Mixture of generalized linear models: identifiability and applications

Benjamin Auder, Elisabeth Gassiat, Mor Absa Loum

Mixture of generalized linear models: identifiability and applications

Benjamin Auder, Elisabeth Gassiat, Mor Absa Loum

October 17, 2018

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, France

Abstract

We consider finite mixtures of generalized linear models with binary output. We prove that cross moments till order 3 are sufficient to identify all parameters of the model. We propose a least squares estimation method and we prove the consistency and the Gaussian asymptotic behavior of the estimator. An R-package is developed to apply our method, we give numerical experiments to compare with likelihood methods. We then provide new identifiability results for several finite mixtures of generalized linear models with binary output and unknown link function including both continuous and categorical covariates, and possibly longitudinal data.

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 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 mixtures models, see for instance the book [5]. Most of them use likelihood methods or Bayesian methods that are likelihood dependent. Indeed, the now 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 many enough initial points. Recently, spectral methods were developed to bypass EM algorithms and they were proved able to recover the directions of the regression parameter in models with known link function and random covariates, see [9].

One aim of this paper is to extend such moment methods using least squares to get estimators of the whole parameters, and to provide theoretical guarantees of this estimation method. The setting is that of regression models with binary outputs, random covariates and known link function, detailed in Section 2. We first prove that cross moments up to order 3 between the output and the regression variables are enough to recover all the parameters of the model, see Theorem 1 for the probit link function and Theorem 2 for general link functions. We then obtain consistency and asymptotic normality of our least squares estimators as usual, see Theorem 3. The algorithm is described at the end of Section 3,

and to apply it, we developed the R-package `morpheus` available on the CRAN ([3]). We then compare experimentally our method to the maximum likelihood estimator computed using the R-package flexmix ([6]). We show that our estimator may be better for the probit link function with finite samples when the dimension increases, though keeping very small computation times when that of flexmix increases with dimension. The experiments are presented in Section 4.

Another aim of this paper is to investigate identifiability in various mixture of non linear regression models with binary outputs. Indeed, identifiability results for such models are still few and not enough to give theoretical guarantees of available algorithms. Let us review what is known up to our knowledge. In [4], the identifiability is proved for finite mixtures of logistic regression models where only the intercept varies with the population [11]. In [7], finite mixtures of multinomial logit models with varying and fixed effects are investigated, the proofs of identifability results use the explicit form of the logit function. In [11], 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 data (Bernoulli models).

We provide in Section 5 several identifiability results, that for example are useful to get theoretical guarantees in applications such as the one in [8]. We prove that with known smooth enough link function, the directions of the covariates 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 non parametrically recovered, see Theorem 5. We then study the simultaneous use of continuous and categorical covariates and further give assumptions under which parameters and 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.

2 Model and notations

Let us denote $[n]$ the set $\{1, 2, \dots, n\}$ 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 $\mathbb{R}^d$ and for $p = 2$, T is a $d \times d$ real matrix. The $(i_1, i_2, \dots, i_p)$-th coordinate of T with respect the canonical basis is denoted $T[i_1, i_2, \dots, i_p]$, $i_1, i_2, \dots, i_p \in [d]$.

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 $\omega_j \geq 0$, $j = 1, \dots, K$ and $\sum_{j=1}^K \omega j = 1$. Define, for $j = 1, \dots, K$, 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 $\omega = (\omega_1, \dots, \omega_K)$, $b = (b_1, \cdots, b_K)$, $\beta = [\beta_1 | \cdots, | \beta_K]$ the $d \times K$ matrix of regression coefficients and denote $\theta = (\omega, \beta, b)$. The

model of population mixture of binary regressions is given by:

$$\mathbb{P}_\theta(Y=1|X=x) = \sum_{k=1}^{K} \omega_k g(<\beta_k, x> + b_k). \tag{1}$$

We assume that the random variable X has a Gaussian distribution. We now focus on the situation where $X \sim \mathcal{N}(0, I_d)$, I_d being the identity $d \times d$ matrix. 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 symetric $d \times d$ matrix.

Define the cross moments between the response Y and the covariable X, up to order 3:

- $M_1(\theta) := \mathbb{E}_\theta[Y.X]$, first-order moment,
- $M_2(\theta) := \mathbb{E}_\theta\Big[Y.\big(X \otimes X - \sum_{j\in[d]} Y.e_j \otimes e_j\big)\Big]$, second-order moment and
- $M_3(\theta) := \mathbb{E}_\theta\Big[Y\big(X \otimes X \otimes X - \sum_{j\in[d]} \big[X \otimes e_j \otimes e_j + e_j \otimes X \otimes e_j + e_j \otimes e_j \otimes X\big]\big)\Big]$ third-order moment.

Let, for $k = 1, \ldots, K$, $\lambda_k = \|\beta_k\|$ and $\mu_k = \beta_k/\|\beta_k\|$. Using Stein's identity, Anandkumar et al. ([9]) prove the following lemma:

Lemma 1 ([9])**.** *Under enough smoothness and integrability of the link function (which hold for the logit and probit link functions, or under our assumption (H3) below) the moments can be rewritten:*

$$\begin{aligned}
M_1(\theta) &= \sum_{k=1}^{K} \omega_k \lambda_k \mathbb{E}[g'\big(\lambda_k \langle X, \mu_k\rangle + b_k\big)]\ \mu_k, \\
M_2(\theta) &= \sum_{k=1}^{K} \omega_k \lambda_k^2 \mathbb{E}[g''\big(\lambda_k < X, \mu_k > + b_k\big)]\ \mu_k \otimes \mu_k, \\
M_3(\theta) &= \sum_{k=1}^{K} \omega_k \lambda_k^3 \mathbb{E}[g^{(3)}\big(\lambda_k < X, \mu_k > + b_k\big)]\ \mu_k \otimes \mu_k \otimes \mu_k.
\end{aligned}$$

It is proved in [9] that the knowledge of $M_3(\theta)$ leads to the knowledge of $\mu_1, \ldots, \mu_K$ up to their sign as soon as they are linearly independent. In the next section, we prove that the knowledge of all cross moments till order 3 allows to recover all parameters for the probit link function under the same assumption on the regression coefficients. We also prove that for a general link function satisfying some weak assumption, the knowledge of all cross moments till order 3 allows to recover all parameters provided they are not too far from 0.

3 Moment identifiability and estimation

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, 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 any functions L_s, $s = 1, 2, 3$, such that $\forall\ z \in \mathbb{R}$, $\forall\ (\lambda, b) \in \mathcal{O}$, we have

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

and further for $s = 1, 2, 3$,

$$\int_{\mathbb{R}} L_s(z) e^{-z^2/2} dz < +\infty.$$

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.
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 $x > 0$.

To prove the limiting Gaussian distribution of our moment estimator, we shall need more assumptions. For $j = 1, \ldots, 5$, let G_j be the $K \times K$ diagonal matrix having the $\mathbb{E}[g^{(j)}\left(\langle \beta_k, X\rangle + b_k\right)]$'s on the diagonal.

– (H4) All diagonal coefficients of G_3 are non zero.
– (H5) All diagonal coefficients of $G_1 G_3 - G_2^2$ are non zero.

3.1 Identifiability results

In the whole section we assume that (H1) holds. Under (H1), we see by Lemma 1 that K is the rank of $M_2(\theta)$.
It is proved in [9] we can recover the μ_k's up to sign from the knowledge of $M_2(\theta)$ and $M_3(\theta)$, but since under (H1) $M_1(\theta)$ is a linear combination of the μ_k's with positive coefficients, 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

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

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

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

Then it is equivalent to prove that the knowledge of

$$\big(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\big) := \left(\int z dQ_{(b,\lambda)}(z); \int z^2 dQ_{(b,\lambda)}(z)\right) \tag{3}$$

implies the knowledge of (b,λ). When g is the probit link function, $Q_{(b,\lambda)}$ is a Gaussian distribution and computations detailed in Section 6.1 leads to the following identifiability result.

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 $M_3(\theta)$.*

In the general situation, identifiability holds at least in an open set. To prove it, one just has 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 $\big(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\big)$ is one-to-one on its image. This leads to the following identifiability result whose proof is postponed to Section 6.2.

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. However, numerical study of the function $(b,\lambda) \mapsto \big(E_{(b,\lambda)}(Z); E_{(b,\lambda)}(Z^2)\big)$ when the link function g is the logit function shows that identifiability seems to hold at least with $L = 8$ and $B = 8$.

3.2 The least squares moment estimator

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

$$\begin{aligned}
\widehat{M_1} &= \frac{1}{n}\sum_{i=1}^n Y_i X_i \\
\widehat{M_2} &= \frac{1}{n}\sum_{i=1}^n \Big[Y_i(X_i\otimes X_i - \sum_{j\in[d]} e_j\otimes e_j)\Big] \\
\widehat{M_3} &= \frac{1}{n}\sum_{i=1}^n \Big[Y_i(X_i\otimes X_i\otimes X_i - \sum_{j\in[d]} \big[X_i\otimes e_j\otimes e_j + e_j\otimes X_i\otimes e_j + e_j\otimes e_j\otimes X_i\big])\Big].
\end{aligned}$$

It is not possible to match the empirical moments exactly, so that we use a least-squares estimator. Define for all θ:

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

and the estimator

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} Q_n(\theta). \tag{4}$$

Theorem 3. *Assume that (H1), (H2), (H3) hold, and that Θ is compact and included in the set of identifiable parameters. Then $\widehat{\theta}_n$ is consistent.*
If moreover (H4) and (H5) hold, then $\sqrt{n}\left(\widehat{\theta}_n - \theta^\star\right)$ converges in distribution under $\mathbb{P}_{\theta^\star}$ to a centered Gaussian distribution.

The proof of Theorem 3 is detailed in Section 6.3 and follows the usual analysis of the asymptotic behavior of Z-estimators, the more delicate part of the proof being to prove that the Hessian of $Q_n(\theta)$ has an invertible limiting value.

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 search the minimizer of Q_n using usual optimization methods. 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 $Q_n(\theta)$ using the estimators of 1. as initial directions
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, we get

$$B(z) = \sum_{k=1}^{K} r^{(3)} \omega_k \lambda_k^3 \mathbb{E}[g^{(3)}\big(\lambda_k < X, \mu_k > + b_k\big)] \langle \mu_k, z \rangle \mu_k^{\otimes 2}.$$

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 $z_1, \ldots, z_P$, $P \geq 2$. Joint diagonalisation of $B(z_1), \ldots B(z_P)$ means finding a matrix V such that the matrices $VB(z_p)V^T$ 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 $\widehat{M_3}$.

4 Simulations

4.1 R package

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

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 $\mathbb{R}^d$)

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

4 : Joint diagonalisation: compute V such that $V\widehat{B}(z_p)V^T$ 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 $O = ginv(U)\widehat{M_1}$

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

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

1. the computation of the directions matrix μ, based on the empirical cross-moments as described in the previous sections;
2. the optimization of all parameters (including μ), 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 Experiments

In this section, we evaluate our algorithm in a first step using mean squared error (MSE). In a second step, we compare experimentally our moments method (morpheus package [3]) and the likelihood method (with felxmix package [6]). We arbitrarily choose the parameters for the simulations, which should be discovered by the algorithms (ours, and the likelihood algorithm).

Experiment 1 (dimension 2):

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

Experiment 2 (dimension 5):

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

Experiment 3 (dimension 10):

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

For all three experiments we use both logit and probit links. Computations are always run on the same data both for our package and flexmix − which is a reference for this kind of estimation, using an iterative algorithm to maximize the log-likelihood. Results are aggregated over $N = 1000$ Monte-Carlo runs.

Mean squared error (MSE). Graphical representations of the MSE versus the sample size (n) are given in figures 1, 2 and 3. In each figure, we represent the MSE associated with each parameter vector versus the sample size. We can see that the goodness of the estimation depends on the sample size and that enough observations is needed to properly estimate the parameters. We have from figure 1 and figure 2 that for our moment method, with dimension less or equal to 5, the necessary sample size is around of 10^5. For large dimension, figures 3 show that 10^6 is not enough to estimate the parameters vectors β. Indeed our estimation method use the spectral estimator, which need more enough data, as starting point.

Algorithms performance. To evaluate algorithms performance, the total number of sample points is fixed to $n = 10^5$. This value still enough to observe correct performances, yet small enough to remain realistic.

On the figures 4 and 5, all (true) parameters (p, b, β) are re-ordered in a real vector, of size $K \times (d+2) - 1$. This vector is plotted as a line to improve visualization experience, but it must be noted that this does not represent any curve data. Dotted lines corresponds to the computed values plus or minus one standard deviation, and computed values themselves are represented with long dashed lines. The leftmost column corresponds to experiment 1

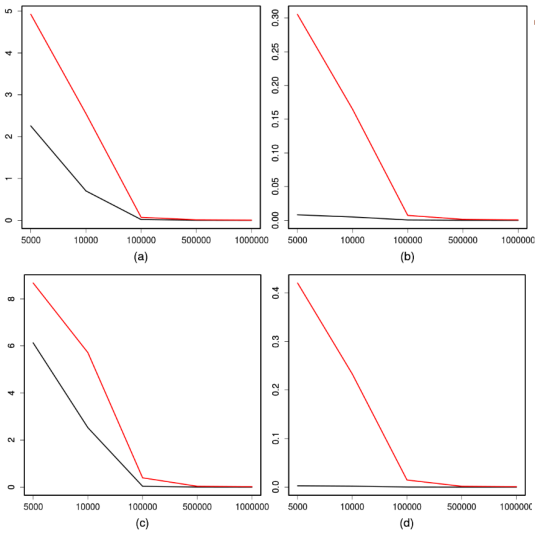

Figure 1: Experiment 1; Top: *logit* link function $\Big($(a) MSE($\widehat{\beta}$), (b) MSE($\widehat{b}$,$\widehat{p}$)$\Big)$, bottom: *probit* link function: $\Big($(c) MSE($\widehat{\beta}$), (d) MSE(($\widehat{b}$,$\widehat{p}$)$\Big)$.

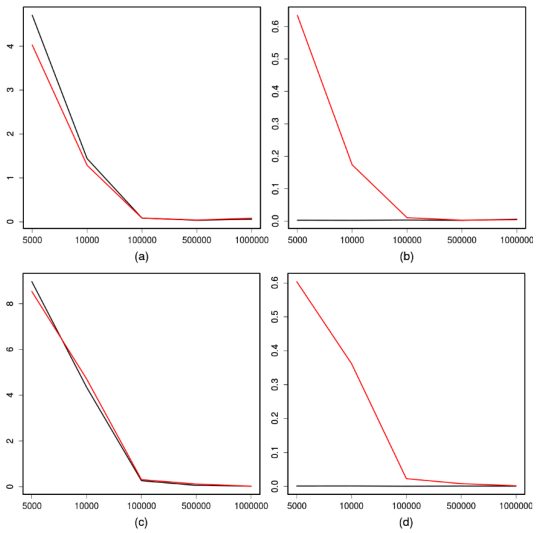

Figure 2: Experiment 2; Top: *logit* link function $\Big($(a) MSE($\widehat{\beta}$), (b) MSE($\widehat{b}$,$\widehat{p}$)$\Big)$, bottom: *probit* link function: $\Big($ (c) MSE($\widehat{\beta}$), (d) MSE(($\widehat{b}$,$\widehat{p}$)$\Big)$.

($d = 2$), the middle one to experiment 2 ($d = 5$), and the rightmost column corresponds to experiment 3 ($d = 10$).
Figure 4: logit link. While most of the times flexmix finds a better solution than our proposed algorithm (smaller variance), both methods are good on average for $d \leq 5$. The case $d = 10$ is not handled well neither by the flexmix package nor by our package (the latter showing

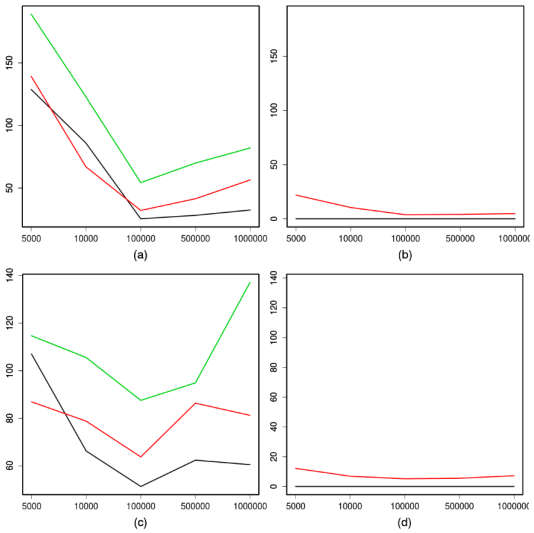

Figure 3: Experiment 3; Top: *logit* link function $\Big($(a) MSE($\widehat{\beta}$), (b) MSE($\widehat{b}$,$\widehat{p}$)$\Big)$, bottom: *probit* link function: $\Big($ (c) MSE($\widehat{\beta}$), (d) MSE(($\widehat{b}$,$\widehat{p}$)$\Big)$.

even poorer accuracy). Indeed in this relatively high dimension the number of observations should be much higher.

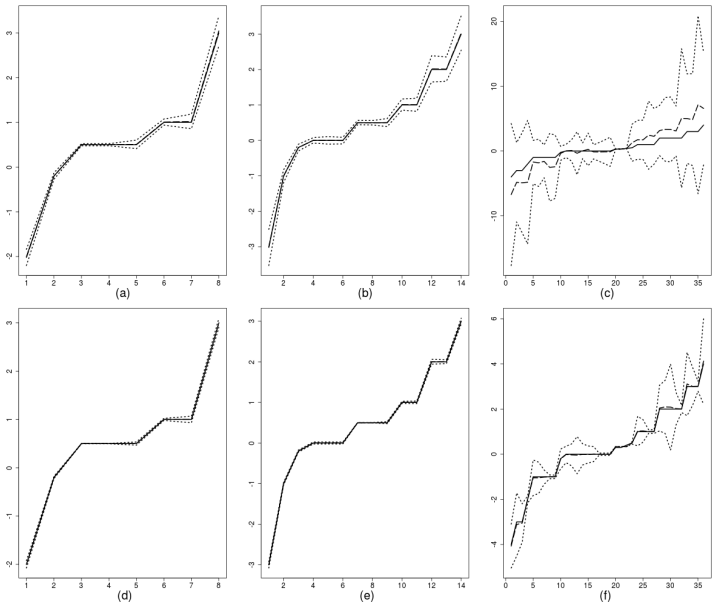

Figure 4: *Logit* link function. Top: our package, bottom: flexmix. From left to right: experiment 1, 2 and 3 respectively

Figure 5: probit link. In this case our algorithm performs slightly better than its flexmix

counterpart for $d \leq 5$. However, again, the variance in the case $d = 10$ is way too high − in fact even the average value is generally wrong, when coefficients are non-zero. We can increase n by a factor 100 to obtain more accurate results.

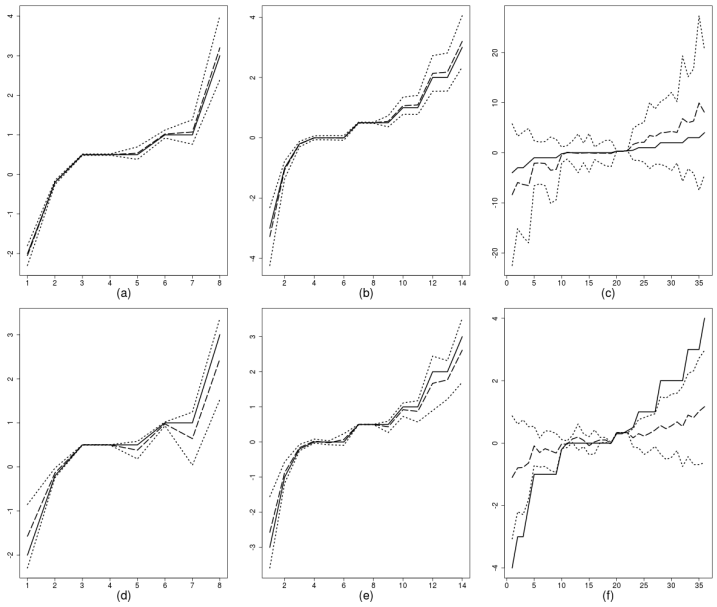

Figure 5: *Probit* link function. Top: our package, bottom: flexmix. From left to right: experiment 1, 2 and 3 respectively

Considering both links with $d \leq 5$, the algorithms performances are comparable with a global small advantage to our method. The case $d = 10$ is handled better by the flexmix algorithm, although clearly not well. Finally, concerning our method we observe a tendancy to overestimate large parameters while underestimating small ones. This observation is inversed for flexmix.

Computational time. It must be noted that our algorithm timing does not depend on n, since it operates on matrices of size $O(d \times K)$. Thus it can be more suitable for very large datasets, where the variability is clearly reduced. Figure 6 (time by seconds versus $\log_{10} n$) illustrate this fact: the timings clearly favor our package − because increasing n has almost no impact on the running time. However, flexmix timings are not that high: just a few minutes for the longest run, on average, on one million sample points. To obtain the data shown on the figure we averaged 1000 runs on random parameters.

5 Some other identifiability results

In this section, we provide several further identifiability results for mixtures of generalized linear models (GLMs) under various assumptions.

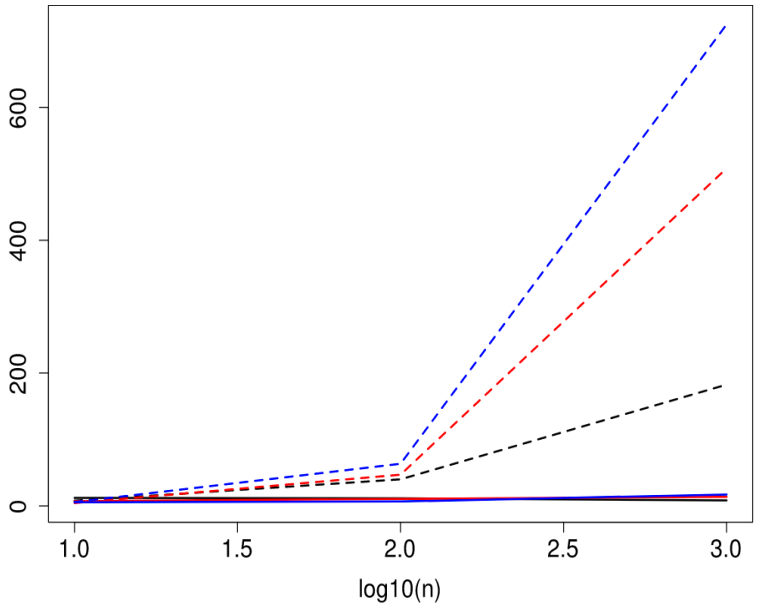

Figure 6: Timings versus sample size. Dotted lines: flexmix; others: our algorithm. Black for $d = 2$, red for $d = 5$ and blue for $d = 10$.

5.1 Continuous covariates

We first consider the setting where the random vector (X, Y), $X \in \mathbb{R}^d$, $Y \in \mathbb{R}$ is such that

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

We assume that for all k, $\omega_k \geq 0$, that $\sum_{k=1}^{K} \omega_k = 1$, and that g takes value in $(0, 1)$. In case Y takes binary values, this is exactly the model we considered in the previous sections. 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 $\mathbb{P}_{g,\omega,\beta,b}$ 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 chioose to fix $g(0)$ and $g(1)$ (with no loss of generality). Denot $\mu_k = \beta_k/\|\beta_k\|$ 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 $\mathbb{P}_{g,\omega,\beta,b}$ allows to recover K and $\mu_1, \ldots, \mu_K$.*

Proof.
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.

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 $\mathbb{P}_{g,\omega,\beta,b}$ allows to recover K, g, ω, β and b.*

Proof.
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: $h_1, \ldots, h_K$ given for $j = 1, \ldots, K$ 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:

$$\begin{aligned} h_j(0) &= \sum_{k=1}^{K} \omega_k g(b_k), \\ \lim_{t \to +\infty} h_j(t) &= \omega_j + \sum_{k=1, k \neq j}^{K} \omega_k g(b_k), \\ \lim_{t \to -\infty} h_j(t) &= \sum_{k=1, k \neq j}^{K} \omega_k g(b_k). \end{aligned}$$

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.

5.2 Continuous and categorical covariates

We now consider the situation where part of the covariates are catergorical, we denote them Z, and $\{z_1, \ldots, z_m\} \subset \mathbb{R}^{d'}$ their possible values. 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 $\mathbb{P}_{g,\omega,\beta,\gamma,b}$ the probability distribution of (X,Y), with $\omega = (\omega_1, \cdots, \omega_K)$, $\beta = [\beta_1|, \cdots, |\beta_K] \in \mathbb{R}^{d\times K}$, $\gamma = [\gamma_1|, \cdots, |\gamma_K] \in \mathbb{R}^{d'\times K}$, and $b = (b_1, \cdots, b_K) \in \mathbb{R}^K$. We introduce

– (S4) The matrix $\begin{pmatrix} 1 & z_1^T \\ 1 & z_2^T \\ \vdots & \vdots \\ 1 & z_m^T \end{pmatrix}$ is full rank.

Remark: (S4) implies that $d' + 1 \leq m$.

It is the continuous covariates that allow to identify g.

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

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

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

We then know for all k

$$A_k = \begin{pmatrix} 1 & z_1^T \\ 1 & z_2^T \\ \vdots & \vdots \\ 1 & z_m^T \end{pmatrix} \begin{pmatrix} b_k \\ \gamma_k \end{pmatrix}$$

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

5.3 Longitudinal observations

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

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

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 $p \geq 3$. If (S1), (S2), (S3) and (S4) hold, then the model is identifiable: the knowledge of $\mathbb{P}_{g,\omega,\beta,\gamma,b}$ allows to recover K, g, ω, β, γ and b.*

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

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

with $\tilde{b}_k(z) = b_k + \langle \gamma_k, z_i \rangle$. Let us firts prove that for all z, the functions $g(\langle \beta_k, \cdot \rangle + \tilde{b}_k(z))$ are linearly independent. Indeed, if $\alpha_1, \dots, \alpha_K$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, 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.

5.4 Some perspectives

Identifiability of a model is a first step to obtain theoretical guarantees for practical estimation procedures. In this paper, we proposed one moment method as an estimation strategy in the particular case of binary outcomes and gaussian covariates, for which we proved the asymptotic Gaussian behaviour. As soon as the identifiability of a model is known, any reasonable estimation strategy leads to consistent estimators. Considering the non parametric estimation of the link function, model selection methods should lead to well behaved estimators. 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 Proofs

6.1 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 (2), 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

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

We can then recover b and λ by

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

and

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

6.2 Proof of Theorem 2

Using (H3) and integration by parts, we get that for (λ, b) in a neighborhood of $(0,0)$

$$\left(\mathbb{E}_{(b,\lambda)}(Z); \mathbb{E}_{(b,\lambda)}(Z^2)\right) = \left(\frac{\lambda \int g''(\lambda z + b) e^{-z^2/2} dz}{\int g'(\lambda z + b) e^{-z^2/2} dz}; 1 + \frac{\lambda^2 \int g^{(3)}(\lambda z + b) e^{-z^2/2} dz}{\int g'(\lambda z + b) e^{-z^2/2} dz}\right). \quad (5)$$

Using (H2),

- (P3) $g'(0) > 0$
- (P4) $g''(0) = g^{(4)}(0) = 0$

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

$$\begin{array}{rccl} K_s & : & \mathbb{R}_+^\star \times \mathbb{R} & \to & \mathbb{R} \\ & & (\lambda, b) & \mapsto & K_s(\lambda, b) = \int g^{(s)}(\lambda z + b) e^{-z^2/2} dz \end{array}$$

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

$$K_s(\lambda, b) = K_s(0,0) + \langle \nabla K_s(0,0), (\lambda, b) \rangle + o(\lambda^2 + b^2). \quad (6)$$

Now

$$\frac{\partial K_s}{\partial \lambda}(0,0) = \int z g^{(s+1)}(0) e^{-z^2/2} dz \quad (7)$$

and

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

so that

$$K_s(\lambda, b) = g^{(s)}(0) \int e^{-z^2/2} dz + g^{(s+1)}(0) \int (\lambda z + b) e^{-z^2/2} dz + o(\lambda^2 + b^2). \quad (9)$$

Using $(P4)$ and (9), we have

$$\int g'(\lambda z + b) e^{-z^2/2} dz = \sqrt{2\pi} g'(0) + o(\lambda^2 + b^2), \quad (10)$$

$$\int g''(\lambda z+b)e^{-z^2/2}dz = \sqrt{2\pi}g^{(3)}(0)b + o(\lambda^2+b^2), \tag{11}$$

and

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

Therefore, replacing (10) to (12) in (5), 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 $(\lambda, b) \mapsto (E_{(b,\lambda)}(Z), E_{(b,\lambda)}(Z^2))$ is invertible in a neighborhood of $(0,0)$.

6.3 Proof of Theorem 3

Let θ^* be the true value of the parameter. For each θ, by the law of large numbers, $Q_n(\theta)$ converges to

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

Define

$$S = \sup_{\theta\in\Theta}\Big|Q_n(\theta) - Q(\theta)\Big|.$$

Since $Q(\theta)$ has θ^* as unique minimum (up to label switching), to prove the consistency of $\widehat{\theta}_n$, it is enough to prove that S converges to 0 in probability, see Theorem 5.7 in [10]. We easily get

$$\begin{aligned}
S \;\leq\; & \sum_{j\in[d]}\Big(\Big|\hat{M}_1[j] - M_1(\theta^\star)[j]\Big|\Big)\Big(\big|\hat{M}_1[j]\big| + \big|M_1(\theta^*)[j]\big| + 2\sup_{\theta\in\Theta}\big|M_1(\theta)[j]\big|\Big) \\
+ \; & \sum_{j,k\in[d]}\Big(\Big|\hat{M}_2[j,k] - M_2(\theta^\star)[j,k]\Big|\Big)\Big(\big|\hat{M}_2[j,k]\big| + \big|M_2(\theta^*)[j,k]\big| + 2\sup_{\theta\in\Theta}\big|M_2(\theta)[j,k]\big|\Big) \\
+ \; & \sum_{j,k,l\in[d]}\Big(\Big|\hat{M}_3[j,k,l] - M_3(\theta^\star)[j,k,l]\Big|\Big)\Big(\big|\hat{M}_3[j,k,l]\big| + \big|M_3(\theta^*)[j,k,l]\big| + 2\sup_{\theta\in\Theta}\big|M_3(\theta)[j,k,l]\big|\Big),
\end{aligned}$$

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

$$\begin{aligned} S &\leq \sum_{j\in[d]} \Big(c_1 + \big|\hat{M}_1[j]\big|\Big)\Big(\Big|\hat{M}_1[j] - M_1(\theta^\star)[j]\Big|\Big) \\ &+ \sum_{j,k\in[d]} \Big(c_2 + \big|\hat{M}_2[j,k]\big|\Big)\Big(\Big|\hat{M}_2[j,k] - M_2(\theta^\star)[j,k]\Big|\Big) \\ &+ \sum_{j,k,l\in[d]} \Big(c_3 + \big|\hat{M}_3[j,k,l]\big|\Big)\Big(\Big|\hat{M}_3[j,k,l] - M_3(\theta^\star)[j,k,l]\Big|\Big) \end{aligned}$$

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 $Z_n(\theta) = \nabla_\theta Q_n(\theta)$. The $r-th$ coordinate of $Z_n(\theta)$ can be obtained by

$$\begin{aligned} \frac{\partial Q_n(\theta)}{\partial \theta_r} &= -2\Big\{ \sum_{j\in[d]} \frac{\partial M_1(\theta)[j]}{\partial \theta_r}\Big[\hat{M}_1[j] - M_1(\theta)[j]\Big] \\ &+ \sum_{j,k\in[d]} \frac{\partial M_2(\theta)[j,k]}{\partial \theta_r}\Big[\hat{M}_2[j,k] - M_2(\theta)[j,k]\Big] \\ &+ \sum_{j,k,l\in[d]} \frac{\partial M_3(\theta)[j,k,l]}{\partial \theta_r}\Big[\hat{M}_3[j,k,l] - M_3(\theta)[j,k,l]\Big]\Big\} \end{aligned}$$

Using Taylor expansion, we get

$$Z_n(\hat{\theta}_n) = Z_n(\theta^*) + \int_0^1 D_1 Z_n\big[\theta^* + t(\hat{\theta}_n - \theta^*)\big]\Big(\hat{\theta}_n - \theta^*\Big)dt \tag{13}$$

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

$$-\sqrt{n}Z_n(\theta^*) = \Big[\int_0^1 D_1 Z_n\big[\theta^* + t(\hat{\theta}_n - \theta^*)\big]dt\ \Big]\sqrt{n}\Big(\hat{\theta}_n - \theta^*\Big) \tag{14}$$

Let us set

$$\hat{M} = \Big(\hat{M}_1[j], \hat{M}_2[j,k], \hat{M}_3[j,k,l]\Big)_{1\leq j,k,l\leq d}$$

and

$$M(\theta^\star) = \Big(M_1(\theta^\star)[j], M_2(\theta^\star)[j,k], M_3(\theta^\star)[j,k,l]\Big)_{1\leq j,k,l\leq d}$$

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

The $(r_1, r2) - th$ coordinate of $D_1 Z_n(\theta) = \nabla^2_\theta Q_n(\theta)$ are given by

$$\begin{aligned} \frac{\partial^2 Q_n(\theta)}{\partial \theta_{r_1}\partial \theta_{r_2}} &= -2\sum_{j\in[d]} \frac{\partial^2 M_1(\theta)[j]}{\partial \theta_{r_1}\partial \theta_{r_2}} \times \big[\hat{M}_1[j] - M_1(\theta)[j]\big] - 2\sum_{j,k\in[d]} \frac{\partial^2 M_2(\theta)[j,k]}{\partial \theta_{r_1}\partial \theta_{r_2}} \times \big[\hat{M}_2[j,k] - M_2(\theta)[j,k]\big] \\ &- 2\sum_{j,k,l\in[d]} \frac{\partial^2 M_3(\theta)[j,k,l]}{\partial \theta_{r_1}\partial \theta_{r_2}} \times \big[\hat{M}_3[j,k,l] - M_3(\theta)[j,k,l]\big] + V_{r_1 r_2}(\theta) \end{aligned}$$

with

$$\begin{aligned} V_{r_1 r_2}(\theta) &= 2\sum_{j\in[d]} \frac{\partial M_1(\theta)[j]}{\partial \theta_{r_1}} \times \frac{\partial M_1(\theta)[j]}{\partial \theta_{r_2}} + 2\sum_{j,k\in[d]} \frac{\partial M_2(\theta)[j,k]}{\partial \theta_{r_1}} \times \frac{\partial M_2(\theta)[j,k]}{\partial \theta_{r_2}} \\ &+ 2\sum_{j,k,l\in[d]} \frac{\partial M_3(\theta)[j,k,l]}{\partial \theta_{r_1}} \times \frac{\partial M_3(\theta)[j,k,l]}{\partial \theta_{r_2}}. \end{aligned}$$

It is not difficult to prove that $\int_0^1 D_1 Z_n\big[\theta^* + t(\hat{\theta}_n - \theta^*)\big]dt$ converges in probability to $V(\theta^*)$ 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. Using the form of V we get that $U^T V U = 0$ if and only if:

$$U^T DM_1(\theta)[j] = 0,\ j = 1,\ldots,q, \tag{15}$$

and

$$U^T DM_2(\theta)[j,l] = 0,\ j,l = 1,\ldots,q, \tag{16}$$

and

$$U^T DM_3(\theta)[j,l,m] = 0,\ j,l,m = 1,\ldots,q. \tag{17}$$

Here, $DM_.[]$ is the gradient vector of the involved coordinate of $M_.$. 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, $\overline{0}\otimes\overline{0}$ the $d\times d$-dimensional zero matrix and $\overline{0}\otimes\overline{0}\otimes\overline{0}$ the $d\times d\times d$-dimensional zero third order tensor. Then, the equation (15) can be rewritten as:

$$U^T DM_1(\theta)[j] = \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_1(\theta)[j]}{\partial \omega_k} + \sum_{k=1}^{K} U(b_k)\frac{\partial M_1(\theta)[j]}{\partial b_k} + \sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_1(\theta)[j]}{\partial \beta_{mk}}, \tag{18}$$

the equation (16) can be rewritten as

$$U^T DM_2(\theta)[j,l] = \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_2(\theta)[j,l]}{\partial \omega_k} + \sum_{k=1}^{K} U(b_k)\frac{\partial M_2(\theta)[j,l]}{\partial b_k} + \sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_2(\theta)[j,l]}{\partial \beta_{mk}}, \tag{19}$$

and the equation (17) can be rewritten as

$$U^T DM_3(\theta)[j,l,m] = \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_3(\theta)[j,l,m]}{\partial \omega_k} + \sum_{k=1}^{K} U(b_k)\frac{\partial M_3(\theta)[j,l,m]}{\partial b_k} + \sum_{k=1}^{K}\sum_{s=1}^{d} U(\beta_{sk})\frac{\partial M_3(\theta)[j,l,m]}{\partial \beta_{sk}}. \tag{20}$$

Using the fact that $\sum_{k=1}^{d}\omega_k = 1$, the first terms of the equations (18) to (20) are rewritten as:

$$\begin{aligned} \sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_1(\theta)[j]}{\partial \omega_k} &= \sum_{k=1}^{K-1} U(\omega_k)\Big\{\mathbb{E}\left[g'\left(\langle x,\beta_k\rangle + b_k\right)\right].\beta_k(j) \\ &- \mathbb{E}\left[g'\left(\langle x,\beta_K\rangle + b_K\right)\right].\beta_K(j)\Big\}, \end{aligned} \tag{21}$$

$$\sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_2(\theta)[j,l]}{\partial \omega_k} = \sum_{k=1}^{K-1} U(\omega_k)\Big\{\mathbb{E}\left[g''\left(\langle x,\beta_k\rangle + b_k\right)\right].\beta_k(j)\beta_k(l) - \mathbb{E}\left[g''\left(\langle x,\beta_K\rangle + b_K\right)\right].\beta_K(j)\beta_K(l)\Big\} \quad (22)$$

and

$$\sum_{k=1}^{K-1} U(\omega_k)\frac{\partial M_3(\theta)[j,l,m]}{\partial \omega_k} = \sum_{k=1}^{K-1} U(\omega_k)\Big\{\mathbb{E}\left[g^{(3)}\left(\langle x,\beta_k\rangle + b_k\right)\right].\beta_k(j)\beta_k(l)\beta_k(m) - \mathbb{E}\left[g^{(3)}\left(\langle x,\beta_K\rangle + b_K\right)\right].\beta_K(j)\beta_K(l)\beta_K(m)\Big\} \quad (23)$$

respectively. Likewise the seconds terms of equations (18) to (20) are rewritten as:

$$\sum_{k=1}^{K} U(b_k)\frac{\partial M_1(\theta)[j]}{\partial b_k} = \sum_{k=1}^{K} \omega_k U(b_k)\mathbb{E}\left[g''\left(\langle x,\beta_k\rangle + b_k\right)\right].\beta(j), \quad (24)$$

$$\sum_{k=1}^{K} U(b_k)\frac{\partial M_2(\theta)[j,l]}{\partial b_k} = \sum_{k=1}^{K} \omega_k U(b_k)\mathbb{E}\left[g^{(3)}\left(\langle x,\beta_k\rangle + b_k\right)\right].\beta(j)\beta(l) \quad (25)$$

and

$$\sum_{k=1}^{K} U(b_k)\frac{\partial M_3(\theta)[j,l,m]}{\partial b_k} = \sum_{k=1}^{K} \omega_k U(b_k)\mathbb{E}\left[g^{(4)}\left(\langle x,\beta_k\rangle + b_k\right)\right].\beta_k(j)\beta(l)_k\beta_k(m) \quad (26)$$

respectively. Derivating with respect to the β_k's coordinates and using Stein's identity, the last terms of equations (18) to (20) are rewritten as:

$$\sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_1(\theta)[j]}{\partial \beta_{mk}} = \sum_{k=1}^{K} \omega_k \mathbb{E}\left[g^{(3)}\left(\langle x,\beta_k\rangle + b_k\right)\right]\langle \beta_k, U(b_k)\rangle \beta_k(j) + \sum_{k=1}^{K} \omega_k \mathbb{E}\left[g'\left(\langle x,\beta_k\rangle + b_k\right)\right] U(\beta_k(j)), \quad (27)$$

$$\sum_{k=1}^{K}\sum_{m=1}^{d} U(\beta_{mk})\frac{\partial M_2(\theta)[j,l]}{\partial \beta_{mk}} = \sum_{k=1}^{K} \omega_k \mathbb{E}\left[g^{(4)}\left(\langle x,\beta_k\rangle + b_k\right)\right]\langle \beta_k, U(b_k)\rangle \beta_k(j)\beta_k(l) + \sum_{k=1}^{K} \omega_k \mathbb{E}\left[g''\left(\langle x,\beta_k\rangle + b_k\right)\right]\Big\{\beta_k(j)U(\beta_k(l)) + \beta_k(l)U(\beta_k(j))\Big\}, \quad (28)$$

and

$$\sum_{k=1}^{K}\sum_{s=1}^{d} U(\beta_{sk})\frac{\partial M_3(\theta)[j,l,m]}{\partial \beta_{sk}} = \sum_{k=1}^{K} \omega_k \mathbb{E}\left[g^{(5)}\left(\langle x,\beta_k\rangle + b_k\right)\right]\langle \beta_k, U(b_k)\rangle \beta_k(j)\beta_k(l)\beta_k(s) + \sum_{k=1}^{K} \omega_k \mathbb{E}\left[g^{(3)}\left(\langle x,\beta_k\rangle + b_k\right)\right]\Big\{\beta_k(j)\beta_k(l)U(\beta_k(s)) + \beta_k(j)U(\beta_k(l))\beta_k(s) + U(\beta_k(j))\beta_k(l)\beta_k(s)\Big\} \quad (29)$$

respectively. Then using equations (18) to (29), we can rewrite equation (15) as:

$$\begin{aligned}\overline{0} &= \sum_{k=1}^{K-1} U(\omega_k)\Big\{\mathbb{E}\big[g'\left(\langle x,\beta_k\rangle+b_k\right)\big]\beta_k - \mathbb{E}\big[g'\left(\langle x,\beta_K\rangle+b_K\right)\big]\beta_K\Big\}\\ &+ \sum_{k=1}^{K}\omega_k U(b_k)\mathbb{E}\big[g''\left(\langle x,\beta_k\rangle+b_k\right)\big]\beta_k + \sum_{k=1}^{K}\omega_k\mathbb{E}\big[g^{(3)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\langle U(\beta_k),\beta_k\rangle\beta_k\\ &+ \sum_{k=1}^{K}\omega_k\mathbb{E}\big[g'\left(\langle x,\beta_k\rangle+b_k\right)\big]U(\beta_k), \end{aligned} \tag{30}$$

rewrite equation (16) as:

$$\begin{aligned}\overline{0}\otimes\overline{0} &= \sum_{k=1}^{K-1} U(\omega_k)\Big[\mathbb{E}\big[g''\left(\langle x,\beta_k\rangle+b_k\right)\big]\beta_k\otimes\beta_k - \mathbb{E}\big[g''\left(\langle x,\beta_K\rangle+b_K\right)\big]\beta_K\otimes\beta_K\Big]\\ &+ \sum_{k=1}^{K}\omega_k U(b_k)\mathbb{E}\big[g^{(3)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\beta_k\otimes\beta_k\\ &+ \sum_{k=1}^{K}\omega_k\mathbb{E}\big[g^{(4)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\langle U(\beta_k),\beta_k\rangle\beta_k\otimes\beta_k\\ &+ \sum_{k=1}^{K}\omega_k\mathbb{E}\big[g''\left(\langle x,\beta_k\rangle+b_k\right)\big]\Big(U(\beta_k)\otimes\beta_k+\beta_k\otimes U(\beta_k)\Big), \end{aligned} \tag{31}$$

and rewrite equation (17) as:

$$\begin{aligned}\overline{0}^{\otimes 3} &= \sum_{k=1}^{K-1} U(\omega_k)\Big[\mathbb{E}\big[g^{(3)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\beta_k^{\otimes 3} - \mathbb{E}\big[g^{(3)}\left(\langle x,\beta_K\rangle+b_K\right)\big]\beta_K^{\otimes 3}\Big]\\ &+ \sum_{k=1}^{K}\omega_k U(b_k)\mathbb{E}\big[g^{(4)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\beta_k^{\otimes 3} + \sum_{k=1}^{K}\omega_k\mathbb{E}\big[g^{(5)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\langle U(\beta_k),\beta_k\rangle\beta_k^{\otimes 3}\\ &+ \sum_{k=1}^{K}\omega_k\mathbb{E}\big[g^{(3)}\left(\langle x,\beta_k\rangle+b_k\right)\big]\Big(U(\beta_k)\otimes\beta_k\otimes\beta_k+\beta_k\otimes U(\beta_k)\otimes\beta_k+\beta_k\otimes\beta_k\otimes U(\beta_k)\Big). \end{aligned} \tag{32}$$

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

Let W be any vector that is orthogonal to this linear space. By multiplying (32) on the right by W, and by using the fact that $\beta_1,\ldots,\beta_K$ are linearly independent by (H1), we get that

$$\forall k=1,\ldots,K,\ \omega_k(G_3)_k\langle U(\beta_k),W\rangle = 0.$$

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

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

Then, if (H4) holds, we get that for any k and any W, $\langle U(\beta_k),W\rangle=0$, which proves that the vectors $U(\beta_1),\ldots,U(\beta_K)$ all belong to the linear space spanned by $\beta_1,\ldots,\beta_K$. Let B be the $d\times K$ matrix having $\beta_1,\ldots,\beta_K$ as colomn vectors. Let $U(\beta)$ be the $d\times K$ matrix

having $U(\beta_1), \ldots, U(\beta_K)$ as colomn vectors. We thus have that there exists a $K \times K$ matrix $A = (A_1, \ldots, A_K)$ such that $UU(\beta) = BA$.
Set

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

$$U(b) = (U(b_1), \ldots, U(b_K))$$

and recall that

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

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 $\Delta = Diag(\omega)$. For $W \in \mathbb{R}^d$, set, $D = Diag\left(\langle \beta_1, W \rangle, \ldots, \langle \beta_K, W \rangle\right)$. Then using the fact that B is full rank, (32) gives that

$$\begin{aligned} &G_3PD + G_4\Delta QD + G_5 + AG_3\Delta D + G_3\Delta DA^T + \Delta Diag\left(\langle U(\beta_1), \beta_1 \rangle, \ldots, \langle U(\beta_K), \beta_K \rangle\right), BA_K \rangle) D \\ &+G_3 Diag\left(\langle U(\beta_1), \beta_1 \rangle, \ldots, \langle U(\beta_K), \beta_K \rangle\right) = \overline{0} \otimes \overline{0}. \end{aligned} \tag{33}$$

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

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

So (33) can be rewriten as

$$G_3PD + G_4\Delta QD + G_5\Delta \tilde{D}D + AG_3\Delta D + G_3\Delta DA^T + G_3\Delta \tilde{D} = \overline{0} \otimes \overline{0}, \tag{34}$$

So that for all $W \in \mathbb{R}^d$, $W \in \mathbb{R}^d$, $AG_3\Delta D + G_3\Delta DA^T$ 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,

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

and (34) can be rewriten as

$$G_3PD + G_4\Delta QD + G_5\Delta A\tilde{B}D + AG_3\Delta D + G_3\Delta DA^T + G_3\Delta A\tilde{B} = \overline{0} \otimes \overline{0}. \tag{35}$$

But by taking, for $k = 1, \ldots, K$, W_k such that $\beta_k^T W_k = 0$, we have $D = 0$. In this case, (35) 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 $U(\beta_k) = 0$, $k = 1, 2, \ldots, K$. Then using the fact that B is full rank, we conclude from (30) and (31) that

$$G_1P + G_2\Delta Q = \overline{0} \otimes \overline{0}, \tag{36}$$

and

$$G_2P + G_3\Delta Q = \overline{0} \otimes \overline{0}. \tag{37}$$

Multipliying (36) by G_3 and (37) by G_2, we have

$$G_1G_3P + G_2G_3\Delta Q = \overline{0} \otimes \overline{0}, \tag{38}$$

and

$$G_2^2 P + G_2 G_3 \Delta Q = \overline{0} \otimes \overline{0}. \tag{39}$$

Taking the difference (38)-(39), 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.

References

[1] Bijan Afsari. Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 30(3):11481171, 09 2008.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variables models. *Journal of machine learning*, 15:2773–2832, 08 2014.

[3] Benjamin Auder and Mor Absa Loum. Morpheus: An R package to estimate parameters of logistic regressions mixtures. *CRAN*, June 2018.

[4] Dean A. Follmann and Diane Lambert. Identifiability of finite mixtures of logistic regression models. *J. Statist. Plann. Inference*, 27(3):375–381, 1991.

[5] Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York, 2006.

[6] Bettina Grün and Friedrich Leisch. Flexmix: An R package for finite mixture modelling. *R News*, 7(1):8–13, April 2007.

[7] Bettina Grün and Friedrich Leisch. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *J. Classification*, 25(2):225–247, 2008.

[8] Arnost Komárek and Lenka Komárková. Clustering for multivariate continuous and discrete longitudinal data. *Ann. Appl. Stat.*, 7(1):177–200, 2013.

[9] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1223–1231, Cadiz, Spain, 09–11 May 2016. PMLR.

[10] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, cambridge edition, 1998.

[11] Shaoli Wang, Weixin Yao, and Mian Huang. A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statist. Probab. Lett.*, 93:41–45, 2014.