Austrian Journal of Statistics
2025, Volume 54, 55–70.
http://www.ajs.or.at/
doi:10.17713/ajs.v54i3.2169

Simultaneous Inference in Finite Mixtures of Regression Models[1]

Friedrich Leisch
LMU München

Torsten Hothorn
LMU München

Abstract

A general framework for simultaneous inference in finite mixtures of generalized linear regression models is presented. Assuming asymptotic normality of the maximum likelihood estimate of all interesting model parameters, confidence regions and p-values using a maximum norm for the multivariate t-statistic are derived. This allows to simultaneously test all regression coefficients whether they are zero. Another application is to test for constant effects across mixture components. Size and power of the new methods are evaluated using artificial data. A real world data set on the productivity of PhD students is used to demonstrate the application of the procedures.

Keywords: R, simultaneous inference, finite mixture model, EM algorithm, latent class regression.

1. Introduction

We consider finite mixture models with K components of form

$$h(y|\mathbf{x}, \mathbf{P}, \mathbf{B}, \mathbf{\Gamma}) = \sum_{k=1}^{K} \pi_k f(y|\mathbf{x}, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k) \tag{1}$$

$$\pi_k > 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

where y is a dependent variable with conditional density h, $\mathbf{x} \in \mathbb{R}^D$ is a D-dimensional vector of independent variables, π_k is the component proportion of component k, $\boldsymbol{\beta}_k$ is the component specific vector of regression coefficients and $\boldsymbol{\gamma}_k$ a component specific vector of nuisance parameters for density function f. Further, let $\mathbf{P} = (\pi_1, \ldots, \pi_K)^\top$ be the vector of all component proportions, $\mathbf{B} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$ the vector of all regression coefficients and $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_K^\top)^\top$ the vector of all other parameters in the model. The reason for splitting

[1]This manuscript was written between 2010 and 2011 while both authors taught at LMU München. It has not been published before and is printed here with only minor editorial changes (typos, table formatting, addressing feedback from the review process, etc.). Hothorn (2025) explains the history and reflects on issues regarding reproducibility of numerical results using this work as an example.

the set of all parameters into vectors $\mathbf{P}$, $\mathbf{B}$ and $\mathbf{\Gamma}$ is that we are only interested in inference on $\mathbf{B}$ in this paper.

If f is a normal density with component-specific mean $\boldsymbol{\beta}_k^\top \mathbf{x}$ and variance σ_k^2, we have $\boldsymbol{\gamma}_k = (\sigma_k^2)$ as component specific nuisance parameters and Equation (1) describes a mixture of standard linear regression models, also called latent class regression. If f is another member of the exponential family (binomial, gamma, Poisson, ...), we get a mixture of generalized linear models (Wedel and DeSarbo 1995; Wang, Puterman, Cockburn, and Le 1996).

The corresponding log-likelihood cannot be maximized analytically and hence numerical methods have to be used. Only in the simplest cases direct optimization using gradient descent methods is feasible, and the most popular method for maximum likelihood estimation of the parameters are variations of the EM algorithm (Dempster, Laird, and Rubin 1977). Below we use the EM algorithm to find a maximum of the log-likelihood. After EM the full log-likelihood of all components is used to obtain a numerical estimate of the covariance matrix of all model parameters. We refer to Grün and Leisch (2008a) for a recent introduction to EM estimation of mixtures of GLMs and identification problems of the model class. Bayesian estimation of GLM mixtures is shown, e.g., in Lenk and DeSarbo (2000) or Frühwirth-Schnatter (2006), but will not be considered below.

A lot of research effort has been devoted to the selection of the right number of mixture components K, also called order selection. In a maximum likelihood (ML) context the most popular method is probably usage of an information criterion like AIC or BIC, see also the introduction of Chen and Khalili (2009) for an overview.

In this manuscript we are not concerned with order selection, but with inference on the regression coefficients of a mixture model with fixed number of components. We will adapt general theory for linear hypothesis tests to finite mixture models and demonstrate them on two important goals of inference procedures:

- Which coefficients are zero such that the corresponding independent variable makes no significant contribution to the respective component?
- Do the coefficients for a single independent variable have the same value over two or more components?

Both types of tests have immediate implications for model selection and interpretation. If coefficients do not significantly differ from zero, one may consider omitting them from the model. If a coefficient has the same value in two or more components, we can reduce the number of estimated parameters by fixing the parameter across these components.

2. Simultaneous inference procedures

In this section we present the underlying model assumptions and review some asymptotic results necessary in the subsequent sections. The concepts presented in this section form the basis for our new software implementation of simultaneous inference procedures. The set of n observations is denoted as $\{(y, \mathbf{x})_1, \ldots, (y, \mathbf{x})_n)\}$. The model contains fixed but unknown regression coefficients $\mathbf{B} \in \mathbb{R}^p$, where $p = K \cdot D$.

We are primarily interested in linear functions $\boldsymbol{\vartheta} := \mathbf{CB}$ of linear combinations of the parameter vector $\mathbf{B}$ as specified through the constant matrix $\mathbf{C} \in \mathbb{R}^{c \times p}$. We describe the underlying model assumptions, the limiting distribution of estimates of our parameters of interest $\boldsymbol{\vartheta}$, as well as the corresponding test statistics for hypotheses about $\boldsymbol{\vartheta}$ and their limiting joint distribution following Hothorn, Bretz, and Westfall (2008).

Suppose $\hat{\mathbf{B}}_n \in \mathbb{R}^p$ is, e.g., the ML estimate of the unknown true $\mathbf{B}_0$ and $\mathbf{S}_n \in \mathbb{R}^{p \times p}$ is an estimate of $\text{cov}(\hat{\mathbf{B}}_n)$ with

$$a_n \mathbf{S}_n \xrightarrow{\mathbb{P}} \Sigma \in \mathbb{R}^{p \times p} \tag{2}$$

for some positive, nondecreasing sequence a_n. Note that convergence rates and asymptotic properties of ML estimators are non-trivial for finite mixture models, see Zhu and Zhang (2004), Chen and Li (2009) and references therein. We assume in the following that all parameters are in the interior of the parameter space, we do not have too many components and the final estimate has been obtained by direct optimization of the full likelihood (see Section 3 below). Note that we only propose inference on the regression *coefficients*, hence avoiding the more problematic inference for mixture proportions (null hypothesis of $\pi_k = 0$ on boundary of parameter space) and variances (unbounded likelihood for decreasing variances).

Furthermore, we assume that a multivariate central limit theorem holds, i.e.,

$$a_n^{1/2}(\hat{\mathbf{B}}_n - \mathbf{B}_0) \xrightarrow{d} \mathcal{N}_p(0, \Sigma). \tag{3}$$

Assuming that (2) and (3) hold for the ML estimate $\hat{\mathbf{B}}_n$ we get $\hat{\mathbf{B}}_n \overset{a}{\sim} \mathcal{N}_p(\mathbf{B}_0, \mathbf{S}_n)$. Then, by Theorem 3.3.A in Serfling (1980), the linear function $\hat{\boldsymbol{\vartheta}}_n = \mathbf{C}\hat{\mathbf{B}}_n$, i.e., an estimate of our parameters of interest, also follows an approximate multivariate normal distribution

$$\hat{\boldsymbol{\vartheta}}_n = \mathbf{C}\hat{\mathbf{B}}_n \overset{a}{\sim} \mathcal{N}_c(\boldsymbol{\vartheta}_0, \mathbf{S}_n^\star)$$

with covariance matrix $\mathbf{S}_n^\star := \mathbf{C}\mathbf{S}_n\mathbf{C}^\top$ for any fixed matrix $\mathbf{C} \in \mathbb{R}^{c\times p}$. Thus we need not to distinguish between elemental parameters $\mathbf{B}$ or derived parameters $\boldsymbol{\vartheta} = \mathbf{C}\mathbf{B}$ that are of interest to the researcher. Instead we have (in analogy to (2) and (3))

$$\hat{\boldsymbol{\vartheta}}_n \overset{a}{\sim} \mathcal{N}_c(\boldsymbol{\vartheta}_0, \mathbf{S}_n^\star) \tag{4}$$

with

$$a_n\mathbf{S}_n^\star \xrightarrow{\mathbb{P}} \Sigma^\star := \mathbf{C}\Sigma\mathbf{C}^\top \in \mathbb{R}^{c\times c}$$

and that the c parameters in $\boldsymbol{\vartheta}$ are themselves the parameters of interest to the researcher. It is assumed that the diagonal elements of the covariance matrix are positive, i.e., $\Sigma_{jj}^\star > 0$ for $j = 1, \ldots, c$.

Then, the standardized estimator $\hat{\boldsymbol{\vartheta}}_n$ is again asymptotically normally distributed

$$\mathbf{T}_n := \mathbf{D}_n^{-1/2}(\hat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}_0) \overset{a}{\sim} \mathcal{N}_c(0, \mathbf{R}_n), \tag{5}$$

where $\mathbf{D}_n = \mathrm{diag}(\mathbf{S}_n^\star)$ is the diagonal matrix given by the diagonal elements of $\mathbf{S}_n^\star$ and

$$\mathbf{R}_n = \mathbf{D}_n^{-1/2}\mathbf{S}_n^\star\mathbf{D}_n^{-1/2} \in \mathbb{R}^{c\times c}$$

is the correlation matrix of the c-dimensional statistic $\mathbf{T}_n$. To finish note that

$$\begin{aligned} \mathbf{T}_n &= \mathbf{D}_n^{-1/2}(\hat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}_0) \\ &= (a_n\mathbf{D}_n)^{-1/2}a_n^{1/2}(\hat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}_0) \\ &\xrightarrow{d} \mathcal{N}_c(0, \mathbf{R}) \end{aligned}$$

with correlation matrix $\mathbf{R}$.

We now focus on the derivation of suitable inference procedures. We start considering the general linear hypothesis (Searle 1971) formulated in terms of our parameters of interest $\boldsymbol{\vartheta}$

$$H_0 : \boldsymbol{\vartheta} := \mathbf{C}\mathbf{B} = \mathbf{m}.$$

Under the conditions of H_0 it holds that

$$\mathbf{T}_n = \mathbf{D}_n^{-1/2}(\hat{\boldsymbol{\vartheta}}_n - \mathbf{m}) \overset{a}{\sim} \mathcal{N}_c(0, \mathbf{R}_n).$$

Note that a small global p-value leading to a rejection of H_0 does not give further indication about the nature of the significant result. Therefore, one is often interested in the individual null hypotheses

$$H_0^j : \vartheta_j = m_j.$$

Testing the hypotheses set $\{H_0^1, \ldots, H_0^c\}$ simultaneously thus requires the individual assessments while maintaining the familywise error rate.

A suitable scalar test statistic for testing the global hypothesis H_0 is to consider the maximum of the individual test statistics $T_{1,n}, \ldots, T_{c,n}$ of the multivariate statistic $\mathbf{T}_n = (T_{1,n}, \ldots, T_{c,n})$, leading to a max-$t$ type test statistic $\max(|\mathbf{T}_n|)$. The distribution of this statistic under the conditions of H_0 can be handled through the c-dimensional distribution

$$\begin{aligned} g_\nu(\mathbf{R}, t) &:= \mathsf{P}(\max(|\mathbf{T}_n|) \le t) \\ &\cong \int_{-t}^{t} \cdots \int_{-t}^{t} \varphi_c(x_1, \ldots, x_c; \mathbf{R}, \nu)\, dx_1 \cdots dx_c \end{aligned} \tag{6}$$

for some $t \in \mathbb{R}$, where φ_c is the density function of either the limiting c-dimensional multivariate normal (with $\nu = \infty$ and the '$\approx$' operator) or the exact multivariate $t_c(\nu, \mathbf{R})$-distribution (with $\nu < \infty$ and the '$=$' operator). Since $\mathbf{R}$ is usually unknown, we plug-in the consistent estimate $\mathbf{R}_n$. The resulting global p-value (exact or approximate, depending on context) for H_0 is $1 - g_\nu(\mathbf{R}_n, \max|\mathbf{t}|)$ when $\mathbf{T} = \mathbf{t}$ has been observed. Efficient methods for approximating the above multivariate normal and t integrals are described in Genz (1992); Genz and Bretz (1999); Bretz, Genz, and Hothorn (2001) and Genz and Bretz (2002).

This max-t type test based on the test statistic $\max(|\mathbf{T}_n|)$ also provides information, which of the c individual null hypotheses $H_0^j, j = 1, \ldots, c$ is significant. Consider testing the c null hypotheses $H_0^1, \ldots, H_0^c$ individually. We require that the familywise error rate, i.e., the probability of falsely rejecting at least one true null hypothesis, is bounded by the nominal significance level $\alpha \in (0, 1)$. In what follows we use adjusted p-values to describe the decision rules. Adjusted p-values are defined as the smallest significance level for which one still rejects an individual hypothesis H_0^j, given a particular multiple test procedure. In the present context of single-step tests, the (at least asymptotic) adjusted p-value for the jth individual two-sided hypothesis $H_0^j : \vartheta_j = m_j, j = 1, \ldots, c$, is given by

$$p_j = 1 - g_\nu(\mathbf{R}_n, |t_j|),$$

where $t_1, \ldots, t_c$ denote the observed test statistics. By construction, we can reject an individual null hypothesis H_0^j, $j = 1, \ldots, c$, whenever the associated adjusted p-value is less than or equal to the pre-specified significance level α, i.e., $p_j \le \alpha$. The adjusted p-values are calculated from expression (6).

Similar results also hold for one-sided testing problems. The adjusted p-values for one-sided cases are defined analogously, using one-sided multidimensional integrals instead of the two-sided integrals (6). Again, we refer to Genz (1992); Genz and Bretz (1999); Bretz *et al.* (2001) and Genz and Bretz (2002) for the numerical details. As specific examples for the general procedure we now give explicit definitions for C and m for the two most important groups of tests.

2.1. Tests for zero coefficients

Using $\mathbf{x} = (x_1, \ldots, x_D)^\top \in \mathbb{R}^D$ and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kD})^\top$ we get the linear predictor of the generalized linear regression model in component k

$$\eta_k = \boldsymbol{\beta}_k^\top \mathbf{x} = \sum_{d=1}^{D} \beta_{kd} x_d.$$

We are now interested in the $c = K \cdot D$ simultaneous tests

$$H_0 : \quad \beta_{kd} = 0, \qquad H_1 : \beta_{kd} \neq 0$$

for $k = 1, \ldots, K$ and $d = 1, \ldots, D$. In this case the contrast matrix $\mathbf{C}$ is an identity matrix of dimension $c \times c$ and $\mathbf{m}$ is a vector of c zeros.

2.2. Tests for constant effects

To test whether coefficients have the same value for two or more components we perform pairwise tests of equality on the set $\{\beta_{1d}, \ldots, \beta_{Kd}\}$. For fixed d we get $K \cdot (K-1)/2$ tests of form

$$H_0: \quad \beta_{kd} = \beta_{ld}, \qquad H_1: \beta_{kd} \neq \beta_{ld}$$

for $k, l = 1, \ldots, K$ and $k \neq l$. An equivalent set of hypotheses is of course given by

$$H_0: \quad \beta_{kd} - \beta_{ld} = 0, \qquad H_1: \beta_{kd} - \beta_{ld} \neq 0.$$

Overall we have D such groups of pairwise tests and hence a total of $c = D \cdot K \cdot (K-1)/2$ simultaneous tests. In this case we have again that $\mathbf{m}$ is a vector of zeros, and each row of $\mathbf{C}$ contains only zeros with exception of elements kd and ld, which are $+1$ and -1, respectively.

3. Size and power simulations

In order to empirically validate our theoretical results, we conducted a comprehensive series of simulation experiments for standard linear models with Gaussian noise and generalized linear models with Poisson dependent variable. We first describe the general design of both series of experiments, then discuss some computational aspects, and finish this section with size and power simulations for tests for zero coefficients and constant effects.

3.1. Simulation design

We consider a finite mixture of standard linear regression models with three components, intercept $x_1 \equiv 1$, and six standard uniform explanatory variables $x_2, \ldots, x_7 \sim U(0,1)$. With $\mathbf{x} = (x_1, x_2, \ldots, x_7)^\top = (1, x_2, \ldots, x_7)^\top$ we get the model for component k as

$$y = \boldsymbol{\beta}_k^\top \mathbf{x} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1/4).$$

The matrix of regression coefficients β_{kd} for the three components is

$$[\beta_{kd}] = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3] = \begin{bmatrix} 3 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 3 & 3 \\ 0 & 2 & 4 \\ 0 & 1 & 5 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix}.$$

Note that the first row corresponds to the intercepts, the remaining six rows are the coefficients for $x_2, \ldots, x_7$. If we stack the three columns of the matrix we get our vector $\mathbf{B}$ of all regression coefficients.

The set β_{kd} was chosen to contain 1/3 zeros (7 entries), available for size evaluation for the test of zero coefficients. The remaining 2/3 non-zero entries (14 in total) will be used to evaluate the power of the test. The asymmetry between the two groups (1/3 vs. 2/3) is deliberate to make room for size and power simulations for the pairwise tests of constant effects, where 3/7 (9 equal pairs) can be used to evaluate size and 4/7 (12 unequal pairs) to evaluate power. E.g., row one (intercept of the model) contains two unequal pairs (first versus second and third column) and one equal pair (second and third column). For both types of test we have more cases to evaluate power because we want to have different levels of disagreement with the null hypothesis.

In addition to the mixture of standard linear regression models we also use a mixture of generalized linear models (GLMs) with Poisson distribution for the dependent variable and log link. The regression coefficients are the same as above, the only difference is that the

distribution of the explanatory variables $x_2, \ldots, x_7$ is now $U(-0.5, 0.5)$. The reason for the shift is the log-link of the GLM, the sum of coefficients in the third component is 21, standard uniform x_i would result in a possible mean value of e^{21} for the Poisson dependent variable.

All computations in this paper were done using the statistical computing environment R (R Core Team 2024) with extension packages **flexmix** (Grün and Leisch 2025; Leisch 2004; Grün and Leisch 2008b) and **multcomp** (Hothorn, Bretz, and Westfall 2025; Hothorn *et al.* 2008). Samples from the data generating processes (DGPs) described above can be obtained using function `ExLinear()` in package **flexmix**.

3.2. Speeding up EM simulations

The two biggest disadvantages of the EM algorithm are its slow convergence and that it may get stuck in a local maximum of the likelihood. Both are problematic for size and power simulations, where we have to replicate the procedure for hundreds of times. Slow convergence can be overcome by brute force of computing power, but getting stuck in local minima will bias simulation results. When fitting the model to a real data set the analyst can manually check for convergence, restart if the algorithm did not converge, and also have a look at the differences between several restarts of the algorithm.

In simulations manual inspection is not an option, hence we have to take care that these problems occur as rarely as possible. To get a valid estimate of the covariance matrix of the model parameters *after convergence* only the distribution at convergence is of interest, it has no influence which path the EM algorithm took to get there. So in theory it should make no difference if we

1. start the EM algorithm using the true cluster memberships of the observations and then find the maximum likelihood estimate for a given new data set, or
2. repeatedly start the EM algorithm with a random initialization and keep the estimates with the best likelihood, assuming at least one run found the global maximum.

Assume $\mathbf{B}_0$ is the true parameter vector of the data generating process, and $\hat{\mathbf{B}}_n$ is the maximum likelihood estimate (MLE) for a given data set of size n. Then the distributions of $\hat{\mathbf{B}}_n$ and the difference

$$\delta(\mathbf{B}_0, \hat{\mathbf{B}}_n) = ||\mathbf{B}_0 - \hat{\mathbf{B}}_n||$$

for some appropriate norm $||\cdot||$ is only a function of n and the data generating process, not of the particular method used to obtain the MLE (assuming the method is capable of finding it).

Starting the EM algorithm in the true solution has several computational advantages:

1. usually only a few iterations are needed to get from $\mathbf{B}_0$ to the MLE $\hat{\mathbf{B}}_n$,
2. with very high probability we cannot get stuck in a local optimum, hence only one EM replication is needed, and
3. the components of $\hat{\mathbf{B}}_n$ have the same order as in $\mathbf{B}_0$, no relabeling is needed.

Despite the theoretical arguments given above we did an extensive simulation study to empirically confirm the arguments for our DGP to be on the safe side. We sample 1000 data sets from the process described above and fit mixture models to obtain regression coefficient estimates

$\hat{\mathbf{B}}^1$: start the EM algorithm with 20 random initializations, run each until convergence, keep solution with best likelihood and relabel components by order of coefficients β_{k4}.

$\hat{\mathbf{B}}^2$: initialize EM algorithm with true cluster memberships and run until convergence.

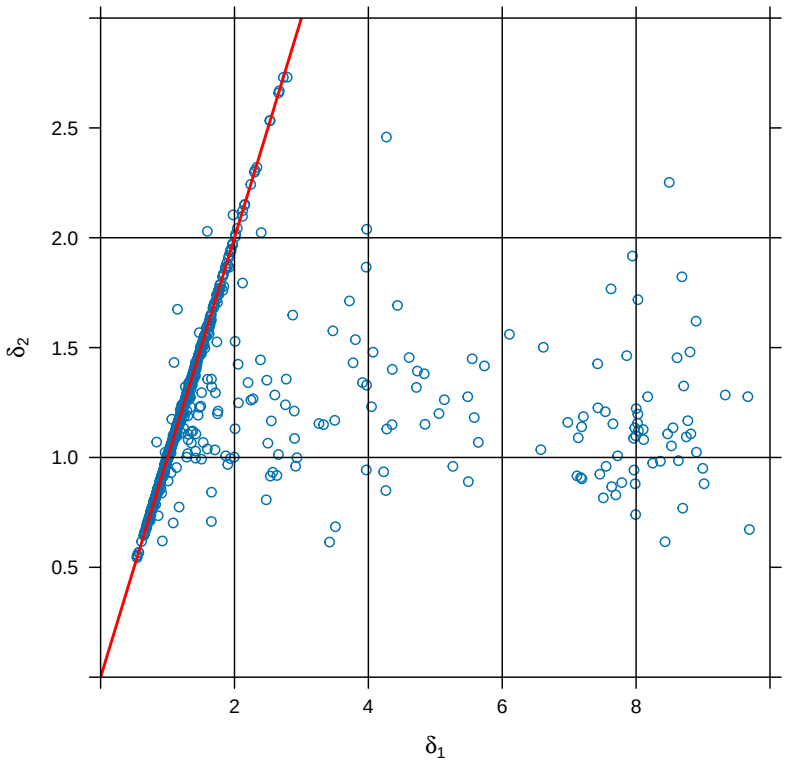

Figure 1: Scatterplot of δ_1 versus δ_2: in approx. 85% of all replications estimates $\hat{\mathbf{B}}^{\mathbf{1}}$ and $\hat{\mathbf{B}}^{\mathbf{2}}$ coincide

We then compute the Euclidean distance between the solutions found and the true matrix $\mathbf{B}_0$:

$$\begin{aligned} \delta_1 &= \delta(\hat{\mathbf{B}}^{\mathbf{1}}, \mathbf{B}_0) = ||\hat{\mathbf{B}}^{\mathbf{1}} - \mathbf{B}_0|| \\ \delta_2 &= \delta(\hat{\mathbf{B}}^{\mathbf{2}}, \mathbf{B}_0) = ||\hat{\mathbf{B}}^{\mathbf{2}} - \mathbf{B}_0|| \end{aligned}$$

Figure 1 shows a scatter plot of δ_1 versus δ_2 for the 1000 replications. The vast majority of points (approx. 85%) lie on a straight line passing through the origin with slope 1. Points with large δ_1 correspond to runs where 20 runs of EM were not enough to find the true solution.

Figure 2 shows the logarithm of δ_1 (left panel) and δ_2 (right panel) versus the replication number. A log scale was chosen to get a clear view of the majority of points and reduce the influence of outliers. It can clearly be seen that the variation within the horizontal bands at the bottom of both panels is identical, the only structural difference is the presence of 15% outliers in the left panel.

Summarizing both figures we conclude that starting the EM algorithm in the true solution preserves the variance of the MLE (the only quantity we are interested in this paper), but is a very effective measure against getting stuck in local maxima of the likelihood. A nice side effect is that this speeds up simulations on artificial data considerably.

3.3. Tests for zero coefficients

To test size and power of our simultaneous inference procedures we draw 10000 data sets of size $n = 300$ and $n = 600$ from the DGPs described above (with 100 or 200 observations in each of the three mixture components). The EM algorithm was initialized using true cluster memberships, then the covariance matrix of all model parameters was numerically estimated from the full likelihood of the model.

The model contains 21 regression coefficients (vector $\mathbf{B}$). Which of these are not significantly different from zero was tested using the simultaneous inference procedure. For comparison

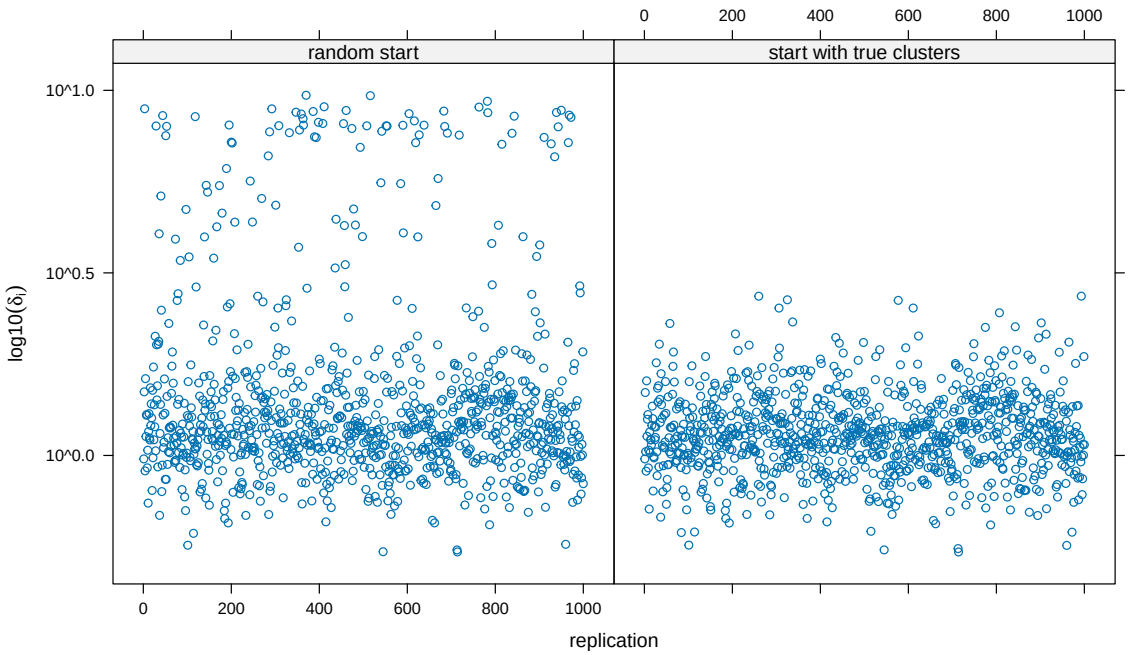

Figure 2: Scatterplot of δ_1 and δ_2 versus replication number

Table 1: Empirical size of tests for zero coefficients for a significance level of $\alpha = 0.05$

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
raw p-values	0.5029	0.3683	0.3292	0.3205
Holm-adj.	0.1804	0.0832	0.0709	0.0581
mult. comp.	0.1162	0.0384	0.0319	0.0230

Table 2: Empirical power of p-values from multiple comparisons procedure for zero coefficients

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
$\beta_{kd} = 1$	0.8162	0.9761	0.9888	1.0000
$\beta_{kd} > 1$	0.9997	1.0000	0.9999	1.0000
found all	0.4165	0.9058	0.9548	0.9998

Table 3: Empirical power of Holm-adjusted p-values for zero coefficients

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
$\beta_{kd} = 1$	0.8622	0.9854	0.9926	1.0000
$\beta_{kd} > 1$	0.9998	1.0000	1.0000	1.0000
found all	0.5358	0.9423	0.9701	1.0000

we also performed separate t-tests which parameters are zero, and adjusted the resulting 21 p-values for multiple testing using Holm's method (Holm 1979).

Table 1 shows the empirical size of the tests for a significance value of $\alpha = 0.05$. Seven of the 21 regression coefficients are zero, the table shows the percentage of replications were at least one of the seven p-values was smaller than 0.05. The first row shows (as reference only) unadjusted p-values of separate tests, these are not corrected for multiple testing and

Table 4: Empirical size and power of multiple comparisons procedure for constant effects

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
Size	0.1189	0.0457	0.0449	0.0271
Power	0.4470	0.8185	0.9397	0.9995

hence the size is much larger than 0.05 in all cases. Both the Holm-adjusted p-values and our simultaneous inference procedure have size problems for $n = 300$ and normal response. Note that we estimate approximately 240 parameters (regression coefficients, component proportions, covariance matrix of all parameters), using $n = 300$ was deliberately chosen as a borderline case. In all other cases (normal $n = 600$, Poisson $n = 300$ and $n = 600$) the simultaneous inference procedure is slightly conservative and has empirical size of 0.02–0.04. The Holm-adjusted p-values are above the nominal size in all cases.

Table 2 shows the power of the simultaneous inference procedure for tests of zero coefficients at a significance level of $\alpha = 0.05$. The first row shows how often β_{kd} with a value of one (two times the standard deviation of the noise) were found significant, the second row shows that β_{kd} that are larger than one were always significant. The last row shows the number of replications were all significant parameters were correctly identified. Except for the problematic case with normal response and $n = 300$ power is generally very good. Table 3 shows the power of Holm-adjusted p-values, which is slightly better than the simultaneous inference procedure. This was to be expected because the size of the Holm-adjusted p-values is larger, it tends to reject the null hypothesis of $\beta_{kd} = 0$ too often even if it is true.

3.4. Tests for constant effects

Next we use the simultaneous inference procedure to test for constant effects. Our simulation design contains nine pairs of parameters which are identical for different components, e.g., $\beta_{21} = \beta_{31} = 1$. These nine pairs can be used to empirically analyze the size of a test for constant effects, the remaining 12 pairs can be used to analyze power.

Table 4 shows the empirical size and power of our simultaneous tests for constant effects. The first row shows the size, i.e., in how many replications was at least one pair of identical coefficients $\beta_{kd} = \beta_{ld}$ wrongly identified to have a significant difference. The second row gives the power, i.e., in how many replications all unequal pairs $\beta_{kd} \neq \beta_{ld}$ were correctly identified. Results are similar to the tests for zero coefficients: except for the problematic case with normal response and $n = 300$ the tests are close to nominal size or slightly conservative, and power is very good.

4. Example: PhD students

In this example we use a sample of 915 biochemistry graduate students from Long (1990) with the following six variables:

art: integer, count of articles produced by the student during last 3 years of Ph.D.
female: nominal, gender of student (Male or Female)
married: nominal, marital status of student (Single or Married)
kid5: integer, number of children aged 5 or younger
phd: metric, prestige of Ph.D. department
ment: integer, count of articles produced by Ph.D. mentor during last 3 years

The goal is to predict the number of articles produced by the students from the other variables. The dependent variable is a count, hence we use a Poisson-GLM. Finite mixtures have become

Table 5: Summary statistics for the predicted number of papers per student in the two-component model

	Comp. 1	Comp. 2
Min.	0.48	1.68
1st Qu.	0.79	2.62
Median	0.95	3.14
Mean	1.04	3.59
3rd Qu.	1.15	3.84
Max.	5.16	32.93

Table 6: Test for significance of regression coefficients using standard asymptotic theory for the two-component model. Estimate, standard error and unadjusted p-value for each component.

	$\widehat{\boldsymbol{\beta_1}}$	SE($\widehat{\boldsymbol{\beta_1}}$)	p-value	$\widehat{\boldsymbol{\beta_2}}$	SE($\widehat{\boldsymbol{\beta_2}}$)	p-value
(Intercept)	−0.509	0.222	0.022	1.184	0.211	0.000
female	−0.140	0.102	0.170	−0.302	0.098	0.002
married	0.251	0.115	0.028	0.090	0.110	0.415
kid5	−0.212	0.074	0.004	−0.181	0.072	0.012
phd	0.092	0.050	0.067	−0.026	0.049	0.590
ment	0.025	0.003	0.000	0.032	0.004	0.000

a popular tool especially for Poisson-GLMs, because they can be used to account for zero-inflation or over-dispersion (e.g., Wang *et al.* 1996).

Fitting a two-component model to the data gives an AIC of 3148 and a BIC of 3211, for a three-component model we get an AIC of 3142 and a BIC of 3238. Hence, the AIC favors three components, the BIC two components, and we will have a look at both below. Models with more components are discarded by both information criteria.

First we have a look at the model with two components. Classifying students based on the a-posteriori probabilities, the first component contains 746 students, the second 169. For each component, the predicted number of papers per student is summarized in Table 5. Note that the mean number of publications in component 2 is three times higher than in component 1.

A test for significance of regression coefficients using standard asymptotic theory is shown in Table 6. Not surprisingly the number of papers per student is positively correlated with the number of papers by the respective mentor with very high significance in both components, similarly a negative correlation with the number of kids of age five and younger. Gender has no significant effect in component 1 (low number of publications), but has a significant effect in component 2, where women belonging to the component are seemingly less productive than men. Based on the classifications to components using the a-posteriori probabilities, the proportion of men and women in both components is very similar, with 47% females in component 1 and 41% females in component 2.

Correcting the tests for multiple testing shows that the significance of gender in component 2 is only borderline, see Table 7. There is also no significant difference between the coefficient for gender in components 1 and 2, see Table 8. Fitting a model with a constant effect for gender across both components gives no significant gender effect (details omitted for brevity).

The reason for the slightly significant gender effect in component 2 above can clearly be seen in the thee-component model with component sizes 314 (152 female), 588 (269 female), and 13 (0 female) based on the classifications obtained using the a-posteriori probabilities. For each component, the predicted number of papers per student is summarized in Table 9.

Table 7: Test for significance of regression coefficients using the simultaneous inference procedure for the two-component model

	$\widehat{\boldsymbol{\beta_1}}$	SE($\widehat{\boldsymbol{\beta_1}}$)	p-value	$\widehat{\boldsymbol{\beta_2}}$	SE($\widehat{\boldsymbol{\beta_2}}$)	p-value
(Intercept)	−0.509	0.222	0.209	1.184	0.211	0.000
female	−0.140	0.102	0.848	−0.302	0.098	0.023
married	0.251	0.115	0.264	0.090	0.110	0.995
kid5	−0.212	0.074	0.046	−0.181	0.072	0.120
phd	0.092	0.050	0.514	−0.026	0.049	1.000
ment	0.025	0.003	0.000	0.032	0.004	0.000

Table 8: Test for significant differences of regression coefficients for the two-component model

	Contrast	SE	p-value
C2.(Intercept)−C1.(Intercept) = 0	1.693	0.274	0.000
C2.female−C1.female = 0	−0.162	0.142	0.769
C2.married−C1.married = 0	−0.161	0.158	0.840
C2.kid5−C1.kid5 = 0	0.031	0.103	0.999
C2.phd−C1.phd = 0	−0.118	0.070	0.388
C2.ment−C1.ment = 0	0.006	0.005	0.708

Table 9: Summary statistics for the predicted number of papers per student in the three-component model

	Comp. 1	Comp. 2	Comp. 3
Min.	0.23	0.64	0.21
1st Qu.	0.51	0.88	0.45
Median	0.80	1.76	3.42
Mean	1.39	1.73	3.70
3rd Qu.	2.08	2.12	6.28
Max.	12.43	12.06	55.19

The model splits the students into three groups of low, medium and very high productivity (although the intercepts in components 1 and 2 are not significantly different). The last group is very small and consists of 13 outliers which happen to be all male. For the big mass of more than 900 remaining students we observe no gender effect, see Table 10. Note that we can estimate gender effects for component 3: Although no women are in this component with higher probability than in the first two, they still have positive probabilities to be there and hence contribute to the mixture likelihood.

What we do observe is a marriage effect. Components 1 and 2 both contain approximately twice as many married students than singles. The main difference between the components is that marriage has a negative impact on the number of publications in component 1, while it is positive in component 2. Besides removing the zero coefficients the three-component model could be further reduced by estimating a constant coefficient for the mentor, because the three coeffiecents are all approximately 0.026 and do not differ significantly, see Table 11.

5. Conclusions

We have presented a general framework for simultaneous inference in finite mixtures of regression models. The asymptotic normality of the maximum likelihood estimate of all interesting

Table 10: Test for significance of regression coefficients using the simultaneous inference procedure for the three-component model

	$\widehat{\boldsymbol{\beta_1}}$	SE($\widehat{\boldsymbol{\beta_1}}$)	p-value	$\widehat{\boldsymbol{\beta_2}}$	SE($\widehat{\boldsymbol{\beta_2}}$)	p-value	$\widehat{\boldsymbol{\beta_3}}$	SE($\widehat{\boldsymbol{\beta_3}}$)	p-value
(Intercept)	0.071	0.341	1.000	−0.373	0.223	0.762	2.166	0.536	0.001
female	−0.056	0.165	1.000	−0.021	0.109	1.000	−2.486	1.272	0.536
married	−1.261	0.317	0.001	1.037	0.235	0.000	0.322	0.385	0.999
kid5	−0.241	0.179	0.937	−0.162	0.062	0.130	0.011	0.256	1.000
phd	0.215	0.084	0.154	−0.011	0.044	1.000	−0.285	0.144	0.510
ment	0.027	0.006	0.000	0.026	0.004	0.000	0.026	0.009	0.047

Table 11: Test for significant differences of regression coefficients for the three-component model

	Contrast	SE	p-value
C2.(Intercept)−C1.(Intercept) = 0	−0.444	0.452	0.983
C3.(Intercept)−C1.(Intercept) = 0	2.095	0.635	0.014
C3.(Intercept)−C2.(Intercept) = 0	2.539	0.591	0.000
C2.female−C1.female = 0	0.035	0.185	1.000
C3.female−C1.female = 0	−2.430	1.312	0.533
C3.female−C2.female = 0	−2.465	1.293	0.492
C2.married−C1.married = 0	2.298	0.283	0.000
C3.married−C1.married = 0	1.583	0.526	0.036
C3.married−C2.married = 0	−0.715	0.461	0.758
C2.kid5−C1.kid5 = 0	0.079	0.200	1.000
C3.kid5−C1.kid5 = 0	0.252	0.314	0.996
C3.kid5−C2.kid5 = 0	0.173	0.252	0.999
C2.phd−C1.phd = 0	−0.227	0.100	0.260
C3.phd−C1.phd = 0	−0.501	0.161	0.027
C3.phd−C2.phd = 0	−0.274	0.154	0.588
C2.ment−C1.ment = 0	−0.001	0.007	1.000
C3.ment−C1.ment = 0	−0.000	0.011	1.000
C3.ment−C2.ment = 0	0.000	0.010	1.000

model parameters is assumed to derive confidence regions and p-values using a maximum norm for the multivariate t-statistic. The new methods are much closer to nominal significance levels in size simulations than classical p-values adjusted for multiple testing while losing almost no power. An example shows that interpretation of model parameters can be quite different when comparing the two approaches. The most important advantage of the new method is that the (possibly very high) correlations between different model parameters are correctly accounted for.

References

Bretz F, Genz A, Hothorn LA (2001). "On the Numerical Availability of Multiple Comparison Procedures." *Biometrical Journal*, **43**(5), 645–656. doi:10.1002/1521-4036(200109)43:5<645::AID-BIMJ645>3.0.CO;2-F.

Chen J, Khalili A (2009). "Order Selection in Finite Mixture Models with a Nonsmooth Penalty." *Journal of the American Statistical Association*, **104**(485), 187–196. doi:10.1198/jasa.2009.0103.

Chen J, Li P (2009). "Hypothesis Test for Normal Mixture Models: The EM Approach." *The Annals of Statistics*, **37**(5a), 2523–2542. doi:10.1214/08-aos651.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM-Alogrithm." *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–38. doi:10.1111/j.2517-6161.1977.tb01600.x.

Frühwirth-Schnatter S (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York, NY, U.S.A. doi:10.1007/978-0-387-35768-3.

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**(2), 141–149. doi:10.1080/10618600.1992.10477010.

Genz A, Bretz F (1999). "Numerical Computation of Multivariate *t*-probabilities with Application to Power Calculation of Multiple Contrasts." *Journal of Statistical Computation and Simulation*, **63**(4), 103–117. doi:10.1080/00949659908811962.

Genz A, Bretz F (2002). "Methods for the Computation of Multivariate *t*-probabilities." *Journal of Computational and Graphical Statistics*, **11**(4), 950–971. doi:10.1198/106186002394.

Grün B, Leisch F (2008a). "Finite Mixtures of Generalized Linear Regression Models." In Shalabh, C Heumann (eds.), *Recent Advances in Linear Models and Related Areas*. Physica Verlag, Heidelberg, Deutschland. doi:10.1007/978-3-7908-2064-5_11.

Grün B, Leisch F (2008b). "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software*, **28**(4), 1–35. doi:10.18637/jss.v028.i04.

Grün B, Leisch F (2025). ***flexmix**: Flexible Mixture Modeling*. doi:10.32614/CRAN.package.flexmix. R package version 2.3-20.

Holm S (1979). "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, **6**(2), 65–70.

Hothorn T (2025). "Did We Practice What We Preached?" *Austrian Journal of Statistics*.

Hothorn T, Bretz F, Westfall P (2008). "Simultaneous Inference in General Parametric Models." *Biometrical Journal*, **50**(3), 346–363. doi:10.1002/bimj.200810425.

Hothorn T, Bretz F, Westfall P (2025). ***multcomp**: Simultaneous Inference in General Parametric Models*. doi:10.32614/CRAN.package.multcomp. R package version 1.4-28.

Leisch F (2004). "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R." *Journal of Statistical Software*, **11**(8), 1–18. doi:10.18637/jss.v011.i08.

Leisch F, Hothorn T (2025). "Reproducibility Material for "Simultaneous Inference in Finite Mixtures of Regression Models"." doi:10.5281/zenodo.14950545.

Lenk PJ, DeSarbo WS (2000). "Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects." *Psychometrika*, **65**(1), 93–119. doi:10.1007/bf02294188.

Long JS (1990). "The Origins of Sex Differences in Science." *Social Forces*, **68**(4), 1297–1315. doi:10.2307/2579146.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Searle SR (1971). *Linear Models*. John Wiley & Sons, New York.

Serfling RJ (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York. doi:10.1002/9780470316481.

Wang P, Puterman ML, Cockburn IM, Le ND (1996). "Mixed Poisson Regression Models with Covariate Dependent Rates." *Biometrics*, **52**(2), 381–400. doi:10.2307/2532881.

Wedel M, DeSarbo WS (1995). "A Mixture Likelihood Approach for Generalized Linear Models." *Journal of Classification*, **12**(1), 21–55. doi:10.1007/bf01202266.

Zhu HT, Zhang H (2004). "Hypothesis Testing in Mixture Regression Models." *Journal of the Royal Statistical Society: Series B (Methodological)*, **66**(1), 3–16. doi:10.1046/j.1369-7412.2003.05379.x.

Reproducibility of simulation results

The main text contains simulation results computed at the time of the initial submission in 2010. The tables below report differences to simulation results obtained from re-running the simulation code in fall 2024. Old results are striked out and new results are printed sans-serif. The code required to reproduce the numerical results is available from Leisch and Hothorn (2025).

Table 12: Empirical size of tests for zero coefficients for a significance level of $\alpha = 0.05$; updated Table 1

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
raw p-values	~~0.5029~~ 0.5052	~~0.3683~~ 0.3726	~~0.3292~~ 0.3326	~~0.3205~~ 0.3123
Holm-adj.	~~0.1804~~ 0.1832	~~0.0832~~ 0.0853	~~0.0709~~ 0.0699	~~0.0581~~ 0.0531
mult. comp.	~~0.1162~~ 0.1163	~~0.0384~~ 0.0383	~~0.0319~~ 0.0344	~~0.0230~~0.0208

Table 13: Empirical power of p-values from multiple comparisons procedure for zero coefficients; updated Table 2

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
$\beta_{kd} = 1$	~~0.8162~~ 0.8130	~~0.9761~~ 0.9771	~~0.9888~~ 0.9877	~~1.0000~~0.9998
$\beta_{kd} > 1$	0.9997	1.0000	~~0.9999~~ 1.0000	1.0000
found all	~~0.4165~~ 0.4095	~~0.9058~~ 0.9094	~~0.9548~~ 0.9505	~~0.9998~~0.9994

Table 14: Empirical power of Holm-adjusted p-values for zero coefficients; updated Table 3

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
$\beta_{kd} = 1$	~~0.8622~~ 0.8626	~~0.9854~~ 0.9860	~~0.9926~~ 0.9918	1.0000
$\beta_{kd} > 1$	0.9998	1.0000	1.0000	1.0000
found all	~~0.5358~~ 0.5372	~~0.9423~~ 0.9447	~~0.9701~~ 0.9672	~~1.0000~~0.9998

Table 15: Empirical size and power of multiple comparisons procedure for constant effects; updated Table 4

	Normal		Poisson	
	$n = 300$	$n = 600$	$n = 300$	$n = 600$
Size	~~0.1189~~ 0.1179	~~0.0457~~ 0.0452	~~0.0449~~ 0.0458	~~0.0271~~0.0290
Power	~~0.4470~~ 0.4551	~~0.8185~~ 0.8167	~~0.9397~~ 0.9375	~~0.9995~~0.9998

The updated simulation results were obtained in the computing environment listed below.

```
R version 4.4.2 (2024-10-31)
Platform: x86_64-pc-linux-gnu
Running under: Debian GNU/Linux 12 (bookworm)

Matrix products: default
```

```
BLAS:   /usr/local/lib/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.11.0

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

time zone: Europe/Berlin
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] multcomp_1.4-28 TH.data_1.1-3   MASS_7.3-61     survival_3.8-3
[5] mvtnorm_1.3-3   flexmix_2.3-19  lattice_0.22-6

loaded via a namespace (and not attached):
 [1] codetools_0.2-20  Matrix_1.7-2      nnet_7.3-19       splines_4.4.2
 [5] modeltools_0.2-23 zoo_1.8-12        stats4_4.4.2      grid_4.4.2
 [9] sandwich_3.1-1    compiler_4.4.2    tools_4.4.2
```

Affiliation:

Friedrich Leisch, Torsten Hothorn
Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstraße 33, 80539 München, Deutschland