovering the underlying mean functions and the class labels. Semiparametric approaches for FMR with similar aims have been recently suggested by Huang et al. (2013b) and Xiang (2014), whereas Huang et al. (2013a) discuss the problem from a functional data-analysis perspective. Huang et al. (2013b) assume that the outcome follows a mixture of Gaussians whose class means and variances are unknown smooth functions of the covariates. The functions are estimated by local smoothing accompanied with a modified EM-algorithm. Similar assumptions are made by Xiang (2014), however with constant class variances. Again, the model is estimated with an EM-algorithm modified for local fitting. Both publications deal only with univariate covariates although results are said to be extendable to mul-

tivariate data. However in both cases, the authors also state that the curse of dimensionality (Hastie et al., 2009) reduces the usefulness of these extensions. Thus, their applicability in high dimensional data sets is questionable.

If the finite mixture regression model is thought of as a supervised learning method, its unsupervised counterpart would be model based clustering. Similar to FMR, these methods assume that the data originates from a mixture of $K$ distributions, typically Gaussians, but are pure clustering methods in the sense that no outcome variable is modelled based on a set of covariates. Instead, all observations originating from a particular class $k$ are assumed to have the same mean $\boldsymbol{\mu}_k$. This corresponds to applying FMR on a multidimensional outcome with only the intercept term in the model. Recent reviews on model based clustering from a variable selection perspective are given by Celeux et al. (2014) and Bouveyron and Brunet-Saumard (2014).

FMR is related to mixtures of experts (MOE) models that are frequently used in the machine learning applications (Yuksel et al., 2012). The idea of MOE is to have a collection of different learners, each specializing to separate parts of the covariate space, to form the overall regression model. For example, in a simple case of a continuous outcome $Y$ and a single covariate $x$, one could consider a linear model $E(Y) = \beta_0 + \beta_1 x$ when $x < a$ and a quadratic fit $E(Y) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ when $x \geq a$. In general, the learners are not limited to such simple polynomial models but can be arbitrarily flexible, e.g. feed-forward neural networks. Furthermore, instead of a hard transition between the models at $x = a$, a certain smoothing technique is used to ensure a soft transition. MOE and FMR can be seen as opposites in this matter: A MOE model invokes a strict (although soft) partition of the covariate space while FMR assumes global and constant class proportions. An extension of MOE, called hierarchical mixtures of experts (HME), allows the class probabilities $\pi_k$ to depend on the covariates. Although there are similarities between the HME approach and the problem posed in this paper, HME models cannot be easily applied to high-dimensional data without compromising flexibility (Yuksel et al., 2012; Hadavandi et al., 2016). Furthermore, the added estimation of the covariate-dependent class probabilities introduces excess variability that is likely to adversely effect the performance of these models under the assumptions of the problem. Similarly to the regular FMR, a penalized HME using generalized linear models as experts (Khalili, 2010) requires the definition of the correct structure for the linear predictor.

The paper is organized as follows. Section 2 provides a brief review of

penalized FMR and random forests, and describes in detail the proposed flexible FMR method, named FMRFLEX. Section 3 considers methods for measuring the predictive power in the FMR setting and the construction of prediction intervals. In section 4, we consider an extensive set of simulation studies where the predictive performance of FMRFLEX is compared to that of reference methods such as FMR and penalized FMR. In section 5, the method is applied to a worker wage prediction dataset. Finally, the paper is concluded with a short discussion in section 6.

## 2. Flexible FMR method

In order to obtain a flexible, nonparametric FMR fit, we replace the linear model $\boldsymbol{x}'\boldsymbol{\beta}_k$ in (1) with an arbitrary function of the covariate $h_k(\cdot) : \mathcal{R}^p \to \mathcal{R}$ and obtain

$$Y|\boldsymbol{x} \sim f(y|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \phi(y; h_k(\boldsymbol{x}), \sigma_k). \tag{2}$$

Our goal is then to estimate $f(y|\boldsymbol{x})$ in order to predict the response for a new observation $\boldsymbol{x}^+$. We achieve this by using a combination of the random forest learner (Breiman, 2001) and variable selection techniques in FMR models. The random forest is used for obtaining a large set of informative dummy variables that are then used as covariates in the subsequent penalized FMR regression model to obtain a flexible fit for the means. Since our main interest is the full density $f(y|\boldsymbol{x})$, the unknown mean functions $h_1(\boldsymbol{x}), \ldots, h_K(\boldsymbol{x})$ need to be estimated only to the extent that (2) estimates $f(y|\boldsymbol{x})$.

In summary, the assumptions made in this paper are:

- the outcome $y$ follows a mixture of Gaussians with means $h_1(\boldsymbol{x}), \ldots, h_K(\boldsymbol{x})$ and variances $\sigma_1^2, \ldots, \sigma_K^2$
- the observations $y_1, \ldots, y_N$ are conditionally independent given the covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$
- the class probabilities $\pi_1, \ldots, \pi_K$ are independent of the covariates $\boldsymbol{x}$.

In order to form a density-based prediction of $Y|\boldsymbol{x}^+$ under model (2), given the observed covariates $\boldsymbol{x}^+$, we need to find a set of component functions, variance terms and mixing probabilities. The key issue is the correct specification of the mean or component functions, especially when one has to go beyond simple linear or quadratic curves. This can be hard or impossible

to accomplish with 10s or 100s of covariates. Choosing the number of latent classes $K$ can also be difficult *a priori*, however model selection methods can be used for this purpose.

The FMRFLEX method uses the random forest to approximate the component functions, and more importantly, to extract a set of informative indicator variables that can be used in the regression. A penalized FMR is then fitted on these indicator variables and the original covariates to extract the final prediction. The motivation for this approach is to combine the flexible, data-driven identification of nonlinearities and interactions achieved with the random forest with the linear associations obtainable with the original covariates and the existing FMR methods. The main ideas of the FMRFLEX method can be summarized as:

1. Fit a random forest using the covariates $\boldsymbol{x}$ and the outcome $y$.
2. Extract the terminal nodes, that is the lowest branches of each of the trees in the random forest, as dummy variables.
3. Fit a penalized FMR model on $y$ using the covariates and the extracted dummy variables.

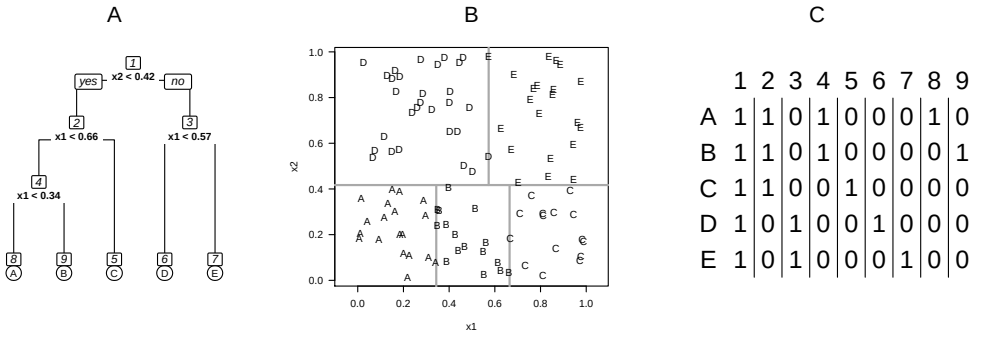
Here, $K$ is chosen using Bayesian information criterion (BIC). The following subsections briefly review and describe the implementation of each of these steps in greater detail.

### *2.1. Random forest*

Random forests (Breiman, 2001) are an especially popular method for both regression and classification. By combining an ensemble of decision trees, each built with a set of randomly selected covariates at each node of each tree and a bootstrap sample of the observations, a highly efficient learner is obtained. Each decision tree in a forest consists of consecutive binary cuts, where observations are divided between two branches based on given splitting rules, and finally assigned to the terminal nodes. One such tree and the resulting segmentation of a 2-dimensional covariate space is illustrated in Figure 1.

Importantly, each branch $b$ in the forest can be expressed as a dummy variable $z_b$, the value of which depends on the covariates (Figure 1):

$$z_b = \prod_{j=1}^{p} I(l_{b,j} < x_j < u_{b,j}), \quad -\infty \leq l_{b,j}, u_{b,j} \leq \infty, \tag{3}$$




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| B | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| C | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| D | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| E | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

Figure 1: **Panel A:** A tree has been fit to a data set with two covariates, resulting in a total of 9 branches of which 5 are terminal nodes (5-9). Each branching is based on a simple splitting criterion for the selected covariate. **Panel B:** The tree partitions the observations into 5 regions according to their covariate value, each corresponding to a terminal node. **Panel C:** Branch memberships can be expressed as a design matrix. All observations belong to branch 1, whereas each observation can belong to only one of the terminal nodes.

where, $l_{b,j}$ and $u_{b,j}$ are the splitting rules for the covariate $x_j$ in branch $b$. Consequently, the regions defined by the whole forest can be expressed as a large set of dummy variables. To simplify, we extract only those dummy variables that correspond to the terminal nodes of the trees (columns 5-9 in the panel C of Figure 1). No information is lost by this procedure, since all the other dummy variables can be obtained as linear combinations of these terminal nodes.

The BIC criterion obtained from the final FMR model is used for finding the values for the key parameter controlling the random forest, which is the minimum proportion of observations in the terminal nodes (minobs). This parameter directly controls the depth of the tree and thus its complexity and flexibility. Deeper trees typically offer increased accuracy but naturally result in a larger set of dummy variables with less data to estimate their regression coefficients. Therefore, slightly larger terminal nodes are preferable compared to the original application of random forests for regression. We experiment with two values minobs $= 0.05, 0.10$ and use a total of 200 trees in a forest.

### *2.2. Penalized FMR*

Joly et al. (2012) sought to compress the random forest in the ordinary one-class regression setting by applying shrinkage methods to the dummy

variables. By fitting a Lasso-penalized (i.e. L1-penalized) (Tibshirani, 1996) linear model between the outcome and the dummy variables, only the most important indicator variables are selected. This reduces the model's space complexity meaning that the model can be, if not fully represented, at least adequately approximated by a simpler form. The authors showed that an equal or even improved predictive power could be obtained with this procedure. Here, we utilize this idea in the FMR context to select those covariates and dummy variables that are important for the given regression problem.

Previous studies of variable selection in FMR models (Khalili and Chen, 2007; Galimberti et al., 2009; Khalili et al., 2010; Städler et al., 2010) have considered shrinkage using some form of Lasso-type penalty. The model parameters are estimated via maximum likelihood while modified EM-algorithms are developed to incorporate the penalization. Such an approach allows for different coefficients to be shrunk in the $K$ different classes. We opt to use the method proposed by Städler et al. (2010) due to its convenient $R$ implementation (R Core Team, 2013; Städler, 2010) and the ability to handle a large amount of covariates.

The most important tuning parameter for the FMR Lasso, in addition to the assumed number of classes $K$, is the penalty term $\lambda$ that controls the amount of penalization. We choose $\lambda$ based on the BIC criterion using a grid search over 100 values as suggested by Städler et al. (2010). The same criterion is used for estimating the number of latent classes $K$ when it is unknown.

In addition to the extracted dummy variables, the original covariates $\boldsymbol{x}$ are included in the linear prediction of the regression to model possible linear relationships efficiently. The dummy variables' role then becomes to model deviations from linearity, and should linearity be true, the model is expected to simplify greatly at the penalization stage.

The proposed method can be expressed as a linear FMR model of the covariates and the extracted dummy variables. Denote the vector of dummy variables $\boldsymbol{z} = (z_1, \ldots, z_r)$, where $r$ is the total number of extracted terminal nodes from all trees. The component functions in (2) are then estimated as

$$\boldsymbol{x}'\hat{\boldsymbol{\beta}}_k + \boldsymbol{z}'\hat{\boldsymbol{\theta}}_k \qquad k = 1, \ldots, K, \tag{4}$$

where the coefficients $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$ along with the standard deviations $\sigma_1, \ldots, \sigma_K$ and the mixing probabilities $\pi_1, \ldots, \pi_K$ are estimated with a Lasso-penalized likelihood.

It has been shown that a reduction in bias can be gained for Lasso-penalized models with an additional fitting step with only the selected covariates (Bühlmann and Van De Geer, 2011; Belloni and Chernozhukov, 2013). The covariates with regression coefficients shrunken to zero are dropped, and the model is refitted with only the remaining covariates without penalization. This additional step is implemented here as well.

*2.3. A modified flexible FMR algorithm to improve computational efficiency*

Steps 1-3 of the method need to be repeated for all combinations of the tuning parameters minobs and $\lambda$, after which the model with the lowest BIC is selected. In practice, this approach is computationally demanding when the number of dummy variables is large. The problem increases with the sample size, the number of trees used in the forest and their depth. In order to reduce the computational demand of the method, we propose an alternative strategy that utilizes an iterative approach:

1. Obtain an initial class assignment for all observations based on a linear FMR fit on the original data using maximum *a posteriori* probability (MAP). Then iterate steps 2-6 $M$ times:
2. Fit the random forest model for each class $k = 1, \ldots, K$ separately using the covariates $\boldsymbol{x}$ as input.
3. Extract the terminal nodes from all $K$ forests for the complete data as dummy variables $\boldsymbol{z}$.
4. Fit a penalized one-class regression for each class separately using the original covariates and the extracted dummy variables.
5. Fit the linear FMR model with $L_1$-penalty using those variables that have non-zero coefficients in at least one of the regression models in the previous step.
6. Aim to reduce the bias of the model by refitting again using only the covariates with non-zero coefficients in (5) in at least one class of the estimated FMR, and without penalization.
7. Update the class assignments based on the model in the previous step.

Following this approach, the computationally heavy variable selection is mainly done in step 4, separately for each class. With traditional penalization methods being highly efficient compared to the mixture alternatives, this greatly reduces the overall computational burden. We use the Lasso-regression method by Friedman et al. (2010) in step 4 and find the amount

of penalization using a grid search and 5-fold cross-validation as suggested by the authors. We found this procedure sufficient and note that this step does not need to be fully optimized to achieve its main goal, which is to filter out most of the unnecessary covariates for step 5. Using cross-validation for the selection of the penalty parameter has been found to have a tendency to over-select variables in practice (Feng and Yu, 2013; Spindler, 2014). This is desirable, because it lowers the chance of unintentionally losing important variables at this step. Five iterations ($M = 5$) of steps 2-7 are run after which the model with the lowest BIC in step 6 is chosen. In our simulations, the best BIC is typically produced in the first two iterations after which the improvement is usually negligible or non-existent. Thus, the relatively small number of iterations is justifiable.

### *2.4. Identifiability of the flexible FMR*

Identifiability is a critical issue when it comes to estimating FMR models. In complex settings, situations can occur, where multiple sets of parameter estimates result in equally good fits for the data. For example, crossing mean functions can lead to such ambiguity if the considered model is flexible enough, as is the case with the proposed method. However, in this example, the full mixture density would still remain identifiable and not compromise the prediction goal. Problems will arise if the target density itself is unidentifiable. Such severe examples are illustrated by Hennig (2000) where two differing sets of linear regression lines give equally good fits for the data and lead to different predictions if interpolated or extrapolated outside the data points. However, with the proposed method, predictions are never obtained for points outside their possible values of 0 and 1, letting us avoid the problem altogether.

## 3. Prediction

### *3.1. Predictive log-likelihood*

We consider a set of simulation studies to assess the predictive performance of our method. Hence, we need to form both a prediction and a metric that quantifies the difference between the prediction and the observed data. In conventional (one-class) regression, this is easily accomplished by evaluating a loss function, for example the mean squared error between observed and predicted values. In an FMR setting however, the assignment of a single-valued prediction is a non-trivial task. Simply assigning one of

the class means does not seem justified, especially when we are assuming constant class proportions for all $\boldsymbol{x}$. A weighted average of the class means would also be a dubious route as such a prediction would not be close to the individual components (in general).

Therefore, instead of providing a single-value prediction, we consider a predictive distribution relying on the assumption of Gaussian errors. Each prediction is thus the density of the mixture distribution defined by the estimated model parameters. A measure of predictive ability is then obtained by evaluating the log-likelihood of the model on a new dataset $(\boldsymbol{x}_1^+, y_1^+), \ldots, (\boldsymbol{x}_{N^+}^+, y_{N^+}^+)$

$$\log \prod_{i=1}^{N^+} \hat{f}(y_i^+ | \boldsymbol{x}_i^+),$$

where $N^+$ is the sample size and $\hat{f}(y_i^+ | \boldsymbol{x}_i^+)$ is the estimated mixture density at $\boldsymbol{x}_i^+$. This approach, called *predictive likelihood* is utilized by Khalili and Chen (2007) for example. Higher values of the predictive likelihood indicate better fits. Note that contrarily to classical criteria like AIC and BIC, no penalty term is required since it is evaluated on new data.

### *3.2. Highest density region predictions*

To obtain an interval-type prediction for a given new observation $\boldsymbol{x}^+$ in the FMR setting, we rely on an estimated predictive distribution density $\hat{f}_+$. Obtaining such a predictive distribution however, requires some additional considerations on top of the model estimation. Simply plugging the point-estimates of the model parameters in to the model definition in (2), as is done for the predictive log-likelihood, leads to liberal intervals since the uncertainty related to the parameter estimates themselves is not accounted for. However, obtaining the variance-covariance matrix of the estimates (which would be needed for even an approximate interval estimate) is difficult to derive explicitly in the penalized FMR context. Here, we rely on the following bootstrap resampling (Efron and Tibshirani, 1994) routine to quantify the estimation uncertainty:

1. Fit the linear FMR model with $L_1$-penalty as done in step 5 of the algorithm proposed in section 2.3. For $b$ in $1, \ldots, B$ perform steps 2-4:
2. Take a bootstrap sample from the data used in the original model. In other words, $N$ observations are sampled randomly with replacement from the data.

3. Refit the FMR model using the same bootstrap sample and the value of the penalty parameter $\lambda$ that was found optimal in the original model.
4. Using the fitted model, find the estimated mixture density for $\boldsymbol{x}^+$ to obtain $\hat{f}_b(\boldsymbol{x}^+)$.
5. Average over the bootstrap estimates of the mixture density to obtain the predictive distribution $\hat{f}_+$:

$$\hat{f}_+ = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\boldsymbol{x}^+).$$

In the same way that the highest posterior density regions are often considered in Bayesian analysis (Gelman et al., 2014), we can find the highest density region of this predictive distribution:

$$\mathcal{A} = \left\{ a \in \mathbb{R} : \int_{\mathcal{A}} \hat{f}_+ = 1-\alpha \quad \cap \quad \hat{f}_+(a) > \hat{f}_+(b), \forall a \in \mathcal{A}, \forall b \in \mathcal{A}^c \right\}.$$

To put the above definition in words, the region contains a $1-\alpha$ proportion of the probability mass and the density $\hat{f}_+$ is always greater for points inside the region than for those outside. This is further illustrated in Figure 2. The region typically consists of up to $K$ disjoint intervals depending on the estimated model parameters, although in theory there could be up to $BK$ disjoint intervals due to the bootstrap sampling. The intervals formed this way contain an average proportion of $1-\alpha$ new observations generated through the same process as the training data.

This procedure leads to approximate $\alpha$-level intervals that account for the uncertainty in the estimates of the linear predictors. A relatively small bootstrap sample of $B = 10$ is used in the simulations. By using a fixed $\lambda$ in step 2, we greatly reduce the computation time needed for the bootstrapped predictions. Additional gains are made by using the posterior probabilities and the point estimates of $\sigma_1, \ldots, \sigma_K$ obtained in step 1 as starting values for the algorithm in step 2.

The calculation of the highest density region is not analytically possible when $\hat{f}_+$ is a mixture of Gaussians. Thus, we consider regions calculated using a Monte Carlo approach as described by Hyndman (1996) utilizing a sample of 5000 values from $\hat{f}_+$.

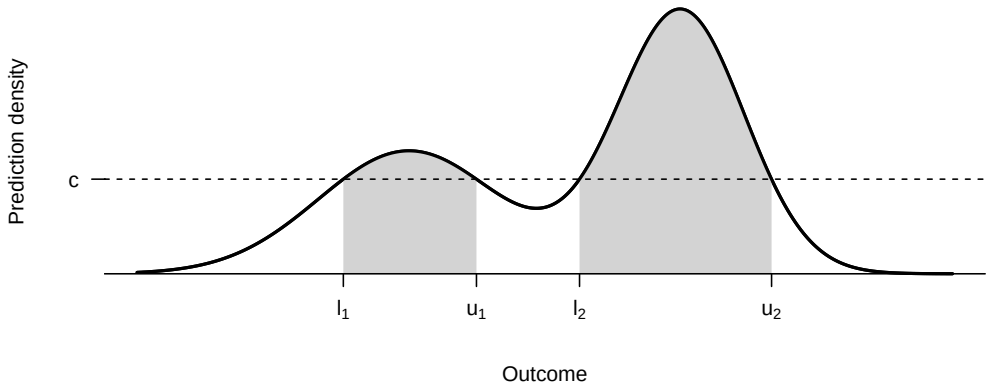



Figure 2: **Highest density region illustrated for mixture of two normals.** A constant $c$ is found such that the sum of gray areas equal the sought level $1-\alpha$. The boundaries of these gray areas determine the $\alpha$-level prediction intervals which in this case are $[l_1, u_1]$ and $[l_2, u_2]$.

## 4. Simulation studies

A set of simulation scenarios are considered that range from a simple linear relationship to more complex models. The simulation scenarios are defined in detail here, and they are further illustrated in Figure 3. We consider only cases with $K = 2$ or $K = 3$ and furthermore assume that $K$ is known *a priori* for simplicity.

(A) We consider a $p$-dimensional covariate $\boldsymbol{x} = (x_1, \ldots, x_p)'$ where each $x_i, i = 1, \ldots, p$ is sampled from the uniform distribution between 0 and 10 with pairwise expected Pearson correlations set to 0.2. The correlation structure is achieved via a Gaussian copula (Nelsen, 2007). The outcome has a mixture distribution of two densities:

$$f(y) = \pi\phi(y; \beta_{0,1} + \boldsymbol{\beta}_1'\boldsymbol{x}, \sigma) + (1-\pi)\phi(y; \beta_{0,2} + \boldsymbol{\beta}_2'\boldsymbol{x}, \sigma).$$

The parameters are fixed to

- $\beta_{0,1} = 0 \qquad \boldsymbol{\beta}_1 = (\beta_{1,1}, \ldots, \beta_{1,p})' = \frac{1}{p}(2, 2, \ldots, 2)$
- $\beta_{0,2} = 5 \qquad \boldsymbol{\beta}_2 = (\beta_{2,1}, \ldots, \beta_{2,p})' = \frac{1}{p}(-1, -1, \ldots, -1)$

- $\pi = 0.5 \qquad \sigma = 1.$

In this scenario, the outcome is linearly dependent on the covariates and a linear FMR, such as the model in (1), is expected to have optimal performance.

(B) Next we consider a quadratic relationship with the outcome. This scenario is identical to A in all ways except that in the second class, a transformation is applied to the covariate vector $\boldsymbol{x^B} = (x_1^2 - 10x_1, \ldots, x_p^2 - 10x_p)'$. The corresponding regression coefficients are set following the same pattern as in A.

(C) We consider a scenario with three classes based on both linear and nonlinear structures. Here, a third nonlinear structure is added to scenario B based on a transformed covariate $\boldsymbol{x^C} = (\sin\frac{1}{2}x_1, \ldots, \sin\frac{1}{2}x_p)'$. The regression coefficients for this third class are set as $\beta_{0,3} = 60$ and $\boldsymbol{\beta}_3 = (-10, 0, ..., 0)'$. The nonlinear form is thus introduced only in the first dimension of the covariate to keep the problem tractable with a reasonable sample size. The complete density is then

$$f(y) = \pi_1\phi(y; \beta_{0,1}+\boldsymbol{\beta}_1'\boldsymbol{x}, \sigma)+\pi_2\phi(y; \beta_{0,2}+\boldsymbol{\beta}_2'\boldsymbol{x^B}, \sigma)+\pi_3\phi(y; \beta_{0,3}+\boldsymbol{\beta}_3'\boldsymbol{x^C}, \sigma),$$

where $\pi_1 = \pi_2 = \pi_3 = 1/3$ and $\sigma = 1$.

(D) Finally, we generate a scenario where the outcome is defined as a function of interaction terms, rather than simple transformations of the original covariate variables. These interaction terms are introduced as a new covariate $\boldsymbol{x^D} = (x_1x_2, x_2x_3, \ldots, x_{p-1}x_p)'$. Similar to A and B, the outcome has a mixture distribution of two densities:

$$f(y) = \pi\phi(y; \beta_{0,1} + \boldsymbol{\beta}_1'\boldsymbol{x^D}, \sigma) + (1 - \pi)\phi(y; \beta_{0,2} + \boldsymbol{\beta}_2'\boldsymbol{x^D}, \sigma),$$

where the parameter values are set to:

- $\beta_{0,1} = -20 \qquad \boldsymbol{\beta}_1 = \frac{1}{2}\mathbf{1}_{p-1} = (1/2, \ldots, 1/2)'$
- $\beta_{0,2} = 0 \qquad \boldsymbol{\beta}_2 = -\frac{1}{2}\mathbf{1}_{p-1} = (-1/2, \ldots, -1/2)'$
- $\pi = 0.5 \qquad \sigma = 1.$

In addition to the $p$-dimensional covariate, a set of $q$ surplus variables are included in the data. These variables are sampled identically to the covariate but are not related to the outcome; their regression coefficients are set to 0. In the simulations, we experiment with all combinations of $p = 2, 5$ and $q = 0, 50$.

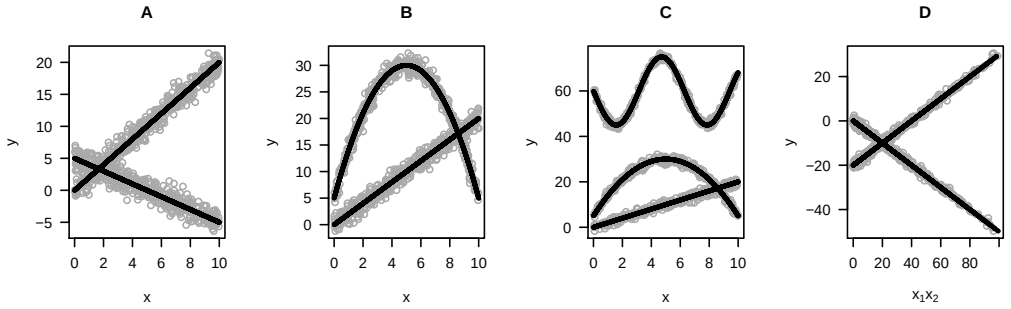



Figure 3: **Simulation scenarios illustrated in one dimension.** The black lines represent the class means and gray circles one realisation of the training set. The data is generated by setting $p = 1, q = 0$ for scenarios A, B and C, and $p = 2, q = 0$ for D. Simulation D is illustrated in terms of the interaction term $x_1x_2$.

In addition to FMRFLEX, we include four existing approaches that represent typical analyses one might choose in this setting:

- A one-class linear model (LM). A simple linear model is fitted on the data using all covariates and least-squares estimation.
- A one-class linear model with Lasso-penalty (LASSO).
- A finite mixture regression model with $K$ classes (FMR).
- A finite mixture regression model with $K$ classes and Lasso-penalty (FMRL). The optimal value for the penalty parameter is found using the BIC-criterion and a grid search over a 100 values as is done with the proposed FMRFLEX method.

We assume Gaussian errors for all models to make the evaluation of the predictive log-likelihood possible.

The different analysis approaches are fitted to training data with 500 $\times K$ observations after which the predictive log-likelihood and the highest density region coverages are calculated using a test set of 10000 observations. For each approach, the log-likelihood is compared to the true log-likelihood, calculated using the true underlying model that was used for generating the data. We also investigated the coverage probabilities and total lengths of the prediction intervals.

The simulations were repeated 100 times for each combination of settings. The covariates and the surplus variables are kept fixed after being sampled, whereas the outcome is resampled for each repetition.

### *4.1. Simulation results*

Examples of prediction intervals obtained both with the proposed method and with the linear FMR from the simulation scenarios A-C with one covariate ($p = 1, q = 0$) are illustrated in Figure 4. Fits for scenario D are also shown but with two covariates $p = 2$ and with the interaction term $x_1x_2$ in the $x$-axis. In scenario A, we see that the FMRFLEX method has clearly found the linear relationships in the data whereas in other scenarios the use of the dummy variables leads to more step-like behaviour. In scenarios B, C and D, the linear FMR provides approximate but highly biased estimates of the mean functions, and consequently, overly wide prediction intervals.

The predictive log-likelihoods from all simulation studies are collected in Figure 5. In all scenarios, we see that the one-class methods (LM, LASSO) provide the worst results as expected, displaying the highest difference to the true model in predictive log-likelihood. The proposed method (FMRFLEX) compares well with the linear alternatives (FMR, FMRL) in scenario A and even outperforms them when surplus variables are introduced ($q = 50$). Here, the method clearly benefits from the iterative approach that alternates between the class assignment and model fitting, which is not implemented in the standard FMR Lasso. Our proposed FRMFLEX method is clearly the top performer in all of the nonlinear scenarios (B, C and D). The conclusion is the same across all combinations of the number of covariates and dummy variables. The negative effect of surplus variables on the prediction performance of the proposed method is not substantial. When compared to the overall variation in the results, the variance across simulations is, with a few exceptions, relatively small demonstrating the stability of all of the methods.

The average coverage and total width of the 95% prediction intervals for the proposed method FMRFLEX and the best performing reference method FMRL are shown in Table 1. The observed coverage is generally close to the target probability of 95%. However, there is a tendency towards conservative intervals, especially in scenario C. This scenario is particularly challenging due to several partially overlapping components. A much larger data set would be required for better performance. More striking differences between the methods are seen in the average widths of the prediction intervals. The proposed method provides more precise predictions in the nonlinear scenarios

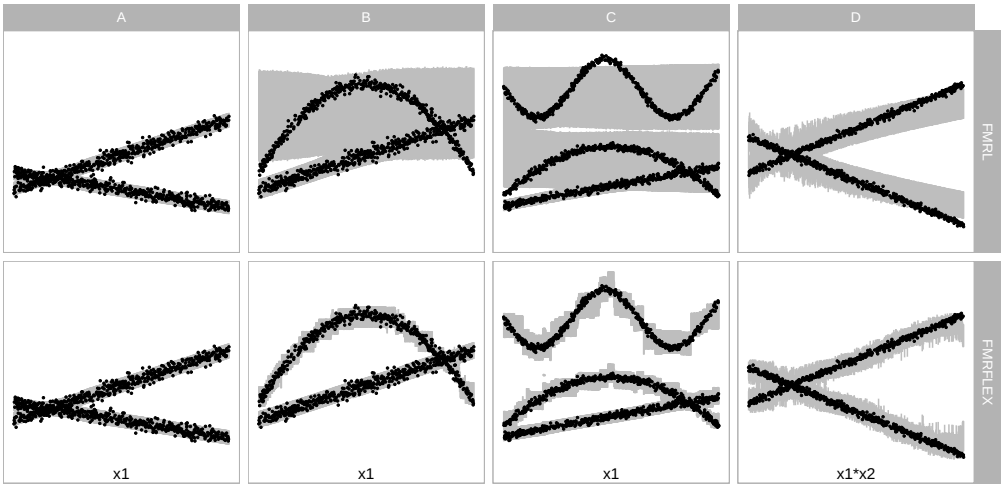



Figure 4: **Examples of the prediction intervals obtained with a linear FMR (top row) and the proposed method (bottom row)** The observed data is shown with black points on top of the gray prediction bands.

B-D as is seen in the relative average widths of the intervals (last column of Table 1). In scenarios B-D, the intervals provided by the proposed method are both more conservative and more accurate and thus are clearly preferable. Only the actual covered region is included in the width, leaving out any possible gaps between the classes.

The number of covariates and dummy variables finally selected by the proposed method depends on the complexity of the scenario as illustrated in Table 2. In the linear case of scenario A, typically no dummy variables were assigned with nonzero regression coefficients. The original $p$ covariates were almost always chosen. The number of dummy variables used grows along with the complexity of the scenarios as expected. There appears to be little difference between the results obtained with no surplus variables ($q = 0$) and those with noisy data ($q = 50$). This finding is in line with the overall simulation results and shows that the method performance is not overly affected by noise.

The computation times for the method greatly depend on the complexity of the problem at hand, mainly on the number of classes. Examples of computation times for $p = 5$ are shown in Table 3. The effects of addi-

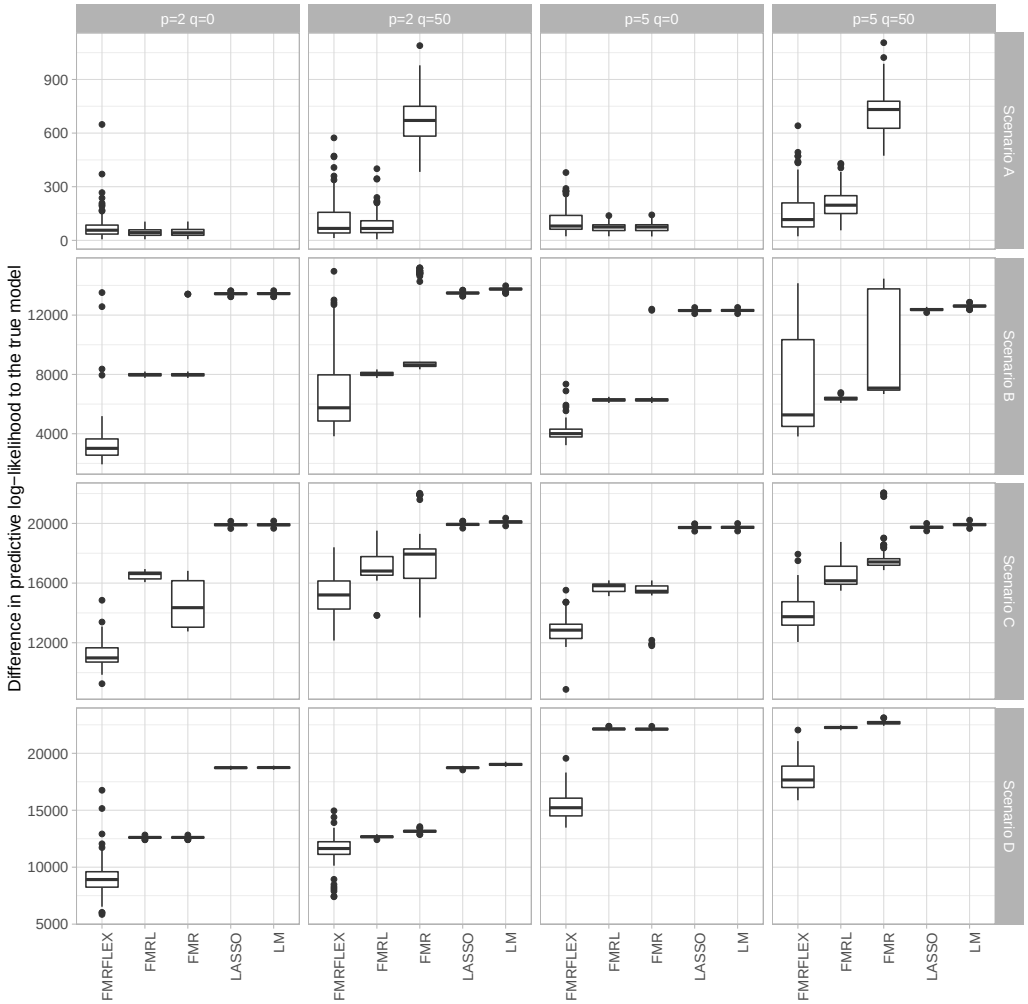



Figure 5: **Simulation results.** Differences in predictive log-likelihood to the true model in all simulation scenarios. Smaller values are better. In all scenarios of A and in those scenarios of D with $p = 5$, the results for the one-class models LM and LASSO are off plot limits.

tional surplus variables and deeper trees are also evident and tend to increase the time needed. Overall, FMRFLEX is much more time consuming than FMRL, however not infeasible for most applications. The measurements were obtained using a single core on an Intel®Xeon®X5450 3.00GHz CPU.

| scenario | $p$ | $q$ | FMRL coverage | FMRFLEX coverage | FMRL length | FMRFLEX length | relative length |
|---|---|---|---|---|---|---|---|
| A | 2 | 0 | 0.948 | 0.954 | 7.3 | 7.6 | 1.039 |
| A | 2 | 50 | 0.948 | 0.956 | 7.3 | 7.6 | 1.035 |
| A | 5 | 0 | 0.946 | 0.953 | 7.5 | 7.7 | 1.028 |
| A | 5 | 50 | 0.951 | 0.953 | 7.7 | 7.7 | 1.000 |
| B | 2 | 0 | 0.942 | 0.948 | 20.8 | 9.4 | 0.454 |
| B | 2 | 50 | 0.943 | 0.934 | 21.0 | 12.9 | 0.616 |
| B | 5 | 0 | 0.955 | 0.947 | 16.8 | 10.5 | 0.625 |
| B | 5 | 50 | 0.956 | 0.942 | 17.0 | 14.3 | 0.846 |
| C | 2 | 0 | 0.973 | 0.986 | 65.3 | 35.0 | 0.536 |
| C | 2 | 50 | 0.959 | 0.965 | 61.1 | 41.0 | 0.671 |
| C | 5 | 0 | 0.974 | 0.990 | 61.2 | 41.9 | 0.684 |
| C | 5 | 50 | 0.971 | 0.981 | 60.6 | 40.5 | 0.667 |
| D | 2 | 0 | 0.949 | 0.983 | 26.5 | 20.2 | 0.761 |
| D | 2 | 50 | 0.953 | 0.973 | 27.0 | 21.6 | 0.803 |
| D | 5 | 0 | 0.949 | 0.966 | 73.3 | 38.5 | 0.526 |
| D | 5 | 50 | 0.951 | 0.958 | 75.1 | 45.0 | 0.599 |

Table 1: Average prediction interval coverage probabilities and total lengths. The last column displays the quotient of the two previous columns.

## 5. Application to a worker wage dataset

To demonstrate the usage of the method in practice, we refer to a population survey dataset on the wages of male workers in the Mid-Atlantic region of the USA. This dataset contains 3000 subjects and is included in the *ISLR* R package (James et al., 2013). The goal of the analysis is to predict the worker's wage based on his age (continuous), job class (binary), health class (binary) and education level (3 classes). The wage is measured in thousands of dollars per year.

The analysis starts by fitting an ordinary one-class random forest regression model on the data. As seen in Figure 6, the residuals of this model are distributed around two centers, most of the density being close to 0 and a smaller portion near 150. This suggests that the model predicts well for most of the data but a finite mixture approach would be needed to account for the group of exceptionally high wages. In order to compare various model candidates, the data were split randomly into two equally sized sets of training and test data. Both the proposed flexible method FMRFLEX and the penalized linear alternative FMRL were tested with varying number of latent classes: $k = 1, 2, 3$.

| Scenario | $p$ | $q$ | Number of covariates kept | Number of dummies kept |
|---|---|---|---|---|
| A | 2 | 0 | 2 (2 - 2) | 0 (0 - 1) |
| A | 2 | 50 | 2 (2 - 2) | 1 (0 - 1) |
| A | 5 | 0 | 5 (5 - 5) | 0 (0 - 1) |
| A | 5 | 50 | 5 (5 - 5) | 0 (0 - 1) |
| B | 2 | 0 | 2 (2 - 2) | 54.5 (49 - 59) |
| B | 2 | 50 | 2 (2 - 2) | 54 (46 - 66) |
| B | 5 | 0 | 5 (5 - 5) | 45 (40 - 50) |
| B | 5 | 50 | 5 (5 - 6) | 42 (38 - 48) |
| C | 2 | 0 | 2 (0 - 2) | 55.5 (40 - 65) |
| C | 2 | 50 | 0 (0 - 2) | 60 (45 - 71) |
| C | 5 | 0 | 0 (0 - 0) | 55 (47 - 63) |
| C | 5 | 50 | 0 (0 - 1) | 65 (49 - 72) |
| D | 2 | 0 | 2 (2 - 2) | 79 (65 - 95) |
| D | 2 | 50 | 2 (2 - 2) | 89.5 (77 - 100) |
| D | 5 | 0 | 5 (5 - 5) | 112.5 (92 - 130) |
| D | 5 | 50 | 7 (6 - 8) | 100 (88 - 117) |

Table 2: Median number(interquartile range) of covariates and dummy variables kept in the final model over the simulation runs.

For the proposed method, the model with lowest BIC was achieved with two latent classes and by setting the minobs parameter to 0.05 yielding a predictive loglikelihood of $-7328.6$ when applied to the test data. This is less than the smallest BIC of $-7357.5$ for the lasso-penalized FMR which was also achieved with two latent classes. The chosen model utilizes 13 dummy variables and only the age covariate in its original form. This suggests that nonlinear relationships and/or interactions between variables need to be addressed in order to tackle the prediction problem. The prediction interval coverage in the test data is measured to be 94.9%, being almost identical with the theoretical value. The estimated class probabilities for the two classes are 0.904 and 0.096 which is in agreement with our earlier impression of the data based on Figure 6. For further analysis, the model was refitted on the complete data. The larger training sample resulted in an increased number of 20 dummy variables.

Interpreting the model results is not straightforward, but the point prediction curves displayed in Figure 7 are particularly revealing. Here we only consider the larger low-wage latent class as the sample size in the smaller high-wage class was found too small to derive any meaningful inference. The method was thus utilized more as an automated filtering tool, rather than as a full description of the data. The fit of a one-class random forest (RF) is also

| scenario | $p$ | $q$ | FMRL | minobs=0.1 | minobs=0.05 |
|---|---|---|---|---|---|
| A | 5 | 0 | 0 min 15 sec | 3 min 29 sec | 3 min 20 sec |
| A | 5 | 50 | 0 min 22 sec | 5 min 8 sec | 7 min 51 sec |
| B | 5 | 0 | 0 min 9 sec | 7 min 14 sec | 8 min 23 sec |
| B | 5 | 50 | 0 min 23 sec | 9 min 10 sec | 19 min 22 sec |
| D | 5 | 0 | 0 min 21 sec | 38 min 40 sec | 30 min 32 sec |
| D | 5 | 50 | 0 min 26 sec | 26 min 7 sec | 20 min 49 sec |
| C | 5 | 0 | 1 min 20 sec | 34 min 41 sec | 46 min 23 sec |
| C | 5 | 50 | 3 min 27 sec | 84 min 6 sec | 108 min 22 sec |

Table 3: Average computation times for the method based on five runs of the simulation scenarios.

shown for comparison. A non-linear age effect is seen in the data explaining the inclusion of dummy variables in the model. Wages increase with age for younger workers after which they remain more or less level. Unsurprisingly, increased education predicts higher average wages but is not independent of the job class. With an advanced degree the highest average salaries are made in the information segment, while with a lower education the average wages are higher in industry. The one-class random forest seems to perform relatively well for these data despite the group of high-wage outliers. Some bias is observed in the younger ages where RF tends to overestimate the wages considerably, but overall, the estimated mean components are very similar between the two models. However, a significant difference is seen in the lengths of their 95% prediction intervals: In a 5-fold cross-validation scheme, the median interval length for the test data using FMRFLEX is 124.7 (median absolute difference = 15.2) while RF gives 138.3 (15.0). The overall coverage of the intervals is comparable, 94.3% for FMRFLEX and 95.2% for RF. Overall, these results lead us to conclude that by using FMRFLEX, one would obtain more accurate predictions for these data, not necessarily in terms of less biased mean estimates but in terms of narrower prediction intervals.

The contextual difference between the two latent classes remains unknown but one can hypothesize that the smaller class of higher wages represents those who have reached leader positions in their workplace. Notably, most of the predictions are not particularly nonlinear, except for the younger ages. This nonlinearity was still enough to justify the selection of the proposed method over the linear alternatives.

To further demonstrate the utility of the method, we construct prediction intervals for two selected professionals. Consider a 25-year old worker

with no college degree, good health and working in an industrial job. The corresponding 95% prediction interval for his wage is $(26.18 - 136.06)$. In contrast, a 40-year old with an advanced degree working in the information field has a predicted wage of $(61.8 - 200.7)$ or $(239.9 - 278.4)$. In addition to just reporting the overall coverage of the two disjoint intervals, one can also place weights on the intervals based on the amount of probability mass of the predictive distribution they contain. In this case, the probability assigned for the first interval equals 0.930 while only 0.021 is placed on the second, reflecting the latent class proportions in the data.

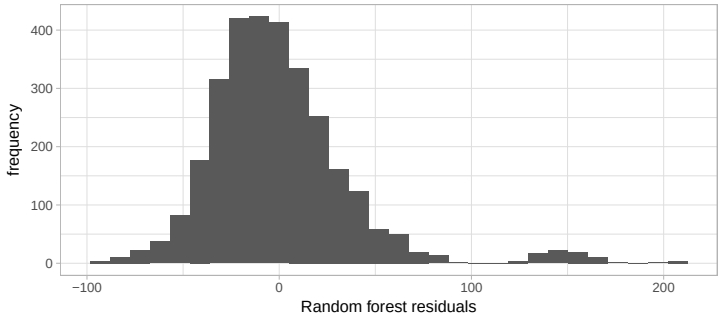



Figure 6: **Residuals of the random forest regression model fitted to the wage data.** Two distinct densities are seen in the distribution.

## 6. Discussion

We have introduced a method for flexible finite mixture regression that combines a random forest learner with a Lasso-penalized finite mixture regression model. We have shown using simulation studies that the method achieves equal performance with existing methods when the true model is in fact linear and superior performance in nonlinear cases.

The method was successfully applied to a wage prediction dataset, which contained two unbalanced latent classes. Superior predictive power compared to the linear alternatives was achieved most likely because the data contained nonlinear relationships and/or interactions between variables that were important for the prediction problem.

A drawback of our proposed FMRFLEX method is that while focusing on improved prediction, we lose the ability to do model-based clustering, which

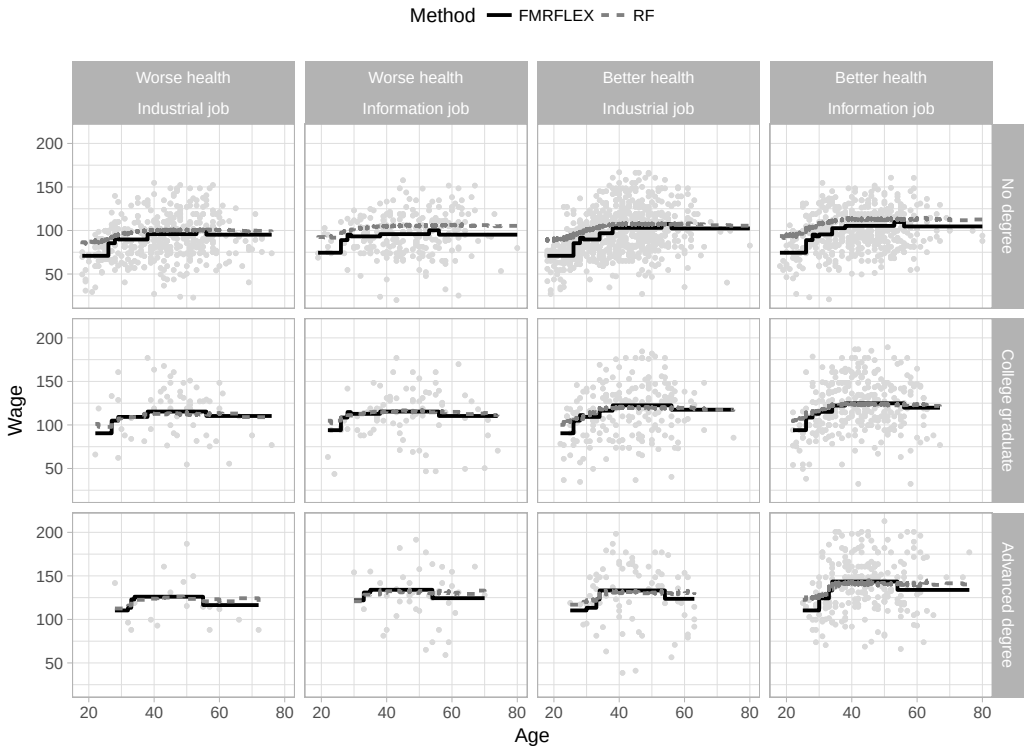



Figure 7: **Predictions of the FMRFLEX method on the wage data.** The plot is split into a grid where varying degrees of education define the rows and the columns correspond to the health and industry variables. Each plot in the grid displays the predictions for those data included in the larger latent class. The two lines correspond to the FMRFLEX and an ordinary one-class random forest fitted on the complete data. The observed data is shown as gray points.

is sometimes the goal of FMR analyses. Even though the clusters obtained by our method are typically relevant and even match the underlying model in the simulation studies, this is not guaranteed. There is no reason to expect that observations originating from the same latent class but differing in their dummy variable arrangement would be clustered together by the method. Furthermore, the balanced class proportions and equal residual variances used in the scenarios result in situations where the class distributions differ only in location. This issue is somewhat alleviated by the first step of the method where the initial classes are assigned using the standard linear FMR

model. This often results in a sensible initial clustering unless the underlying model is particularly nonlinear.

Another loss compared to standard FMR models is the difficulty of interpretation of the model parameters. In the presence of dummy variables, it is difficult to assess the net effect of a particular covariate in the model, especially since the dummy variables typically depend on multiple covariates. Luckily, both of the above-mentioned drawbacks relate only to nonlinear cases where dummy variables are used extensively. If the true association is in fact linear, the dummy variables do not typically get carried through the penalization and the method reduces to a linear FMR. If, however, the true association is nonlinear, a linear FMR is not valid anyway making its easy interpretation irrelevant.

The scope of the simulation studies covered cases with strong non-linearity and partly overlapping latent classes. Simulations in even more extreme settings could be performed to gain more information about the limitations of the proposed method. However, some assumptions can be made based on its structure. Due to the flexibility of the random forest learner, there is no fundamental reason to assume that the proposed method could not handle even severe cases of non-linearity or even discontinuous mean components. However, it is conceivable that weakly-separated latent classes combined with non-linearity could prove problematic. A particular concern in this case is the first step of the algorithm, where the initial class assignments are sought with a linear fit.

In this paper, we investigated exclusively a modified version of the original FMRFLEX idea for computational reasons. It is not guaranteed that the two alternatives perform identically when applied to the same set of data and in fact, we have observed that a small loss in accuracy due to the modification is possible. It is therefore not correct to treat these alternatives as two fully exchangeable methods. However, the magnitude of the difference in performance is negligible compared to the reduction in computational burden.

There exists many opportunities for future research around this topic. For example, an extension of the model to HMEs with covariate-dependent mixing proportions could be developed, as well as extensions for different types of outcome variables. Furthermore, the properties of BIC in selecting the number latent classes $K$ should be investigated in more detail.

A number of R (R Core Team, 2013) packages were used in different parts of the analysis. Random forests were obtained with *randomForest* (Liaw and

Wiener, 2002), the linear FMR with *flexmix* (Leisch, 2004) and the Lasso-penalized FMR with *fmrlasso* (Städler, 2010). Parallel simulation runs using multiple computing units were achieved with *snowfall* (Knaus, 2013). Finally, the *nor1mix* (Mächler, 2014) package was used for handling miscellaneous calculations involving univariate mixtures of Gaussians.

## 7. Acknowledgements

We thank the reviewers for their valuable comments that helped us greatly to improve the quality of the paper. We thank the Jenny and Antti Wihuri Foundation, Magnus Ehrnrooth Foundation, Turku University Foundation, the MATTI Graduate School, the Natural Sciences and Engineering Research Council of Canada (NSERC) and Fondation HEC Montreal for their financial support.

## References

Belloni, A., Chernozhukov, V., 2013. Least Squares After Model Selection in High-Dimensional Sparse Models. Bernoulli 19 (2), 521–547.

Bouveyron, C., Brunet-Saumard, C., 2014. Model-Based Clustering of High-Dimensional Data: A Review. Computational Statistics and Data Analysis 71, 52–78.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Bühlmann, P., Van De Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Science & Business Media.

Celeux, G., Martin-Magniette, M.-L., Maugis-Rabusseau, C., Raftery, A. E., 2014. Comparing Model Selection and Regularization Approaches to Variable Selection in Model-Based Clustering. Journal de la Societe Francaise de Statistique (2009) 155 (2), 57.

De Veaux, R. D., 1989. Mixtures of Linear Regressions. Computational Statistics & Data Analysis 8 (3), 227–245.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (methodological), 1–38.

Efron, B., Tibshirani, R. J., 1994. An Introduction to the Bootstrap. Chapman and Hall/CRC.

Feng, Y., Yu, Y., 2013. Consistent Cross-Validation for Tuning Parameter Selection in High-Dimensional Variable Selection. arXiv preprint arXiv:1308.5390.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33 (1), 1.

Galimberti, G., Montanari, A., Viroli, C., 2009. Penalized Factor Mixture Analysis for Variable Selection in Clustered Data. Computational Statistics & Data Analysis 53 (12), 4301–4310.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2014. Bayesian Data Analysis. Vol. 2. Taylor & Francis.

Hadavandi, E., Shahrabi, J., Hayashi, Y., 2016. SPMoE: a Novel Subspace-Projected Mixture of Experts Model for Multi-Target Regression Problems. Soft Computing 20 (5), 2047–2065.

Hastie, T., Tibshirani, R., Friedman, J., et al., 2009. The Elements of Statistical Learning, 2nd Edition. Springer.

Hennig, C., 2000. Identifiablity of Models for Clusterwise Linear Regression. Journal of Classification 17 (2), 273–296.

Huang, M., Li, R., Wang, H., Yao, W., 2013a. Estimating Mixture of Gaussian Processes by Kernel Smoothing. Journal of Business & Economic Statistics.

Huang, M., Li, R., Wang, S., 2013b. Nonparametric Mixture of Regression Models. Journal of the American Statistical Association 108 (503), 929–941.

Hyndman, R. J., 1996. Computing and Graphing Highest Density Regions. The American Statistician 50 (2), 120–126.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. ISLR: Data for An Introduction to Statistical Learning with Applications in R. R package

version 1.0.
URL http://CRAN.R-project.org/package=ISLR

Joly, A., Schnitzler, F., Geurts, P., Wehenkel, L., 2012. L1-Based Compression of Random Forest Models. In: 20th European Symposium on Artificial Neural Networks.

Khalili, A., 2010. New Estimation and Feature Selection Methods in Mixture-of-Experts Models. Canadian Journal of Statistics 38 (4), 519–539.

Khalili, A., Chen, J., 2007. Variable Selection in Finite Mixture of Regression Models. Journal of the American Statistical Association 102 (479).

Khalili, A., Chen, J., Lin, S., 2010. Feature Selection in Finite Mixture of Sparse Normal Linear Models in High-Dimensional Feature Space. Biostatistics.

Knaus, J., 2013. snowfall: Easier Cluster Computing (Based on snow). R package version 1.84-4.
URL http://CRAN.R-project.org/package=snowfall

Leisch, F., 2004. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. Journal of Statistical Software.
URL http://www.jstatsoft.org/v11/i08/

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2 (3), 18–22.
URL http://CRAN.R-project.org/doc/Rnews/

Mächler, M., 2014. nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods). R package version 1.2-0.
URL http://CRAN.R-project.org/package=nor1mix

McLachlan, G., Peel, D., 2004. Finite Mixture Models. John Wiley & Sons.

Nelsen, R. B., 2007. An Introduction to Copulas. Springer Science & Business Media.

Quandt, R. E., 1972. A New Approach to Estimating Switching Regressions. Journal of the American Statistical Association 67 (338), 306–310.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Spindler, M., 2014. Lasso for Instrumental Variable Selection: A Replication study. Journal of Applied Econometrics.

Städler, N., 2010. fmrlasso: Lasso for Finite Mixture of Regressions. R package version 1.0.

Städler, N., Bühlmann, P., Van De Geer, S., 2010. L1-penalization for Mixture Regression Models. Test 19 (2), 209–256.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

Xiang, S., 2014. Semiparametric Mixture Models. Ph.D. thesis, Kansas State University.

Yuksel, S. E., Wilson, J. N., Gader, P. D., 2012. Twenty Years of Mixture of Experts. IEEE Transactions on Neural Networks and Learning Systems 23 (8), 1177–1193.