Noname manuscript No.
(will be inserted by the editor)

Least squares moment identification of binary regression mixture models

Benjamin Auder · **Elisabeth Gassiat** · **Mor Absa Loum**

Received: date / Accepted: date

Abstract We consider finite mixtures of generalized linear models with binary output. We prove that cross moments (between the output and the regression variables) up to order three are sufficient to identify all parameters of the model. We propose a least-squares estimation method based on those moments and we prove the consistency and the Gaussian asymptotic behavior of the estimator. We provide simulation results and comparisons with likelihood methods. Numerical experiments were conducted using the R-package morpheus that we developed for our least-squares moment method and with the R-package flexmix for likelihood methods. We then give some possible extensions to finite mixtures of regressions with binary output including both continuous and categorical covariates, and possibly longitudinal data.

Keywords Generalized linear model · Mixture Model · Moment method · Spectral method · Binary regression

Benjamin Auder
E-mail: benjamin.auder@universite-paris-saclay.fr
Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, 91405 Orsay, France

Elisabeth Gassiat
E-mail: elisabeth.gassiat@universite-paris-saclay.fr
Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, 91405 Orsay, France

Mor Absa Loum (corresponding author)
E-mail: morabsa.loum5@gmail.com
Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, 91405 Orsay, France

1 Introduction

Logistic models, or more generally multinomial regression models that fit covariates to discrete responses through a link function, are very popular for use in various application fields. When the data under study come from several latent groups that have different characteristics, using mixture models is also a very popular way to handle heterogeneity. Thus, many algorithms were developed to deal with various mixture models, see for instance the books [6] and [7]. Most of them use likelihood or Bayesian methods that are likelihood dependent. Indeed, the well known expectation-maximization (EM) methodology or its randomized versions makes it often easy to build algorithms. However, one problem of such methods is that they can converge to local spurious maxima so that it is necessary to explore multiple initial points as explained in [25]. This is one reason for renewed interest in moment based methods and for the recent development of so-called tensor methods mostly studied in the machine learning community. Tensor methods are based on the fact that in many latent variables models, low order moments contain enough information to make the inverse problem efficiently solvable, with a solution that uses simple computational steps; see [2] and references therein.

The main goal of this paper is to extend such moment methods in the setting of mixtures of binary regression models. Let us denote $\mathbb{R}$ the set of real numbers and $M^\top$ the transpose of a given matrix M. Let $X \in \mathbb{R}^d$ be the vector of covariates and $Y \in \{0, 1\}$ be the binary output. A binary regression model assumes that for some link function g, the probability that $Y = 1$ conditionally to $X = x$ is given by $g(\langle \beta, x\rangle + b)$, where $\beta \in \mathbb{R}^d$ is the vector of regression coefficients and $b \in \mathbb{R}$ is the intercept. Popular examples of link functions are the logit link function where for any real z, $g(z) = e^z/(1 + e^z)$ and the probit link function where $g(z) = \Phi(z)$, with Φ the cumulative distribution function of the standard normal $\mathcal{N}(0, 1)$. If now we want to modelise heterogeneous populations, let K be the number of populations and $\omega = (\omega_1, \cdots, \omega_K)$ their weights such that $\omega_j \geq 0$, $j = 1, \ldots, K$ and $\sum_{j=1}^K \omega_j = 1$. Define, for $j = 1, \ldots, K$, the regression coefficients in the j-th population by $\beta_j \in \mathbb{R}^d$ and the intercept in the j-th population by $b_j \in \mathbb{R}$. Let $\omega = (\omega_1, \ldots, \omega_K)$, $b = (b_1, \cdots, b_K)$, $\beta = [\beta_1|\cdots,|\beta_K]$ the $d \times K$ matrix of regression coefficients and denote $\theta = (\omega, \beta, b)$. The model of a population mixture of binary regressions is given by:

$$P_\theta(Y = 1|X = x) = \sum_{k=1}^{K} \omega_k g(\langle \beta_k, x\rangle + b_k). \tag{1}$$

Provable guarantees of estimation methods for mixtures of binary regressions are few, see Chapters 9 and 12 of [7] and references therein, or [22] and references therein. In particular in [22] it is proved that cross moments up to order 3 between the covariates and the output allow to recover the directions $\pm\beta_j/\|\beta_j\|$, $j = 1, \ldots, K$.

Our first main result is that cross moments up to order 3 between the output and the regression variables are enough to recover all the parameters of the model. This holds for the probit link function, see Theorem 1 and for general link functions under weak assumptions, see Theorem 2. These moment identifiability theorems are detailed in Section 2.

Our second contribution is a new least squares moment method to estimate the parameter θ with theoretical guarantees. The estimator $\widehat{\theta}$ is the one that minimizes a generalized square distance of theoretical moments to empirical moments. The estimation method is detailed in Section 3 together with the algorithm to compute the estimator. We developed the associated R-package `morpheus` available on the CRAN ([3]). Consistency and asymptotic normality of our least squares moment estimator are proved in Theorem 3. We then compare experimentally our method to the maximum likelihood estimator (computed using the R-package flexmix proposed in [8]). The experiments are presented in Section 4, where we give indications about the advantages or disadvantages of both methods.

We end the paper by investigating the identifiability in other population mixture models of binary regressions. For generalized linear models with outputs that are not binary, there is a large literature including estimation methods and practical algorithms such as in [4], [9], [12], [13], [14], [17], [18], [19], [21], [20];

see also the recent book [7] and references therein. However, the identifiability problem has not been fully explored in these articles and theoretical guarantees are largely missing. In [5], the identifiability is proved for finite mixtures of logistic regression models where only the intercept varies with the population. In [10], finite mixtures of multinomial logit models with varying and fixed effects are investigated, the proofs of the identifiability results use the explicit form of the logit function. In [24], further non parametric identifiability of the link function is proved, but only for models where the base exponential models are identifiable for mixtures, which does not apply to binary outcome. We provide in Section 6 several identifiability results which are useful as a first step to obtain theoretical guarantees in applications, such as in [15]. We prove that with a known smooth enough link function, the directions of the regression vectors may be recovered under the only assumption that they are distinct; see Theorem 4. Then, under the strengthened assumption that they are linearly independent, we prove that the link function may be recovered in a non parametric way ; see Theorem 5. We then study the simultaneous use of continuous and categorical covariates and further propose assumptions under which the parameters and the link function may be recovered; see Theorem 6. We finally prove that, with longitudinal data having at least 3 repetitions for each individual, the whole model is identifiable under the weakest assumption that the regression directions are distinct; see Theorem 7.

All the proofs of the different Theorems are deferred to Section 7.

2 Moment identifiability

We now define the cross moments that will be used to identify model (1).
Let us denote $[n]$ the set $\{1, 2, \ldots, n\}$ and $e_i \in \mathbb{R}^d$, the i-th canonical basis vector of $\mathbb{R}^d$. Denote also $I_d \in \mathbb{R}^{d\times d}$ the identity matrix in $\mathbb{R}^d$. The tensor product of p Euclidean spaces $\mathbb{R}^{d_i}$, $i \in [p]$ is noted $\bigotimes_{i=1}^p \mathbb{R}^{d_i}$. T is called a real p-th order tensor if $T \in \bigotimes_{i=1}^p \mathbb{R}^{d_i}$. For $p = 1$, T is a vector in $\mathbb{R}^{d_1}$ and for $p = 2$, T is a $d_1 \times d_2$ real matrix. The $(i_1, i_2, \ldots, i_p)$-th coordinate of T with respect the canonical basis is denoted $T[i_1, i_2, \ldots, i_p]$, $(i_1, i_2, \ldots, i_p) \in [d_1] \times \cdot \times [d_p]$. Define the cross moments between the scalar response Y and the covariable X:

- $M_1(\theta) := E_\theta[Y X]$,
- $M_2(\theta) := E_\theta\Big[Y\big(X \otimes X - \sum_{j\in[d]} e_j \otimes e_j\big)\Big]$, and
- $M_3(\theta) := E_\theta\Big[Y\big(X \otimes X \otimes X - \sum_{j\in[d]} \left[X \otimes e_j \otimes e_j + e_j \otimes X \otimes e_j + e_j \otimes e_j \otimes X\right]\big)\Big]$.

We assume that the random variable X has a Gaussian distribution. We now focus on the situation where $X \sim \mathcal{N}(0, I_d)$. All results may be easily extended to the situation where $X \sim \mathcal{N}(m, \Sigma)$, $m \in \mathbb{R}^d$, Σ a positive and symmetric $d \times d$ matrix by writing $X = m + \sigma^{1/2}\tilde{X}$ with $\tilde{X} \sim \mathcal{N}(0, I_d)$, and estimating empirically m and Σ. We shall explain below how the results can be extended to the case where X has any smooth enough distribution with support $\mathbb{R}^d$.

To prove our moment identifiability result, we shall use the following assumptions:

- (H1) The vectors $\beta_1, \ldots, \beta_K$ are linearly independent and the weights are positive: $\omega_k > 0$, $k = 1, \ldots, K$.
- (H2) The link function g is strictly increasing from 0 at $-\infty$ to 1 at $+\infty$, it has continuous derivatives till order 4, strictly decreasing first derivative on $[0, +\infty[$, and it satisfies

$$\forall z \in \mathbb{R},\ g(z) + g(-z) = 1.$$

- (H3) There exists a neighborhood $\mathcal{O}$ of $(0, 0)$ in $\mathbb{R}_+^\star \times \mathbb{R}$ and functions L_s, $s = 1, 2, 3$, such that for $s = 1, 2, 3$, $\int_R L_s(z)e^{-z^2/2}dz < +\infty$, and $\forall\ z \in \mathbb{R}$, $\forall\ (\lambda, b) \in \mathcal{O}$, we have

$$(|z| + 1)\left|\frac{\partial g^{(s+1)}}{\partial\lambda}(\lambda z + b)\right| \leq L_s(z).$$

Notice that (H1) implies that $d \geq K$, and that (H2) and (H3) hold in particular for the logistic link function and the probit link function.

Theorem 1 (Probit identifiability) *If (H1) holds and if g is the probit link function, one may recover K and $\theta = (\omega, \beta, b)$ from the knowledge of $M_1(\theta)$, $M_2(\theta)$ and $M_3(\theta)$.*

For general link functions, identifiability holds at least in an open set. The following identifiability result and Theorem 1 are proved in Section 7.1.

Theorem 2 (General identifiability) *If (H1), (H2), (H3) hold and $g^{(3)}(0) \neq 0$, there exist $L > 0$ and $B > 0$ such that as soon as $\|\beta_k\| < L$ and $|b_k| < B$ for all $k = 1, \dots, K$, then one may recover K and $\theta = (\omega, \beta, b)$ from the knowledge of $M_1(\theta)$, $M_2(\theta)$ and $M_3(\theta)$.*

Since the proof uses Taylor expansions, it only proves the existence of *small enough* positive L and B such that the result holds. For the logit function as link function g, we investigated numerically the function defined in (5) in the proof for which the property of being one-to-one is enough to deduce identifiability, and we found that Theorem 2 seems to be true at least with $L = 8$ and $B = 8$.

For population mixture models of linear regressions with continuous outcomes, it is only needed that the vectors of regression coefficients are distinct. However, for population mixture models of binary regressions, identifiability results exist only for the logit link function, see [5] and [10]. In Section 6 we prove that the directions of the regression vectors are identifiable under the only assumption that they are distinct. One consequence of Theorem 2 is that it provides provable guarantees for the maximum likelihood estimator so that in particular it is consistent.

The Gaussian assumption for the distribution of X may be relaxed by defining score-adapted cross-moments in the same way as in Section 4 of [22].

3 The least squares moment estimator

3.1 Definition of the estimator

In Section 2 we showed that the parameters can be recovered by matching the cross-moments till order 3. Those moments are unknown, so we estimate them empirically using:

$$\widehat{M_1} = \frac{1}{n} \sum_{i=1}^{n} Y_i X_i$$

$$\widehat{M_2} = \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i (X_i \otimes X_i - \sum_{j \in [d]} e_j \otimes e_j) \Big]$$

$$\widehat{M_3} = \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i (X_i \otimes X_i \otimes X_i - \sum_{j \in [d]} \big[X_i \otimes e_j \otimes e_j + e_j \otimes X_i \otimes e_j + e_j \otimes e_j \otimes X_i \big]) \Big].$$

It could be impossible to match the empirical moments exactly, so we use a least-squares estimator. Let Θ be the set of parameters, and define for all θ:

$$Q_n(\theta) = \sum_{j \in [d]} \Big\{ \widehat{M_1}[j] - M_1(\theta)[j] \Big\}^2 + \sum_{j,k \in [d]} \Big\{ \widehat{M_2}[j,k] - M_2(\theta)[j,k] \Big\}^2 + \sum_{j,k,l \in [d]} \Big\{ \widehat{M_3}[j,k,l] - M_3(\theta)[j,k,l] \Big\}^2.$$

The least squares moment method (LSMM) defines the estimator as follows:

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}}\ Q_n(\theta). \tag{2}$$

The terms in the sum to compute $Q_n(\theta)$ contribute unevenly to the result, because there are more indices combinations for higher order moments, and thus more (sub-)terms. Given this observation, pondering each term might lead to improvements in the estimation. Following [11] we propose the following generalized least squares moment method (GLSMM). Define for all θ and $i = 1, \dots, n$:

$$\tilde{M}_i(\theta) = Y_i \begin{pmatrix} X_i \\ X_i \otimes X_i - \sum_{j\in[d]} e_j \otimes e_j \\ X_i \otimes X_i \otimes X_i - \sum_{j\in[d]} \left[X_i \otimes e_j \otimes e_j + e_j \otimes X_i \otimes e_j + e_j \otimes e_j \otimes X_i\right] \end{pmatrix} - \begin{pmatrix} M_1(\theta) \\ M_2(\theta) \\ M_3(\theta) \end{pmatrix}$$

as a column vector, by flattening the tensorial products (for example put column j between indices $j \times d$ and $(j+1) \times d - 1$). Let W be a symmetric and positive definite square matrix of size $d + d^2 + d^3$, set

$$Q_n^W(\theta) = \left(\frac{1}{n}\sum_{i=1}^n \tilde{M}_i^\top(\theta)\right) W \left(\frac{1}{n}\sum_{i=1}^n \tilde{M}_i(\theta)\right),$$

and define

$$\widehat{\theta}_n^W = \operatorname*{argmin}_{\theta\in\Theta} Q_n^W(\theta). \tag{3}$$

Notice that when W is the identity matrix I_d, $\widehat{\theta}_n^{I_d} = \widehat{\theta}_n$.

3.2 Asymptotic properties

Define $\theta^\star$ as the true unknown parameter.
To prove the consistency of the estimator, identifiability of the model is obviously necessary. Thus, if the link function is not probit, then we choose Θ such that all parameters in Θ satisfy $\|\beta_k\| < L$ and $|b_k| < B$ for all $k = 1, \dots, K$, L and B being defined in Theorem 2.
To prove the limiting Gaussian distribution of our least-squares moment estimator, we shall need more assumptions. For $j = 1, \dots, 5$, let G_j be the $K \times K$ diagonal matrix having the $E[g^{(j)}(\langle\beta_k^\star, X\rangle + b_k^\star)]$'s on the diagonal.

- (H4) All diagonal coefficients of G_3 are non zero.
- (H5) All diagonal coefficients of $G_1G_3 - G_2^2$ are non zero.

Let $q = K(2+d) - 1$ be the dimension of θ. Let Z_n be the gradient vector of $Q_n(\theta)$ at $\theta = \theta^*$. Z_n is a linear combination of empirical moments, see the explicit formula in Section 7.2. Define $\Gamma(\theta^\star)$ the $q \times q$ matrix which is the variance matrix of $\sqrt{n}Z_n$. For each θ, define $\tilde{M}(\theta)$ as the expectation of $\tilde{M}_i(\theta)$, that is:

$$\tilde{M}(\theta) = \begin{pmatrix} M_1(\theta^\star) \\ M_2(\theta^\star) \\ M_3(\theta^\star) \end{pmatrix} - \begin{pmatrix} M_1(\theta) \\ M_2(\theta) \\ M_3(\theta) \end{pmatrix}$$

as a column vector in the same way as $\tilde{M}_i(\theta)$. Define also the $q \times q$ matrix $V(\theta)$ such that for all $r_1, r_2 = 1, \dots, q$:

$$V_{r_1r_2}(\theta) = 2\left(\frac{\partial\tilde{M}(\theta)^\top}{\partial\theta_{r_1}}\right) W \left(\frac{\partial\tilde{M}(\theta)}{\partial\partial\theta_{r_2}}\right).$$

As usual in mixture models, permutations over the populations do not change the value of the contrast function Q_n^W, so that permuted estimators are equivalent: this is the label-switching issue. To compare the estimator to the unknown true value, we thus in the statement of the following theorem and in its proof fix the permutation of the components of the estimator to the one that minimizes (over permutations of populations) the distance to the true parameter.

Theorem 3 *Assume that (H1), (H2), (H3) hold, that Θ is compact and that $\theta^\star \in \Theta$. Then $\widehat{\theta}_n^W$ is consistent.*
If moreover (H4) and (H5) hold, then $V(\theta^\star)$ is invertible, and $\sqrt{n}\left(\widehat{\theta}_n^W - \theta^\star\right)$ converges in distribution under $P_{\theta^\star}$ to a centered Gaussian distribution with variance $\Sigma(\theta^\star) = V(\theta^\star)^{-1}\Gamma(\theta^\star)V(\theta^\star)^{-1}$.

The proof of Theorem 3 is detailed in Section 7.2 and follows the usual analysis of the asymptotic behavior of Z-estimators, the more delicate part of the proof being to prove that the limiting value $V(\theta^\star)$ of the Hessian of $Q_n^W(\theta)$ is invertible.
To apply this Theorem one may estimate consistently $\Sigma(\theta^\star)$ by plug-in.

A natural question is the choice of the matrix W. It is proved in [11] that the matrix W minimizing the asymptotic variance of the estimator is given by $W(\theta^\star)$, with $W(\theta) = (\mathrm{E}[\tilde{M}_i(\theta)\tilde{M}_i^\top(\theta)])^{-1}$. This optimal matrix can be estimated empirically as follows:

$$\hat{W}(\hat{\theta}) = \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{M}_i(\hat{\theta})\tilde{M}_i^\top(\hat{\theta})\right)^{-1},$$

with $\hat{\theta}$ the parameters estimated from the LSMM for example, or another GLSMM (with another choice of the matrix W). $W(\theta)$ can also be re-estimated as many times as needed in a loop algorithm.

3.3 Algorithm

The estimator $\widehat{\theta}_n$ is computed by using the representation of the regression vectors β_k through their direction μ_k and norm λ_k, $k = 1, \ldots, K$. In a first step, we compute a preliminary estimate of $[\mu_1, \ldots, \mu_K]$ using a spectral method. In a second step, we minimize $Q_n^{W_{\mathtt{init}}}$ using usual optimization methods with $W_{\mathtt{init}} = Id$, or any user-defined matrix. W can then be recomputed using the optimized parameters, and a final minimization of Q_n^W follows. It would be possible to iterate this procedure again, but in practice it did not lead to improved results. The preliminary estimator for the directions is used as initial point for the directions in the optimization procedure.

Algorithm $M3LS$: Estimation of all parameters
input: X, Y, K, g
1) Estimate the directions $\mu_1, \ldots, \mu_K$ using Algorithm $InitDir$
2) Optimize $Q_n^{W_{\mathtt{init}}}(\theta)$ using the estimators of 1) as initial directions. Stop here if $W_{\mathtt{init}}$ is considered good enough
3) Re-compute W using the optimized parameters p, β and b
4) Execute step 2) again
Output: The estimated parameter $\widehat{\theta}$

The preliminary estimation of the directions is based on the spectral method. For any vector $z \in \mathbb{R}^p$, define $B(z)$ the $d \times d$ matrix such that

$$B(z)[i,j] := \sum_{s=1}^{d} M_3(\theta)[i,j,s] z_s,$$

so that, using Lemma 1 (presented in Section 7), we get

$$B(z) = \sum_{k=1}^{K} \omega_k \lambda_k^3 \mathbb{E}\left[g^{(3)}\big(\lambda_k \langle X, \mu_k\rangle + b_k\big)\right] \langle \mu_k, z\rangle \; \mu_k \otimes \mu_k \, .$$

It is proved in [1] that it is possible to recover the directions by joint diagonalisation of $B(z_1), \ldots B(z_P)$ for distinct vectors $z_1, \ldots, z_P$, $P \geq 2$ (In practice, we consider $z_1, \ldots, z_P$ as the canonical basis, i.e $P = d$). Joint diagonalisation of $B(z_1), \ldots B(z_P)$ means finding a matrix V such that the matrices $VB(z_p)V^\top$ are the most diagonal possible. The normalized vectors $\mu_1, \ldots, \mu_K$ are obtained up to sign and label switching by taking the first K vectors of V^{-1}. Let us denote U the matrix of these K vectors. Let $O = U_*^{-1} M_1(\theta) \in \mathbb{R}^K$, with U_*^{-1} the general inverse of U. The real numbers $\omega_k \lambda_k \mathbb{E}[g'\big(\lambda_k \langle X, \mu_k\rangle + b_k\big)]$, $k = 1, \ldots, K$, are given up to sign by the elements of O. Since they are positive, the sign of the μ_k's are obtained by multiplying -1 all the vectors associated to the negative values of O.

In pratice, the vectors $\mu_1, \ldots, \mu_K$ are estimated using the joint diagonalisation method applied to the matrices $\widehat{B}(z_p)$, $p = 1, \ldots, P$, computed using $\widehat{M_3}$.

Algorithm *InitDir*: Joint diagonalisation algorithm to estimate the directions

input: X, Y, K

1: Estimate the cross moments $\widehat{M_1}$, $\widehat{M_2}$ and $\widehat{M_3}$ as explained in section 3.2

2: Choose vectors $\{z_1, z_2 \ldots, z_P\} \subseteq \mathbb{R}^d$ (for instance: the canonical basis $e_1, e_2, \ldots, e_P$ of $\mathbb{R}^d$ if $P = d$)

3: Compute $\widehat{B}(z_p)$ for all $p \in \{1, 2, \ldots, P\}$

4: Joint diagonalisation: compute V such that $V\widehat{B}(z_p)V^\top$ are the most diagonal possible

5: Compute $U = V^{-1}[1 : K]$ the K-first vectors of V^{-1} (by ordering the diagonal values in decreasing absolute value)

6: Compute $O = \texttt{generalized\_inverse\_of}(U)\widehat{M_1}$

7: Multiply by -1 all the vectors of U corresponding to the negative values of O, $U[, O < 0] = -U[, O < 0]$

Output: The preliminary estimators of $\mu_1, \ldots, \mu_K$

4 Simulations

We first present the R-package implementing the generalized least-squares moment method (GLSMM) in Section 4.1. Then, in Section 4.2, the algorithms GLSMM and the maximum likelihood method (MLM) are compared in terms of performance (accuracy and computation time). We also compare in Section 4.3 the initialization as indicated in the introduction below, with an initialization based on random starts.

4.1 R **package**

The R-package is called `morpheus` [3] and is divided into two main parts:

1. the computation of the matrix μ (containing the normalized columns of β), based on the empirical cross-moments as described in section 3.3;
2. the optimization of all parameters (including β), using the initially estimated directions as a starting point.

The former is a straightforward translation of the mathematical formulas (file R/computeMu.R), while the latter calls R constrOptim() method on the objective function expression and its derivative (file R/optimParams.R). For usage examples, please refer to the package help.

4.2 **Comparison of the GLSMM and MLM algorithms**

In this section, we compare experimentally our GLSMM algorithm (morpheus package [3]) to the MLM algorithm (with flexmix package [8] which is a reference for this kind of estimation, using an iterative algorithm to maximize the log-likelihood). We present at first estimation errors and we compare computational time for the two different approaches. The parameters for the simulations are chosen arbitrarily as indicated below; these should be recovered by the algorithms (GLSMM, and the MLM algorithm).

Experiment 1 (2 dimensions):

$$K = 2,\ \omega = (0.5, 0.5),\ b = (-0.2, 0.5),\ \beta^\top = \begin{pmatrix} 1 & -2 \\ 3 & 1 \end{pmatrix}$$

Experiment 2 (5 dimensions):

$$K = 2,\ \omega = (0.5, 0.5),\ b = (-0.2, 0.5),\ \beta^\top = \begin{pmatrix} 1 & 1 & 2 & -1 & 0 & 3 \\ 1 & 2 & -3 & 0 & 1 & 0 \end{pmatrix}$$

Experiment 3 (10 dimensions):

$$K = 3,\ \omega = (0.3, 0.3, 0.4),\ b = (-0.2, 0, 0.5),\ \beta^\top = \begin{pmatrix} 1 & 2 & -1 & 0 & 3 & 4 & -1 & -3 & 0 & 2 \\ 2 & -3 & 0 & 1 & 0 & -1 & -4 & 3 & 2 & 0 \\ -1 & 1 & 3 & -1 & 0 & 0 & 2 & 0 & 1 & -2 \end{pmatrix}$$

An experiment consists in a data generation step, followed by the parameters estimation. The same link function is used for these two stages. For all three experiments we use both logit and probit links, which amounts to a total of six different computations. The covariates vector X follows a multivariate Gaussian distribution ($X \sim \mathcal{N}(0, I_d)$). Within each experiment, the sample size varies from 10^3 to 10^6 (intermediate sizes: 5×10^3, 10^4, 5.10^4, 10^5 and 5×10^5). Computations are always run on the same data both for GLSMM and MLM. The procedure to reproduce the results is available online at `https://www.imo.universite-paris-saclay.fr/~auder/reproduce/morpheus/index.html`.

Estimation errors.
Figures 1 and 2 show how the average absolute estimation error evolves as the sample size n increases, respectively for the logit and probit links. Since flexmix fails for $n = 1000$ while our algorithm doesn't output very accurate estimations, values of n below 1000 were not tested. The parameters are estimated by averaging the outputs of $N = 100$ replications for different values of n in each experiment. The starting point for the optimization stage are the estimated directions μ in the GLSMM case. The starting point is not documented for the MLM in flexmix package and users do not have the opportunity to choose it in the package. ([8]). Since the output groups appear in a random order, the Hungarian algorithm is run to fix the label switching issue. Its input is the distance matrix D computed from the reference parameters R: $D_{i,j} = \texttt{dist}(R_i, P_j)$ where R_i and P_j are respectively the i^{th} and j^{th} columns of the reference matrix and some parameters matrix in output.

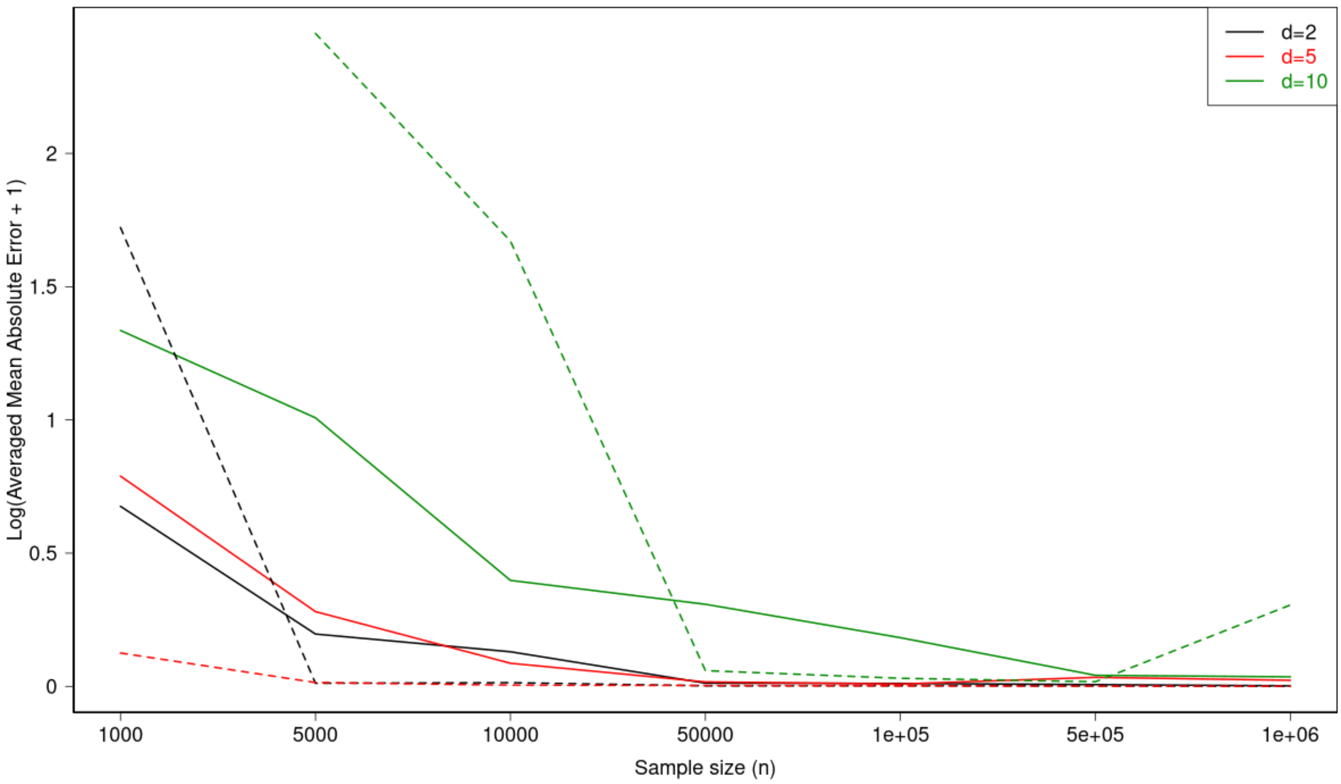

Fig. 1 *logit* link function. Average absolute estimation error evolution as n increases. Solid line represents GLSMM, while dotted line is MLM. In black $d = 2$, in red $d = 5$ and in green $d = 10$.

Logit link function:
The MLM algorithm (dotted lines) appears to converge faster than GLSMM, while also being less robust: MLM often fails for small values of n (below 10^4), and shows an unexpectedly high error at $n = 10^6$ for $d = 10$ (green dotted line). The curve corresponding to the third experiment ($d = 10$, in green) starts at $n = 5000$ for MLM (dotted line) because all run failed in this case.

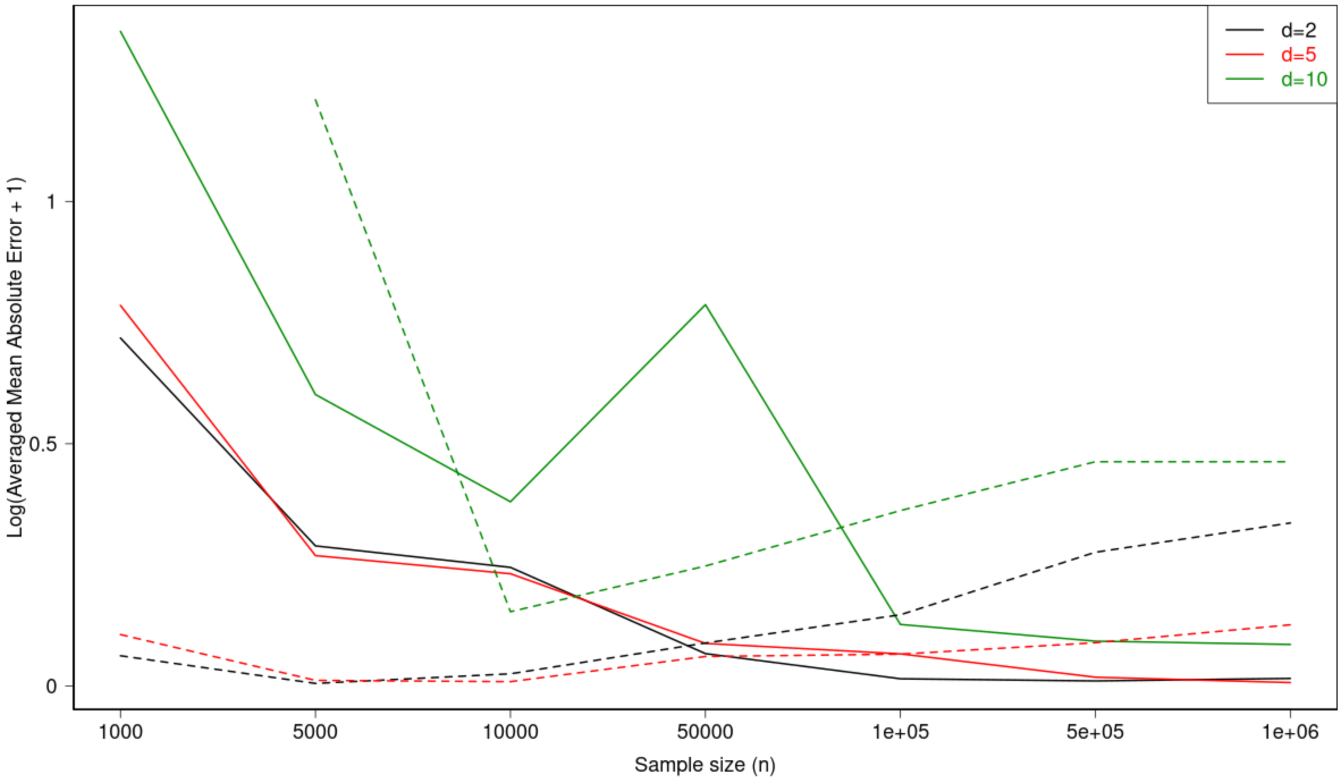

Fig. 2 *probit* link function. Average absolute estimation error evolution as n increases. Solid line represents GLSMM, while dotted line is MLM. In black $d = 2$, in red $d = 5$ and in green $d = 10$.

Probit link function:
From $n = 10^4$ the MLM errors tend to slightly increase with n, while the GLSMM curves globally decrease (except a peak at $n = 5.10^4$ when $d = 10$). This is in favor of our algorithm, although the performance when n is small is relatively poor.

Tables 1 and 2 contain a more detailed information: the Mean Absolute Error summed over all components of each parameter instead of a global average. That is to say, for ω (resp. b) the error showed is $\sum_{k=1}^{K} |\omega_k - \hat{\omega}_k|$ (resp. $\sum_{k=1}^{K} |b_k - \hat{b}_k|$). Concerning β, we show the error column per column: $\texttt{err}_k = \sum_{i=1}^{d} |\beta_{i,k} - \hat{\beta}_{i,k}|$ with $k \in [1, K]$. Note that the second line for the MLM algorithm ends with very high values (third experiment, $d = 10$): this is due to a high percentage of failed runs (about one third).

	Experiment 1 ($d = 2$)				Experiment 2 ($d = 5$)				Experiment 3 ($d = 10$)				
n	ω	β_1	β_2	b	ω	β_1	β_2	b	ω	β_1	β_2	β_3	b
10^3	1.8e-2	3.9e+0	5.4e+0	1.1e+0	4.7e-3	1.6e+1	1.4e+1	1.2e+0	1.3e-1	7.3e+1	5.2e+1	8.9e+1	3.7e+0
	1.1e-2	4.0e-1	4.5e+1	4.6e+0	1.4e-2	2.2e+0	9.1e-1	1.7e-1	-	-	-	-	-
5.10^3	4.5e-2	1.2e+0	1.0e+0	1.2e-1	3.8e-4	2.3e+0	2.5e+0	9.8e-1	4.9e-2	8.6e+0	8.7e+0	2.4e+1	1.2e+1
	4.0e-3	7.1e-2	6.0e-2	1.6e-3	6.0e-4	2.8e-1	8.5e-2	1.3e-2	1.3e-1	3.5e+2	5.0e+2	1.4e+0	1.0e+1
10^4	1.4e-2	6.3e-1	7.2e-1	1.4e-1	2.6e-4	6.5e-1	1.0e+0	2.1e-1	2.7e-2	3.7e+0	4.3e+0	8.1e+0	2.8e+0
	2.0e-3	2.6e-2	9.5e-2	2.2e-2	1.8e-3	5.3e-2	4.7e-2	6.4e-3	1.3e-1	3.0e+2	1.8e+0	4.5e-1	8.9e+0
5.10^4	1.1e-2	6.9e-2	2.2e-2	1.7e-2	3.4e-3	4.2e-1	1.1e-2	1.2e-2	1.9e-2	1.0e+1	7.3e+0	9.2e+0	5.6e-1
	2.3e-4	7.2e-3	1.2e-2	3.5e-3	3.7e-4	1.6e-2	2.9e-2	1.0e-2	1.3e-1	2.7e-1	1.8e-1	2.4e+0	1.3e-1
10^5	1.9e-3	1.1e-2	8.9e-2	1.5e-2	1.9e-2	3.1e-2	5.3e-2	1.8e-2	3.3e-2	3.0e+0	2.7e+0	5.6e+0	6.3e-1
	7.7e-4	7.0e-3	1.1e-2	5.3e-3	5.9e-5	2.3e-2	2.6e-2	2.1e-3	1.3e-1	1.6e-1	7.8e-1	9.9e-2	4.0e-2
5.10^5	4.3e-3	3.0e-2	2.0e-2	1.1e-2	5.0e-2	3.7e-1	2.6e-1	3.0e-2	9.9e-3	1.7e+0	4.6e-1	5.8e-1	1.0e-1
	1.2e-4	3.6e-3	6.7e-3	4.1e-3	2.3e-4	6.5e-3	3.7e-3	3.0e-3	1.3e-1	2.8e-2	1.7e-1	3.2e-2	2.4e-3
10^6	8.2e-5	1.5e-3	9.6e-3	6.6e-3	4.0e-2	2.8e-1	1.8e-1	5.7e-3	2.3e-2	1.1e+0	6.4e-2	9.7e-1	9.4e-2
	2.8e-5	4.9e-4	1.6e-3	1.0e-3	7.0e-5	2.5e-3	6.0e-3	1.9e-3	1.3e-1	2.5e+1	1.5e+0	3.0e-2	4.1e-1

Table 1 *logit* link function. Summed errors for GLSMM (top number) and MLM (bottom number), for increasing values of n. Parameters estimations are averaged over $N = 100$ runs.

	Experiment 1 ($d = 2$)				Experiment 2 ($d = 5$)				Experiment 3 ($d = 10$)				
n	ω	β_1	β_2	b	ω	β_1	β_2	b	ω	β_1	β_2	β_3	b
5.10^3	2.9e-2	5.5e+0	5.2e+0	9.1e-1	7.3e-3	1.6e+1	1.5e+1	9.4e-1	1.3e-1	6.6e+1	6.2e+1	8.2e+1	4.6e+0
	1.1e-2	2.1e-1	4.3e-1	5.8e-2	4.8e-3	1.1e+0	1.5e+0	1.6e-1	-	-	-	-	-
10^4	3.0e-2	2.0e+0	1.4e+0	3.1e-1	5.4e-3	3.3e+0	2.7e+0	6.5e-1	5.2e-2	1.7e+0	1.6e+0	8.8e+0	6.2e+0
	3.3e-3	1.8e-2	2.8e-2	6.2e-3	2.0e-3	8.3e-2	1.2e-1	2.8e-2	1.4e-1	7.6e+1	1.1e+2	1.5e+0	2.1e+0
10^5	1.1e-2	1.1e+0	1.7e+0	2.5e-1	2.4e-2	2.0e+0	2.3e+0	6.9e-1	2.2e-2	3.1e+0	3.5e+0	8.4e+0	2.6e+0
	1.2e-3	8.2e-2	1.5e-1	3.5e-2	4.5e-4	1.5e-1	8.8e-2	6.4e-3	1.4e-1	5.9e+0	6.2e+0	2.5e-1	1.2e-1
5.10^5	6.7e-4	1.2e-1	5.5e-1	7.9e-2	2.1e-2	1.0e+0	1.1e+0	9.6e-2	1.3e-2	2.6e+1	2.5e+1	3.4e+1	2.2e+0
	1.7e-3	2.9e-1	5.8e-1	1.2e-1	2.7e-4	7.5e-1	7.2e-1	8.3e-2	1.4e-1	7.9e+0	7.7e+0	4.9e+0	3.4e-1
10^6	6.3e-3	1.3e-1	2.7e-2	6.1e-3	2.9e-2	7.1e-1	3.4e-1	1.7e-1	1.9e-2	1.8e+0	1.3e+0	3.6e+0	5.4e-1
	1.2e-3	5.0e-1	9.9e-1	2.0e-1	5.4e-5	8.1e-1	7.6e-1	9.3e-2	1.3e-1	1.3e+1	1.2e+1	7.8e+0	5.4e-1
5.10^5	3.7e-3	3.6e-2	6.8e-2	7.9e-3	3.1e-2	1.4e-1	2.2e-1	7.1e-3	1.2e-2	1.5e+0	4.9e-1	4.1e+0	2.5e-1
	2.4e-4	1.0e+0	2.0e+0	4.1e-1	5.4e-5	1.1e+0	1.1e+0	1.3e-1	1.3e-1	1.7e+1	1.6e+1	1.1e+1	7.4e-1
10^6	4.9e-3	4.7e-2	1.1e-1	1.4e-2	2.0e-2	4.7e-2	3.6e-2	6.8e-3	1.0e-2	2.6e+0	1.9e+0	1.9e+0	1.5e-1
	1.2e-4	1.3e+0	2.5e+0	5.1e-1	1.0e-5	1.6e+0	1.5e+0	1.8e-1	1.3e-1	1.7e+1	1.6e+1	1.1e+1	7.4e-1

Table 2 *probit* link function. Summed errors for GLSMM (top number) and MLM (bottom number), for increasing values of n. Parameters estimations are averaged over $N = 100$ runs.

Computational time.

The running time of LSMM (GLSMM without the generalized least squares step) does not depend directly on n, since after computing the empirical moments it operates on matrices of size at most $O(d^3 \times K)$. So, we observed excellent running times in this case, outperforming MLM by far especially for $n = 10^6$, as seen on Table 3. The times were averaged over 100 runs using random true parameters, with the logit link function. Only one core is available for each run (they are executed in parallel). The top numbers correspond to the GLSMM algorithm, and the numbers below correspond to the flexmix package (MLM).

	$n = 5.10^3$	$n = 10^4$	$n = 5.10^4$	$n = 10^5$	$n = 5.10^5$	$n = 10^6$
$d = 2$	0.83 4.1	0.82 5.7	0.76 21	0.93 35	1.2 146	1.4 315
$d = 5$	1.4 3.4	1.4 6.3	1.6 26	1.6 64	2.5 279	3.4 537
$d = 10$	5.6 4.3	6.2 7.1	7.4 36	7.5 88	12 446	15 882

Table 3 Average running time over 100 runs *with a fixed W matrix* on one core, for the logit link, for different values of n and d. Top numbers: GLSMM; bottom numbers: MLM.

However, the accuracy of the initial method in high dimension was clearly bad, so the final step computing and using the empirical matrix W is required. This final step turns out to be quite costly, so the timings in this case get closer, with still an advantage to GLSMM as seen on Table 4. This time we used four cores per run to benefit from a parallel loop in the computation of W, in addition to the parallelization of the runs. As we can see this also helps the flexmix package.

	$n = 5.10^3$	$n = 10^4$	$n = 5.10^4$	$n = 10^5$	$n = 5.10^5$	$n = 10^6$
$d = 2$	14.5 1.2	16.2 2.3	15.2 10.6	13.3 16.1	11.7 64.8	9.8 (*) 116.6
$d = 5$	7.1 1.5	9.9 2.6	11.3 12.1	13.0 23.5	19.2 112.0	24.5 181.3
$d = 10$	12.7 1.7	26.3 3.2	38.3 15.7	52.4 30.2	132.1 151.9	220.3 298.3

Table 4 Average running time over 100 runs *with optimisation of W* on four cores, for the logit link, for different values of n and d. Top numbers: GLSMM; bottom numbers: MLM.

(*) The computational time decreases with higher n in this case for GLSMM. This could be explained by a quicker convergence when n increases, doing more than compensating for the slightly longer computation time for the moments and the matrix W. When d is larger, the optimisation step is no longer the only factor.

We can imagine to benefit from the initial speed of the algorithm by estimating W on a reduced dataset, before the parameters are estimated from the full dataset starting from a custom $W_{\texttt{init}}$. Overall, the results are quite promising, especially if we find a way to optimize the computation of W.

4.3 **Random initialization**

Instead of choosing μ as a starting point for the optimization stage, we could just sample a few points at random in the unit sphere S^{d-1} and keep the initial point leading to the smallest value of $Q_n(\theta)$. So we compare here these two settings, choosing arbitrarily three random points on the sphere. This version of GLSMM is written GLSMM3 in this section. Table 5 summarizes the results, showing in each cell the total error for ω, b or a column of the matrix β. First two lines correspond to the logit link, respectively for GLSMM and GLSMM3, whereas last two lines correspond to the probit link function.

Experimental setup:

- $n = 10^5$ sample points,
- parameters according to the experiments 1, 2 and 3 respectively ($d = 2$, then 5 and 10).

	Experiment 1				Experiment 2				Experiment 3				
link	ω	β_1	β_2	b	ω	β_1	β_2	b	ω	β_1	β_2	β_3	b
logit	6.2e-3	2.8e-2	8.5e-3	5.7e-3	1.2e-2	4.6e-2	1.0e-1	4.2e-2	1.6e-2	7.3e-1	1.6e+0	2.7e+0	2.8e-1
	2.6e-2	5.7e-2	5.4e-2	3.9e-3	1.7e-3	7.4e-2	1.4e-1	4.1e-2	1.2e-2	5.8e-1	2.0e+0	1.5e+0	2.3e-1
probit	5.6e-3	5.8e-2	1.1e-2	6.5e-3	1.5e-2	1.3e-1	4.4e-1	1.7e-1	1.1e-2	1.2e+0	5.6e-1	3.9e+0	3.6e-1
	2.6e-2	2.2e-1	1.8e-1	3.1e-2	6.0e-4	1.2e-1	6.1e-1	6.8e-2	8.7e-3	4.6e-1	3.2e-1	4.0e+0	3.8e-1

Table 5 Summed errors for GLSMM (top numbers) and GLSMM3 (bottom numbers), for $n = 10^5$ and both logit and probit link functions. Parameters estimations are averaged over $N = 100$ runs.

The random starts version of GLSMM is slightly more accurate, although it is not easily spotted on the tables. The total sums of errors for the logit link are respectively 5.5 (GLSMM) and 4.7 (GLSMM3), and respectively 6.9 and 6.4 for the probit link function. However, using three random starting points is costly, because of the current computational burden to obtain the weights matrix W. Moreover, using only one random starting point (GSLMM1) leads to a total sum of errors quite higher than GSLMM: 9.4 for the logit link, and 9.1 for the probit link. As a conclusion, using μ as a starting point for the optimization step is a good compromise.

5 Real data

We chose the classic Diamonds dataset to experiment the algorithms on real data. It is available for example on Kaggle: `https://www.kaggle.com/shivam2503/diamonds`. There are about 50k individual data, each line containing characteristics of a diamond. After a preliminary step of variables selections by trials and errors, we keep the next three as numerical covariables:

- The carat weight of the diamond.
- The total depth percentage, which is calculated from diamond's dimensions.
- The width of top of diamond relative to widest point ("table").

We build a binary output from the "cut" covariable, which indicates the quality of the diamond. Values "Fair", "Good" and "Very Good" correspond to 0, while values "Premium" and "Ideal" correspond to 1. The goal is thus to predict the diamond's quality based on a few numerical variables.

Dataset summary:

- 53940 lines,

- 3 numerical covariables considered Gaussian,
- one binary output: 18598 '0' and 35342 '1'.

Parameters are estimated using both GSLMM and MLM algorithms, with different values of K ($K = 2, 3$), and different link functions (logit and probit). Data are split into a training set and a testing set, in respective proportions 70% and 30%. The classification results on the testing set and running times are given in Table 6. The flexmix package gives a slightly better accuracy, at the cost of a quite larger running time. The results suggest that this dataset does not contain more than two groups.

	$K = 2$		$K = 3$	
link	GSLMM	MLM	GSLMM	MLM
logit	0.75 (4s)	0.78 (9s)	0.73 (3s)	0.78 (21s)
probit	0.75 (4s)	0.78 (9s)	0.74 (4s)	0.78 (35s)

Table 6 Classification results on Diamonds dataset, for $K = 2, 3$ and both link functions, using GSLMM and MLM. Computation times are in parenthesis.

6 Some other identifiability results

In this section, we provide several further identifiability results for various population mixture models of binary regressions.

Identifiability of a model is a first step to obtain theoretical guarantees for practical estimation procedures. Our identifiability results open the way to build estimators for which theoretical guarantees could be obtained. In particular, for parametric maximum likelihood estimators in mixture models for which algorithms already exist, consistency is a consequence of our identifiability theorems by applying the usual theory.

6.1 Continuous covariates

We first consider the setting of the previous sections where

$$E\left(Y|X\right) = \sum_{k=1}^{K} \omega_k g\left(\langle \beta_k, X\rangle + b_k\right).$$

We assume that for all k, $\omega_k \geq 0$, that $\sum_{k=1}^{K} \omega_k = 1$, and that g takes value in $(0, 1)$. We show below that the directions of the regression vectors may be recovered as soon as they are distinct, even if the link function is unknown.
Denote $P_{g,\omega,\beta,b}$ the probability distribution of (X, Y), with $\omega = (\omega_1, \cdots, \omega_K)$, $\beta = [\beta_1|, \cdots, |\beta_K] \in \mathbb{R}^{d\times K}$, and $b = (b_1, \cdots, b_K) \in \mathbb{R}^K$. When g is unknown, obviously it is needed to fix origin and scale; we choose to fix $g(0)$ and $g(1)$ (without loss of generality). Denote $\mu_k = \beta_k/\|\beta_k\|$ and $\lambda_k = \|\beta_k\|$, $k = 1, \ldots, K$, so that $\beta_k = \lambda_k\mu_k$.
We introduce the assumptions:

- (S1) The support of the law of X is $\mathbb{R}^d$.
- (S2) For all $j \neq k$, $\mu_j \neq \mu_k$ and $\mu_j \neq -\mu_k$.
- (S3) The function $g : \mathbb{R} \to]0, 1[$ is increasing, has limit 0 in $-\infty$, limit 1 in $+\infty$, and it is continuously derivable with derivative having limit 0 in $-\infty$ and in $+\infty$. Also, $g(0) < g(1)$ are fixed.

Remark: There is no assumption on K with respect to d.

Theorem 4 *Under assumptions (S1), (S2) and (S3), knowledge of $P_{g,\omega,\beta,b}$ allows to recover K and $\mu_1, \ldots, \mu_K$.*

The proof of this theorem is given in Section 7.3.
Under the more stringent assumption that the regression vectors are linearly independent, it is possible to recover all parameters and the link function.

– (S2bis) The vectors $\mu_1, \ldots, \mu_K$ are linearly independent.

Remark: (H2bis) implies that $K \leq d$.

Theorem 5 *Under assumptions (S1), (S2bis) and (S3), the mixture model is identifiable: the knowledge of $P_{g,\omega,\beta,b}$ allows to recover K, g, ω, β and b.*

The proof of this theorem is given in Section 7.4.

6.2 Continuous and categorical covariates

We now consider the situation where part of the covariates are categorical and are coded with m binary variables in a vector Z taking values in $\{0,1\}^m$. Let $\mathcal{Z} \subset \{0,1\}^m$ be the set of possible values of Z (depending on the possible combinations of covariates). We still denote $X \in \mathbb{R}^d$ the continuous covariates. Now

$$E(Y|X,Z) = \sum_{k=1}^{K} \omega_k g(\langle \beta_k, X\rangle + \langle \gamma_k, Z\rangle + b_k),$$

and we denote $P_{g,\omega,\beta,\gamma,b}$ the probability distribution of (X,Y), with $\omega = (\omega_1, \cdots, \omega_K)$, $\beta = [\beta_1|, \cdots, |\beta_K] \in \mathbb{R}^{d\times K}$, $\gamma = [\gamma_1|, \cdots, |\gamma_K] \in \mathbb{R}^{m\times K}$, and $b = (b_1, \cdots, b_K) \in \mathbb{R}^K$. We assume that there exist $z_1, \ldots, z_{m+1}$ in $\mathcal{Z}$ such that

– (S4) The matrix $\begin{pmatrix} 1 & z_1^\top \\ 1 & z_2^\top \\ \vdots & \vdots \\ 1 & z_{m+1}^\top \end{pmatrix}$ is full rank.

Remark: (S4) implies that $\mathcal{Z}$ contains at least m linearly independent vectors.

The continuous covariates allow to identify g.

Theorem 6 *Under assumptions (S1), (S2bis), (S3) and (S4), the model is identifiable: the knowledge of $P_{g,\omega,\beta,\gamma,b}$ allows to recover K, g, ω, β, γ et b.*

The proof of this theorem is given in Section 7.5.

6.3 Longitudinal observations

We now consider the situation where for each individual Y, conditional to the membership of a population, we have q independent experiments with several covariates $X_1, \ldots, X_q$ and $Z_1, \ldots, Z_q$. Thus the random variable Y has dimension m, and

$$E(Y|X,Z) = \sum_{k=1}^{K} \omega_k \left(g(\langle \beta_k, X_j\rangle + \langle \gamma_k, Z_j\rangle + b_k)\right)_{1\leq j\leq q}.$$

As soon as the number of experiments is at least 3, we do not need the linear independence of the regression vectors to get identifiability.

Theorem 7 *Assume that $q \geq 3$. If (S1), (S2), (S3) and (S4) hold, then the model is identifiable: the knowledge of $P_{g,\omega,\beta,\gamma,b}$ allows to recover K, g, ω, β, γ and b.*

The proof of this theorem is given in Section 7.6.

7 Proofs

7.1 Proof of Theorem 1 and Theorem 2

Let, for $k = 1, \dots, K$, $\lambda_k = \|\beta_k\|$ and $\mu_k = \beta_k / \|\beta_k\|$. Using Stein's identity, Sedghi et al. ([22]) prove the following lemma:

Lemma 1 ([22]) *Under (H3) the moments can be rewritten:*

$$M_1(\theta) = \sum_{k=1}^{K} \omega_k \lambda_k E[g'\left(\lambda_k \langle X, \mu_k \rangle + b_k\right)]\ \mu_k,$$
$$M_2(\theta) = \sum_{k=1}^{K} \omega_k \lambda_k^2 E[g''\left(\lambda_k \langle X, \mu_k \rangle + b_k\right)]\ \mu_k \otimes \mu_k,$$
$$M_3(\theta) = \sum_{k=1}^{K} \omega_k \lambda_k^3 E[g^{(3)}\left(\lambda_k \langle X, \mu_k \rangle + b_k\right)]\ \mu_k \otimes \mu_k \otimes \mu_k.$$

Under (H1), we see by Lemma 1 that K is the rank of $M_2(\theta)$.
It is proved in [22] that we can recover the μ_k's up to sign from the knowledge of $M_2(\theta)$ and $M_3(\theta)$, but since $M_1(\theta)$ is a linear combination of the μ_k's with positive coefficients, under (H1) the knowledge of $M_1(\theta)$ allows to recover the signs. It is then seen that using $M_1(\theta)$, $M_2(\theta)$ and $M_3(\theta)$, one may recover the 3-uples

$$\left(\omega_k E[g'\left(\langle \beta_k, X \rangle + b_k\right)]\lambda_k ; \omega_k E[g''\left(\langle \beta_k, X \rangle + b_k\right)]\lambda_k^2 ; \omega_k E[g^{(3)}\left(\langle \beta_k, X \rangle + b_k\right)]\lambda_k^3\right),$$

$k = 1, \dots, K$. Thus, one gets identifiability as soon as the function from $]0, +\infty[\times\mathbb{R}\times]0, +\infty[$ to its image that associates (ω, b, λ) to

$$\left(\omega\lambda \int g'(\lambda z + b) e^{-z^2/2} dz ; \omega\lambda^2 \int g''(\lambda z + b) e^{-z^2/2} dz ; \omega\lambda^3 \int g^{(3)}(\lambda z + b) e^{-z^2/2} dz\right)$$

is one-to-one. Using integration by parts this is equivalent to the fact that the function from $]0, +\infty[\times\mathbb{R}\times]0, +\infty[$ to its image that associates (ω, b, λ) to

$$\lambda\left(\omega \int g'(\lambda z + b) e^{-z^2/2} dz ; \omega \int z g'(\lambda z + b) e^{-z^2/2} dz ; \omega \int z^2 g'(\lambda z + b) e^{-z^2/2} dz\right)$$

is one-to-one. This is again equivalent to the fact that the function from $\mathbb{R}\times]0, +\infty[$ to its image that associates (b, λ) to

$$\left(\frac{\int z g'(\lambda z + b) e^{-z^2/2} dz}{\int g'(\lambda z + b) e^{-z^2/2} dz} ; \frac{\int z^2 g'(\lambda z + b) e^{-z^2/2} dz}{\int g'(\lambda z + b) e^{-z^2/2} dz}\right)$$

is one-to-one. For any $(b, \lambda) \in \mathbb{R}\times]0, +\infty[$, define

$$dQ_{(b,\lambda)}(z) = \frac{g'(\lambda z + b) e^{-z^2/2}}{\int g'(\lambda z + b) e^{-z^2/2} dz} dz. \tag{4}$$

Then it is equivalent to prove that the knowledge of

$$\left(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\right) := \left(\int z dQ_{(b,\lambda)}(z); \int z^2 dQ_{(b,\lambda)}(z)\right) \tag{5}$$

implies the knowledge of (b, λ).

7.1.1 End of the proof of Theorem 1

When the link function g is probit, then $g'(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. Replacing in equation (4), we have

$$dQ_{(b,\lambda)}(z) = \frac{e^{-\frac{1}{2}\left((\lambda z+b)^2+z^2\right)}}{\int e^{-\frac{1}{2}((\lambda z+b)^2+z^2)}dz}dz,$$

which after some computations leads to

$$Q_{(b,\lambda)} = \mathcal{N}\left(-\frac{\lambda b}{\lambda^2+1}; \frac{1}{\lambda^2+1}\right).$$

Its first two moments are then given by

$$(\alpha_1, \alpha_2) := \left(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\right) = \left(-\frac{\lambda b}{\lambda^2+1}; \frac{\lambda^2 b^2 + \lambda^2 + 1}{(\lambda^2+1)^2}\right).$$

We can then recover b and λ by

$$b = -\alpha_1 \frac{(\lambda^2+1)}{\lambda}$$

and

$$\lambda = \sqrt{(\alpha_2 - \alpha_1^2)^{-1} - 1}.$$

7.1.2 End of the proof of Theorem 2

It remains to prove that for some $B >$ and $L > 0$, if $(H2)$ holds, then, the function that associates $(b, \lambda) \in]-B, B[\times]0, L[$ to $\left(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\right)$ is one-to-one on its image.
Using (H3) and integration by parts, we get that for (b, λ) in a neighborhood of $(0, 0)$

$$\left(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\right) = \left(\frac{\lambda \int g''(\lambda z+b)e^{-z^2/2}dz}{\int g'(\lambda z+b)e^{-z^2/2}dz}; 1 + \frac{\lambda^2 \int g^{(3)}(\lambda z+b)e^{-z^2/2}dz}{\int g'(\lambda z+b)e^{-z^2/2}dz}\right). \tag{6}$$

From (H2), one gets that

– (P1) The function g' is positive and satisfies $g'(x) = g'(-x)$ for all $x \in \mathbb{R}$,
– (P2) The function g'' satisfies $g''(x) = -g''(-x)$ for all $x \in \mathbb{R}$ and $g''(x) < 0$ for $x > 0$,
– (P3) $g'(0) > 0$,
– (P4) $g''(0) = g^{(4)}(0) = 0$.

Let us define the functions K_s, $s = 1, 2, 3$ such that

$$\begin{aligned} K_s : \mathbb{R}\times]0, +\infty[&\to \mathbb{R} \\ (b, \lambda) &\mapsto K_s(b, \lambda) = \int g^{(s)}(\lambda z+b)e^{-z^2/2}dz \end{aligned}$$

Using (H3), the functions K_s, $s = 1, 2, 3$, are differentiable in a neighborhood of $(0, 0)$ and Taylor expansion writes:

$$K_s(b, \lambda) = K_s(0, 0) + \langle \nabla K_s(0, 0), (b, \lambda)\rangle + o(\lambda^2 + b^2). \tag{7}$$

Now

$$\frac{\partial K_s}{\partial \lambda}(0, 0) = \int z g^{(s+1)}(0)e^{-z^2/2}dz \tag{8}$$

and

$$\frac{\partial K_s}{\partial b}(0, 0) = \int g^{(s+1)}(0)e^{-z^2/2}dz \tag{9}$$

so that

$$K_s(b,\lambda) = g^{(s)}(0)\int e^{-z^2/2}dz + g^{(s+1)}(0)\int (\lambda z + b)e^{-z^2/2}dz + o(\lambda^2 + b^2). \tag{10}$$

Using $(P4)$ and (10), we have

$$\int g'(\lambda z + b)e^{-z^2/2}dz = \sqrt{2\pi}g'(0) + o(\lambda^2 + b^2), \tag{11}$$

$$\int g''(\lambda z + b)e^{-z^2/2}dz = \sqrt{2\pi}g^{(3)}(0)b + o(\lambda^2 + b^2), \tag{12}$$

and

$$\int g^{(3)}(\lambda z + b)e^{-z^2/2}dz = \sqrt{2\pi}g^{(3)}(0) + o(\lambda^2 + b^2). \tag{13}$$

Therefore, replacing (11) to (13) in (6), we get

$$E_{(b,\lambda)}(Z) = \frac{g^{(3)}(0)}{g'(0)}\lambda b + o(\lambda^2 + b^2)$$

and

$$E_{(b,\lambda)}(Z^2) = 1 + \frac{g^{(3)}(0)}{g'(0)}\lambda^2 + o(\lambda^2 + b^2),$$

which easily leads to

$$\lambda^2 = \frac{g'(0)}{g^{(3)}(0)}\left(E_{(b,\lambda)}(Z^2) - 1\right) + o(|E_{(b,\lambda)}(Z^2) - 1| + |E_{(b,\lambda)}(Z)|)$$

and

$$\lambda b = \frac{g'(0)}{g^{(3)}(0)}E_{(b,\lambda)}(Z) + o(|E_{(b,\lambda)}(Z^2) - 1| + |E_{(b,\lambda)}(Z)|).$$

This proves that the function $(b,\lambda) \mapsto (E_{(b,\lambda)}(Z), E_{(b,\lambda)}(Z^2))$ is invertible in a neighborhood of $(0,0)$.

7.2 Proof of Theorem 3

By the law of large numbers, $Q_n^W(\theta)$ converges to

$$Q^W(\theta) := \tilde{M}^\top(\theta)W\tilde{M}(\theta).$$

Define

$$S_n = \sup_{\theta\in\Theta}\left|Q_n^W(\theta) - Q^W(\theta)\right|.$$

Since $Q^W(\theta)$ has θ^* as unique minimum (up to label switching), to prove the consistency of $\widehat{\theta}_n$, it is enough to prove that S_n converges to 0 in probability, see Theorem 5.7 in [23]. Let λ_{max} be the largest eigenvalue of W and $F = (d + d^2 + d^3)^2$. We easily get

$$\begin{aligned}
S_n \leq{}& \lambda_{max}F\sum_{j\in[d]}\left(\left|\hat{M}_1[j] - M_1(\theta^\star)[j]\right|\right)\left(\left|\hat{M}_1[j]\right| + \left|M_1(\theta^*)[j]\right| + 2\sup_{\theta\in\Theta}\left|M_1(\theta)[j]\right|\right)\\
&+ \lambda_{max}F\sum_{j,k\in[d]}\left(\left|\hat{M}_2[j,k] - M_2(\theta^\star)[j,k]\right|\right)\left(\left|\hat{M}_2[j,k]\right| + \left|M_2(\theta^*)[j,k]\right| + 2\sup_{\theta\in\Theta}\left|M_2(\theta)[j,k]\right|\right)\\
&+ \lambda_{max}F\sum_{j,k,l\in[d]}\left(\left|\hat{M}_3[j,k,l] - M_3(\theta^\star)[j,k,l]\right|\right)\left(\left|\hat{M}_3[j,k,l]\right| + \left|M_3(\theta^*)[j,k,l]\right| + 2\sup_{\theta\in\Theta}\left|M_3(\theta)[j,k,l]\right|\right),
\end{aligned}$$

and since the functions $\theta \mapsto M_r(\theta),\ r=1,2,3$ are continuous and Θ is compact, then there exist $c_1,\ c_2$ and c_3 such that

$$\begin{aligned}S_n \le \lambda_{max} D \sum_{j\in[d]} \left(c_1 + \left|\hat{M}_1[j]\right|\right)\left(\left|\hat{M}_1[j] - M_1(\theta^\star)[j]\right|\right)\\ + \lambda_{max} D \sum_{j,k\in[d]} \left(c_2 + \left|\hat{M}_2[j,k]\right|\right)\left(\left|\hat{M}_2[j,k] - M_2(\theta^\star)[j,k]\right|\right)\\ + \lambda_{max} D \sum_{j,k,l\in[d]} \left(c_3 + \left|\hat{M}_3[j,k,l]\right|\right)\left(\left|\hat{M}_3[j,k,l] - M_3(\theta^\star)[j,k,l]\right|\right)\end{aligned}$$

which converges to 0 by the law of large numbers, which ends the proof of the consistency of $\widehat{\theta}_n$.

Let us define Z_n as $Z_n(\theta) = \nabla_\theta Q_n^W(\theta)$. The $r-th$ coordinate of $Z_n(\theta)$ can be obtained by

$$\frac{\partial Q_n^W(\theta)}{\partial \theta_r} = 2\left(\frac{\partial \tilde{M}(\theta)^\top}{\partial \theta_r}\right) W \left(\frac{1}{n}\sum_{i=1}^n \tilde{M}_i(\theta)\right).$$

Using Taylor expansion, we get

$$Z_n(\hat{\theta}_n) = Z_n(\theta^\star) + \int_0^1 D_1 Z_n\left[\theta^\star + t(\hat{\theta}_n - \theta^\star)\right]\left(\hat{\theta}_n - \theta^\star\right) dt \tag{14}$$

where $D_1 Z_n$ is the first derivative matrix of Z_n. Since $Z_n(\hat{\theta}_n) = 0$, we have

$$-\sqrt{n} Z_n(\theta^\star) = \left[\int_0^1 D_1 Z_n\left[\theta^\star + t(\hat{\theta}_n - \theta^\star)\right] dt\ \right] \sqrt{n}\left(\hat{\theta}_n - \theta^\star\right) \tag{15}$$

Applying the central limit theorem and the delta method we get that $\sqrt{n} Z_n(\theta^\star)$ is asymptotically Gaussian.

The $(r_1, r_2) - th$ coordinate of $D_1 Z_n(\theta) = \nabla_\theta^2 Q_n(\theta)$ are given by

$$\frac{\partial^2 Q_n^W(\theta)}{\partial \theta_{r_1}\partial \theta_{r_2}} = 2\left(\frac{\partial^2 \tilde{M}(\theta)^\top}{\partial \theta_{r_1}\partial \theta_{r_2}}\right) W \left(\frac{1}{n}\sum_{i=1}^n \tilde{M}_i(\theta)\right) + V_{r_1 r_2}(\theta)$$

with

$$V_{r_1 r_2}(\theta) = 2\left(\frac{\partial \tilde{M}(\theta)^\top}{\partial \theta_{r_1}}\right) W \left(\frac{\partial \tilde{M}(\theta)}{\partial \theta_{r_2}}\right).$$

It is not difficult to prove that $\int_0^1 D_1 Z_n\left[\theta^\star + t(\hat{\theta}_n - \theta^\star)\right] dt$ converges in probability to $V(\theta^\star)$ so that the proof is completed by showing that the matrix $V = V(\theta^\star)$ is invertible.

V is a $q \times q$ matrix with $q = K(2+d) - 1$. Let $U \in \mathbb{R}^q$. We shall denote the coordinates of U according to the parameters. Since W is a symmetric and positive definite square matrix , using the form of V we get that $U^\top V U = 0$ if and only if:

$$U^\top D M_1(\theta^\star)[j] = 0,\ j = 1, \ldots, q, \tag{16}$$

and

$$U^\top D M_2(\theta^\star)[j,l] = 0,\ j,l = 1, \ldots, q, \tag{17}$$

and

$$U^\top D M_3(\theta^\star)[j,l,m] = 0,\ j,l,m = 1, \ldots, q. \tag{18}$$

Here, DM_i is the gradient vector of the involved coordinate of M_i. Denote $U(\beta_k)$ the d-dimensional vector involving the coordinates of U according to parameter β_k. Denote $\overline{0}$ the d-dimensional zero vector,

$\overline{0} \otimes \overline{0} = \overline{0}^{\otimes 2}$ the $d \times d$-dimensional zero matrix and $\overline{0} \otimes \overline{0} \otimes \overline{0} = \overline{0}^{\otimes 3}$ the $d \times d \times d$-dimensional zero third order tensor. Then, the equation (16) can be rewritten as:

$$U^\top DM_1(\theta^\star)[j] = \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_1(\theta^\star)[j]}{\partial \omega_k} + \sum_{k=1}^{K} U(b_k)\frac{\partial M_1(\theta^\star)[j]}{\partial b_k} + \sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_1(\theta^\star)[j]}{\partial \beta_{mk}}, \tag{19}$$

the equation (17) can be rewritten as

$$U^\top DM_2(\theta^\star)[j,l] = \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_2(\theta^\star)[j,l]}{\partial \omega_k} + \sum_{k=1}^{K} U(b_k)\frac{\partial M_2(\theta^\star)[j,l]}{\partial b_k} + \sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_2(\theta^\star)[j,l]}{\partial \beta_{mk}}, \tag{20}$$

and the equation (18) can be rewritten as

$$U^\top DM_3(\theta^\star)[j,l,m] = \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_3(\theta^\star)[j,l,m]}{\partial \omega_k} + \sum_{k=1}^{K} U(b_k)\frac{\partial M_3(\theta^\star)[j,l,m]}{\partial b_k} + \sum_{k=1}^{K}\sum_{s=1}^{d} U(\beta_{sk})\frac{\partial M_3(\theta^\star)[j,l,m]}{\partial \beta_{sk}}. \tag{21}$$

Using the fact that $\sum_{k=1}^{d} \omega_k = 1$, the first terms of the equations (19) to (21) are rewritten as:

$$\begin{aligned}\sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_1(\theta^\star)[j]}{\partial \omega_k} &= \sum_{k=1}^{K-1} U(\omega_k)\Big\{E\left[g'\left(\langle X, \beta_k^\star\rangle + b_k^\star\right)\right].\beta_k^\star(j) \\ &\quad - E\left[g'\left(\langle X, \beta_K^\star\rangle + b_K^\star\right)\right].\beta_K^\star(j)\Big\},\end{aligned} \tag{22}$$

$$\begin{aligned}\sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_2(\theta^\star)[j,l]}{\partial \omega_k} &= \sum_{k=1}^{K-1} U(\omega_k)\Big\{E\left[g''\left(\langle X, \beta_k^\star\rangle + b_k^\star\right)\right].\beta_k^\star(j)\beta_k(l) \\ &\quad - E\left[g''\left(\langle X, \beta_K^\star\rangle + b_K^\star\right)\right].\beta_K^\star(j)\beta_K^\star(l)\Big\}\end{aligned} \tag{23}$$

and

$$\begin{aligned}\sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_3(\theta^\star)[j,l,m]}{\partial \omega_k} &= \sum_{k=1}^{K-1} U(\omega_k)\Big\{E\left[g^{(3)}\left(\langle X, \beta_k^\star\rangle + b_k^\star\right)\right].\beta_k^\star(j)\beta_k^\star(l)\beta_k^\star(m) \\ &\quad - E\left[g^{(3)}\left(\langle X, \beta_K^\star\rangle + b_K^\star\right)\right].\beta_K^\star(j)\beta_K^\star(l)\beta_K^\star(m)\Big\}\end{aligned} \tag{24}$$

respectively. Likewise the seconds terms of equations (19) to (21) are rewritten as:

$$\sum_{k=1}^{K} U(b_k)\frac{\partial M_1(\theta^\star)[j]}{\partial b_k} = \sum_{k=1}^{K} \omega_k^\star U(b_k) E\left[g''\left(\langle X, \beta_k^\star\rangle + b_k^\star\right)\right].\beta_k^\star(j), \tag{25}$$

$$\sum_{k=1}^{K} U(b_k)\frac{\partial M_2(\theta^\star)[j,l]}{\partial b_k} = \sum_{k=1}^{K} \omega_k^\star U(b_k) E\left[g^{(3)}\left(\langle X, \beta_k^\star\rangle + b_k^\star\right)\right].\beta_k^\star(j)\beta_k^\star(l) \tag{26}$$

and

$$\sum_{k=1}^{K} U(b_k)\frac{\partial M_3(\theta^\star)[j,l,m]}{\partial b_k} = \sum_{k=1}^{K} \omega_k^\star U(b_k) E\left[g^{(4)}\left(\langle X, \beta_k^\star\rangle + b_k^\star\right)\right].\beta_k^\star(j)\beta_k^\star(l)\beta_k^\star(m) \tag{27}$$

respectively. Derivating with respect to the β_k's coordinates and using Stein's identity, the last terms of equations (19) to (21) are rewritten as:

$$\sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_1(\theta^\star)[j]}{\partial \beta_{mk}} = \sum_{k=1}^{K}\omega_k^\star E\left[g^{(3)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\right]\langle \beta_k^\star, U(b_k)\rangle \beta_k^\star(j) + \sum_{k=1}^{K}\omega_k^\star E\left[g'\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\right]U(\beta_k(j)), \tag{28}$$

$$\sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_2(\theta^\star)[j,l]}{\partial \beta_{mk}} = \sum_{k=1}^{K}\omega_k^\star E\left[g^{(4)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\right]\langle \beta_k^\star, U(b_k)\rangle \beta_k^\star(j)\beta_k(l) + \sum_{k=1}^{K}\omega_k^\star E\left[g''\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\right]\Big\{\beta_k^\star(j)U(\beta_k(l)) + \beta_k^\star(l)U(\beta_k(j))\Big\}, \tag{29}$$

and

$$\sum_{k=1}^{K}\sum_{s=1}^{d} U(\beta_{sk})\frac{\partial M_3(\theta^\star)[j,l,m]}{\partial \beta_{sk}} = \sum_{k=1}^{K}\omega_k^\star E\left[g^{(5)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\right]\langle \beta_k^\star, U(b_k)\rangle \beta_k^\star(j)\beta_k(l)\beta_k^\star(s) + \sum_{k=1}^{K}\omega_k^\star E\left[g^{(3)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\right]\Big\{\beta_k^\star(j)\beta_k^\star(l)U(\beta_k^\star(s)) + \beta_k^\star(j)U(\beta_k(l))\beta_k^\star(s) + U(\beta_k(j))\beta_k^\star(l)\beta_k^\star(s)\Big\} \tag{30}$$

respectively. Then using equations (19) to (30), we can rewrite equation (16) as:

$$\begin{aligned}\overline{0} &= \sum_{k=1}^{K-1} U(\omega_k)\Big\{E\big[g'\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\beta_k^\star - E\big[g'\left(\langle X,\beta_K^\star\rangle + b_K^\star\right)\big]\beta_K\Big\} \\ &+ \sum_{k=1}^{K}\omega_k^\star U(b_k)E\big[g''\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\beta_k^\star + \sum_{k=1}^{K}\omega_k E\big[g^{(3)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\langle U(\beta_k),\beta_k^\star\rangle\beta_k^\star \\ &+ \sum_{k=1}^{K}\omega_k^\star E\big[g'\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]U(\beta_k),\end{aligned} \tag{31}$$

rewrite equation (17) as:

$$\begin{aligned}\overline{0}\otimes\overline{0} &= \sum_{k=1}^{K-1} U(\omega_k)\Big[E\big[g''\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\beta_k^\star\otimes\beta_k - E\big[g''\left(\langle X,\beta_K^\star\rangle + b_K^\star\right)\big]\beta_K^\star\otimes\beta_K\Big] \\ &+ \sum_{k=1}^{K}\omega_k U(b_k)E\big[g^{(3)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\beta_k^\star\otimes\beta_k^\star \\ &+ \sum_{k=1}^{K}\omega_k^\star E\big[g^{(4)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\langle U(\beta_k),\beta_k^\star\rangle\beta_k^\star\otimes\beta_k^\star \\ &+ \sum_{k=1}^{K}\omega_k^\star E\big[g''\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\Big(U(\beta_k)\otimes\beta_k^\star + \beta_k^\star\otimes U(\beta_k)\Big),\end{aligned} \tag{32}$$

and rewrite equation (18) as:

$$\begin{aligned}\overline{0}^{\otimes 3} &= \sum_{k=1}^{K-1} U(\omega_k)\Big[E\big[g^{(3)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\beta_k^{\star\otimes 3} - E\big[g^{(3)}\left(\langle X,\beta_K^\star\rangle + b_K^\star\right)\big]\beta_K^{\star\,\otimes 3}\Big] \qquad (33)\\ &+ \sum_{k=1}^{K}\omega_k^\star U(b_k)E\big[g^{(4)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\beta_k^{\star\otimes 3} + \sum_{k=1}^{K}\omega_k^\star E\big[g^{(5)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\langle U(\beta_k),\beta_k^\star\rangle\beta_k^{\star\otimes 3}\\ &+ \sum_{k=1}^{K}\omega_k E\big[g^{(3)}\left(\langle X,\beta_k^\star\rangle + b_k^\star\right)\big]\Big(U(\beta_k)\otimes\beta_k^\star\otimes\beta_k + \beta_k^\star\otimes U(\beta_k)\otimes\beta_k^\star + \beta_k^\star\otimes\beta_k^\star\otimes U(\beta_k)\Big).\end{aligned}$$

We shall first prove that the vectors $U(\beta_1),\dots,U(\beta_K)$ all belong to the linear space spanned by $\beta_1,\dots,\beta_K$.

Let J be any vector that is orthogonal to this linear space. By multiplying (33) on the right by J, and by using the fact that $\beta_1,\dots,\beta_K$ are linearly independent by (H1), we get that

$$\forall k = 1,\dots,K,\ \omega_k^\star (G_3)_k\langle U(\beta_k),J\rangle = 0.$$

Using (H1) we have $\omega_k^\star > 0$, $k = 1,\dots,K$ so that we get

$$\forall k = 1,\dots,K,\ (G_3)_k\langle U(\beta_k),J\rangle = 0.$$

Then, if (H4) holds, we get that for any k and any J, $\langle U(\beta_k),J\rangle = 0$, which proves that the vectors $U(\beta_1),\dots,U(\beta_K)$ all belong to the linear space spanned by $\beta_1,\dots,\beta_K$. Let B be the $d\times K$ matrix having $\beta_1,\dots,\beta_K$ as colomn vectors. Let $U(\beta)$ be the $d\times K$ matrix having $U(\beta_1),\dots,U(\beta_K)$ as colomn vectors. We thus have that there exists a $K\times K$ matrix $A=(A_1,\dots,A_K)$ such that $U(\beta)=BA$.
Set

$$U(\omega) = \left(U(\omega_1),\dots,U(\omega_{K-1}),-\sum_{k=1}^{K-1}U(\omega_k)\right),$$
$$U(b) = (U(b_1),\dots,U(b_K))$$

and recall that

$$\omega^\star = \left(\omega_1^\star,\dots,\omega_{K-1}^\star, 1-\sum_{k=1}^{K-1}\omega_k^\star\right).$$

Whenever R is a K-dimensional vector, denote $Diag(R)$ the $K\times K$ diagonal matrix having the R_k's on the diagonal.
Let P, Q and Δ be diagonal matrices such that $P = Diag(U(\omega))$, $Q = Diag(U(b))$ and $\Delta = Diag(\omega^\star)$. For $T\in\mathbb{R}^d$, set, $D = Diag\left(\langle\beta_1,T\rangle,\dots,\langle\beta_K,W\rangle\right)$. Then using the fact that B is full rank, (33) gives that

$$\begin{aligned}&G_3PD + G_4\Delta QD + G_5 + AG_3\Delta D + G_3\Delta DA^\top + \Delta Diag\left(\langle U(\beta_1),\beta_1\rangle,\dots,\langle U(\beta_K),\beta_K\rangle\right), BA_K\rangle)D\\ &+G_3 Diag\left(\langle U(\beta_1),\beta_1\rangle,\dots,\langle U(\beta_K),\beta_K\rangle\right) = \overline{0}\otimes\overline{0}. \qquad (34)\end{aligned}$$

Since $U(\beta)=BA$, then $U(\beta_k)=\sum_{r=1}^K \beta_r A_{rk} = BA_k$. This implies that

$$Diag(\langle U(\beta_1),\beta_1\rangle,\dots,\langle U(\beta_K),\beta_K\rangle) = Diag(\langle BA_1,\beta_1\rangle,\dots,\langle A_K,\beta_K\rangle) = \tilde{D}.$$

So (34) can be rewriten as

$$G_3PD + G_4\Delta QD + G_5\Delta\tilde{D}D + AG_3\Delta D + G_3\Delta DA^\top + G_3\Delta\tilde{D} = \overline{0}\otimes\overline{0}, \qquad (35)$$

so that for all $T\in\mathbb{R}^d$, $AG_3\Delta D + G_3\Delta DA^\top$ is a diagonal matrix. Since $G_3\Delta$ has no zero entries, this proves that, under (H1) and (H3), A is a diagonal matrix. In such a case,

$$\tilde{D} = A\tilde{B} \text{ with } \tilde{B} = Diag\left(\|\beta_1\|^2,\dots,\|\beta_K\|^2\right)$$

and (35) can be rewriten as

$$G_3PD + G_4\Delta QD + G_5\Delta A\tilde{B}D + AG_3\Delta D + G_3\Delta DA^\top + G_3\Delta A\tilde{B} = \overline{0}\otimes\overline{0}. \qquad (36)$$

But by taking, for $k = 1, \dots, K$, T_k such that $\beta_k^\top T_k = 0$, we have $D = 0$. In this case, (36) is given by

$$G_3 \Delta A \tilde{B} = \overline{0} \otimes \overline{0},$$

and using the fact that we get that G_3, Δ and $\tilde{B}$ have no zero entries we get that $A = 0$. This implies that $U(\beta_k) = 0, \;\; k = 1, 2, \dots, K$. Then using the fact that B is full rank, we conclude from (31) and (32) that

$$G_1 P + G_2 \Delta Q = \overline{0} \otimes \overline{0}, \tag{37}$$

and

$$G_2 P + G_3 \Delta Q = \overline{0} \otimes \overline{0}. \tag{38}$$

Multipliying (37) by G_3 and (38) by G_2, we have

$$G_1 G_3 P + G_2 G_3 \Delta Q = \overline{0} \otimes \overline{0}, \tag{39}$$

and

$$G_2^2 P + G_2 G_3 \Delta Q = \overline{0} \otimes \overline{0}. \tag{40}$$

Taking the difference (39)-(40), we get

$$\left(G_1 G_3 - G_2^2\right) P = \overline{0} \otimes \overline{0},$$

and since $G_1 G_3 - G_2^2$ has no zero entries, this leads to $P = 0$. Moreover, since $G_3 \Delta$ has no zero entries, this leads also $Q = 0$. Thus, under (H1), (H4) and (H5), the matrix V is full rank.

7.3 Proof of Theorem 4

If one knows the law of (Y, X), then the function

$$x \mapsto H(x) = \sum_{k=1}^{K} \omega_k g(\lambda_k \langle \mu_k, x \rangle + b_k)$$

is known on the support of X, thus on $\mathbb{R}^d$. Then the function

$$DH(x) = \sum_{k=1}^{K} \omega_k g'(\lambda_k \langle \mu_k, x \rangle + b_k) \mu_k$$

is known, and if $V \in \mathbb{R}^d$, $\lim_{t \to +\infty} \|DH(tV)\| = 0$ except in case V is orthogonal to at least one of the μ_k's. The set of $V \in \mathbb{R}^d$ such that $\lim_{t \to +\infty} \|DH(tV)\| \neq 0$ is then $\cup_{k=1}^{K} \langle \mu_k \rangle^\perp$, union of disjoint vectorial spaces of dimension $d-1$, which allows to recover the orthogonal space of $\langle \mu_k \rangle^\perp$ for all k, thus to recover $K$and all one-dimensional spaces $\langle \mu_k \rangle$. Since for all k, $\omega_k g'(b_k) > 0$, this allows to recover the μ_k's.

7.4 Proof of Theorem 5

Using Theorem 4, one knows K and the μ_k's. Since the μ_k's are linearly independent, by considering the spaces that are orthogonal to all U_k's except one, we see that the following functions are known: $h_1, \dots, h_K$ given for $j = 1, \dots, K$ by:

$$t \mapsto h_j(t) = \omega_j g(\lambda_j t + b_j) + \sum_{k=1, k \neq j}^{K} \omega_k g(b_k).$$

Then:

$$h_j(0) = \sum_{k=1}^{K} \omega_k g(b_k),$$

$$\lim_{t\to+\infty} h_j(t) = \omega_j + \sum_{k=1,k\neq j}^{K} \omega_k g(b_k),$$

$$\lim_{t\to-\infty} h_j(t) = \sum_{k=1,k\neq j}^{K} \omega_k g(b_k).$$

This allows to recover ω_j and $g(b_j)$ for $j = 1, \ldots, K$. Thus the functions

$$t \mapsto \ell_j(t) = g(\lambda_j t + b_j)$$

are known. Since $g(0) = \ell_j(-b_j/\lambda_j)$ and $g(1) = \ell_j((1-b_j)/\lambda_j)$ are fixed, one can find λ_j and b_j, and then the function g.

7.5 Proof of Theorem 6

Using Theorem 4 applied to the distributions of Y conditional to X and $Z = z$ for all $z \in \{z_1, \ldots, z_{m+1}\}$, the knowledge of $P_{g,\omega,\beta,\gamma,b}$ allows to recover K, g, ω, β, and $A_k = (a_{k,i})_{1\leq i\leq m+1}$, $k = 1, \ldots, K$, with

$$a_{k,i} = b_k + \langle \gamma_k, z_i \rangle.$$

We then know for all k

$$A_k = \begin{pmatrix} 1 & z_1^\top \\ 1 & z_2^\top \\ \vdots & \vdots \\ 1 & z_{m+1}^\top \end{pmatrix} \begin{pmatrix} b_k \\ \gamma_k \end{pmatrix}$$

which allows to recover the b_k's and γ_k's when (S4) holds.

7.6 Proof of Theorem 7

If one knows the law of Y, then, for all fixed $z \in \{z_1, \ldots, z_m\}$, one knows the function $H : (\mathbb{R}^d)^p \to (0,1)^p$ given by

$$H(x_1, \ldots, x_p) = \sum_{k=1}^{K} \omega_k \left(g(\langle \beta_k, x_j \rangle + \tilde{b}_k(z)) \right)_{1\leq j\leq p}$$

with $\tilde{b}_k(z) = b_k + \langle \gamma_k, z_i \rangle$. Let us first prove that for all z, the functions $g(\langle \beta_k, \cdot \rangle + \tilde{b}_k(z))$ are linearly independent. Indeed, if $\alpha_1, \ldots, \alpha_K$ are such that for all $x \in \mathbb{R}^d$,

$$\sum_{k=1}^{K} \alpha_k \ g(\langle \beta_k, x \rangle + \tilde{b}_k(z)) = 0,$$

then by taking the derivative, for all $x \in \mathbb{R}^d$,

$$\sum_{k=1}^{K} \alpha_k \ g'(\langle \beta_k, x \rangle + \tilde{b}_k(z))\beta_k = 0.$$

Since (S2) holds, there exists $V \in \langle \beta_k \rangle^\perp$ such that $V \notin \langle \beta_j \rangle^\perp$, $j \neq k$. Then taking $x = tV$ and t tending to infiny, we get that $\alpha_k g'(\tilde{b}_k(z))\beta_k = 0$, and then $\alpha_k = 0$.

Now, following the spectral method of proof developed in [2] to prove that multidimensional mixtures are identifiable (see also [7] Chapter 14) , we see that the knowledge of H allows to recover K, the ω_k's and, for all z, the functions $g(\langle\beta_k,\cdot\rangle+\tilde{b}_k(z))$.
Then, if one knows the function $x\mapsto g(\lambda_k\langle\mu_k,x\rangle+\tilde{b}_k(z))$ one can recover μ_k by taking the derivative, then g and the $\tilde{b}_k(z)$'s as in the proof of Theorem 5, then the γ_k's and the b_k's as in the proof of Theorem 6.

References

1. Afsari, B.: Sensitivity analysis for the problem of matrix joint diagonalization. SIAM Journal on Matrix Analysis and Applications **30**(3), 1148–1171 (2008). DOI https://doi.org/10.1137/060655997. URL `https://doi.org/10.1137/060655997`
2. Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M., Telgarsky, M.: Tensor decompositions for learning latent variables models. Journal of machine learning **15**, 2773–2832 (2014). URL `http://jmlr.org/papers/volume15/anandkumar14b/anandkumar14b.pdf`
3. Auder, B., Loum, M.A.: Morpheus: An R package to estimate parameters of logistic regressions mixtures. CRAN (2018). URL `https://cran.r-project.org/web/packages/morpheus`
4. Dang, U.J., Punzo, A., McNicholas, P.D., Ingrassia, S., Browne, R.P.: Multivariate response and parsimony for gaussian cluster-weighted models. Journal of Classification **34**(1), 4–34 (2017). DOI 10.1007/s00357-017-9221-2. URL `https://doi.org/10.1007/s00357-017-9221-2`
5. Follmann, D.A., Lambert, D.: Identifiability of finite mixtures of logistic regression models. J. Statist. Plann. Inference **27**(3), 375–381 (1991). DOI 10.1016/0378-3758(91)90050-O. URL `https://doi-org.revues.math.u-psud.fr/10.1016/0378-3758(91)90050-O`
6. Frühwirth-Schnatter, S.: Finite mixture and Markov switching models. Springer Series in Statistics. Springer, New York (2006)
7. Frühwirth-Schnatter, S., Celeux, G., Robert, C.P. (eds.): Handbook of mixture analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL (2019)
8. Grün, B., Leisch, F.: Flexmix: An R package for finite mixture modelling. R News **7**(1), 8–13 (2007). URL `http://CRAN.R-project.org/doc/Rnews/`
9. Grün, B., Leisch, F.: Finite mixtures of generalized linear regression models. In: Recent advances in linear models and related areas, pp. 205–230. Springer, Heidelberg (2008). DOI 10.1007/978-3-7908-2064-5_11. URL `https://doi-org.revues.math.u-psud.fr/10.1007/978-3-7908-2064-5_11`
10. Grün, B., Leisch, F.: Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. J. Classification **25**(2), 225–247 (2008). DOI 10.1007/s00357-008-9022-8. URL `https://doi-org.revues.math.u-psud.fr/10.1007/s00357-008-9022-8`
11. Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica **50**(4), 1029–1054 (1982). DOI 10.2307/1912775. URL `https://www.jstor.org/stable/1912775`
12. Hennig, C.: Identifiablity of models for clusterwise linear regression. Journal of Classification **17**(2), 273–296 (2000). DOI 10.1007/s003570000022. URL `https://doi.org/10.1007/s003570000022`
13. Ingrassia, S., Punzo, A.: Decision boundaries for mixtures of regressions. Journal of the Korean Statistical Society **45**(2), 295 – 306 (2016). DOI https://doi.org/10.1016/j.jkss.2015.11.005. URL `http://www.sciencedirect.com/science/article/pii/S1226319215000939`
14. Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S.C.: The generalized linear mixed cluster-weighted model. Journal of Classification **32**(1), 85–113 (2015). DOI 10.1007/s00357-015-9175-1. URL `https://doi.org/10.1007/s00357-015-9175-1`
15. Komárek, A., Komárková, L.: Clustering for multivariate continuous and discrete longitudinal data. Ann. Appl. Stat. **7**(1), 177–200 (2013). DOI 10.1214/12-AOAS580. URL `https://doi-org.revues.math.u-psud.fr/10.1214/12-AOAS580`
16. Loum, M.A., Poursat, M.A., Sow, A., Amadou Alpha Sall, C.L., Gassiat, E.: Multinomial logistic model for coinfection diagnosis between arbovirus and malaria in kedougou. The International Journal of Biostatistics **15**(2) (2019). DOI https://doi.org/10.1515/ijb-2017-0015. URL `https://www.degruyter.com/view/journals/ijb/15/2/article-20170015.xml`
17. Maruotti, A., Punzo, A.: Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. Computational Statistics & Data Analysis **113**, 475 – 496 (2017). DOI https://doi.org/10.1016/j.csda.2016.05.024. URL `http://www.sciencedirect.com/science/article/pii/S0167947316301360`
18. Mazza, A., Punzo, A.: Mixtures of multivariate contaminated normal regression models. Statistical Papers (2017). DOI 10.1007/s00362-017-0964-y. URL `https://doi.org/10.1007/s00362-017-0964-y`
19. Mazza, A., Punzo, A., Ingrassia, S.: flexcwm: A flexible framework for cluster-weighted models. Journal of Statistical Software, Articles **86**(2), 1–30 (2018). DOI 10.18637/jss.v086.i02. URL `https://www.jstatsoft.org/v086/i02`
20. Punzo, A., Ingrassia, S., Maruotti, A.: Multivariate generalized hidden markov regression models with random covariates: Physical exercise in an elderly population. Statistics in Medicine **37**, 2797–2808. (2018). DOI https://doi.org/10.1002/sim.7687. URL `https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7687`
21. Punzo, A., McNicholas, P.D.: Robust clustering in regression analysis via the contaminated gaussian cluster-weighted model. Journal of Classification **34**(2), 249–293 (2017). DOI 10.1007/s00357-017-9234-x. URL `https://doi.org/10.1007/s00357-017-9234-x`
22. Sedghi, H., Janzamin, M., Anandkumar, A.: Provable tensor methods for learning mixtures of generalized linear models. In: A. Gretton, C.C. Robert (eds.) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 51, pp. 1223–1231. PMLR, Cadiz, Spain (2016)

23. Vaart, A.W.v.d.: Asymptotic Statistics, cambridge edn. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (1998). DOI 10.1017/CBO9780511802256
24. Wang, S., Yao, W., Huang, M.: A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. Statist. Probab. Lett. **93**, 41–45 (2014). DOI 10.1016/j.spl.2014.06.010. URL `https://doi-org.revues.math.u-psud.fr/10.1016/j.spl.2014.06.010`
25. Wu, C.F.J.: On the convergence properties of the em algorithm. The Annals of Statistics **11**(1), 95–103 (1983). URL `http://www.jstor.org/stable/2240463`