One-step closed-form estimator for generalized linear model with categorical explanatory variables

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan, Tom Rohmer

One-step closed-form estimator for generalized linear model with categorical explanatory variables

Alexandre Brouste[1], Christophe Dutang[2], Lilit Hovsepyan[1], and Tom Rohmer*[3]

[1]Laboratoire Manceau de Mathématiques, Le Mans Université, F-72000 Le Mans
[2]Université Grenoble Alpes, CNRS, Grenoble INP, LJK, F-38000 Grenoble
[3]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan

October 20, 2023

Abstract. The parameters of generalized linear models are generally estimated by the maximum likelihood estimator (MLE), computed using a Newton-Raphson type algorithm that can be time-consuming for a large number of variables or modalities, or a large sample size. Explicit estimators exist for these models but they are not always asymptotically efficient, especially for simple effects models, although they are fast to calculate compared to the MLE. The article proposes a fast and asymptotically efficient estimation of the parameters of generalized linear models with categorical explanatory variables. It is based on a one-step procedure where a single step of the gradient descent is performed on the log-likelihood function initialized from the explicit estimators. This work presents the theoretical results obtained, the simulations carried out and an application to car insurance pricing.

Keywords. Generalized linear models, explicit estimators, categorical explanatory variables, one-step procedure

1 Introduction

Generalized linear models (GLMs) are regression models where the distribution of the response variable belongs to the exponential family with a natural parameter which is a linear combination of explanatory variables (up to a link function, see e.g. McCullagh & Nelder (1989) for an introduction). The unknown parameters of such models are generally estimated by the maximum likelihood estimator (MLE) for which the asymptotic normality (and efficiency) have been established by Fahrmeir & Kaufmann (1985). Since the MLE has no closed-form formula in general, it is numerically computed by a Newton-type method (known as the Fisher scoring algorithm which can be rewritten as an iteratively re-weighted least square method (IWLS), see McCullagh & Nelder 1989). This computation method is time-consuming for large datasets or for numerous explanatory variables.

The setting of sole categorical explanatory variables is singular due to the non-identifiability of the model and linear identifiability conditions are usually imposed. In this setting, for some

*Corresponding author: tom.rohmer@inrae.fr

specific models, closed-form MLE has been proposed by Brouste, Dutang & Rohmer (2020) which is asymptotically efficient. For other models, as the single effect models, closed-form alternative estimator which is consistent, asymptotically normal but not asymptotically efficient has also been built (see e.g. Brouste et al. 2022).

Dealing with categorical explanatory variables is of particular interest in practical applications. A finite number of risk groups relying on categorical variables is used for the pricing of guarantees (Denuit et al. 2020) or for modeling disease, fertility and milk production in dairy cattle in Kadarmideen et al. (2000). The parameters of GLMs can be fastly estimated with the closed formula and large datasets and/or large number of modalities can be handled.

But in most situations, explanatory variables are used as single effect ((McCullagh & Nelder 1989, Chapters 4 and 6), Lindsey (1997), (Denuit et al. 2020, Chapter 4), (Wuethrich & Merz 2021, Chapter 5)). Consequently, the loss of asymptotical efficiency for the gain of computation speed could be questioned. We propose in this paper a fast and asymptotically efficient procedure to estimate the parameters of GLM with categorical variables based on the one-step procedure.

The one-step procedure was initially considered for the estimation of parameters in independent and identically distributed (i.i.d.) samples (see Le Cam (1956)). In such procedure, an initial guess estimator is proposed which is fast to be computed but not asymptotically efficient. Then, a single step of the gradient descent method is done on the log-likelihood function in order to correct the initial estimation and reach asymptotic efficiency, see Brouste et al. (2021). With some recent developments, the one-step procedure has been successfully generalized to more sophisticated statistical experiments as diffusion processes in Kamatani & Uchida (2015), Gloter & Yoshida (2021), ergodic Markov chains in Kutoyants & Motrunich (2016), inhomogeneous Poisson counting processes in Dabye et al. (2018), fractional Gaussian and stable noises observed at high frequency in Brouste & Masuda (2018), Brouste, Soltane & Votsi (2020).

The paper falls into 5 parts. The notations for GLMs are given in Section 2. The notion of restricted parameter is described in Section 3, which is new from our previous studies in Brouste, Dutang & Rohmer (2020), Brouste et al. (2022), with the asymptotic properties of the MLE and the closed-form estimator. The fast one-step estimation procedure is also presented in Section 3 with a convergence result. Monte-Carlo simulations on samples of finite size are done in Section 4 and an application to car insurance is given in Section 5. Technical lemmas and the proof of the main are postponed in Appendix.

2 Preliminaries on GLMs

2.1 Notation for GLMs

Bold notations are given for the vectors. The index $i \in I = \{1, \dots, n\}$ is reserved for the observations, while the indexes j, k are used for the explanatory variables.

The observation sample $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ is composed of independent random variables valued in $\mathbb{Y} \subset \mathbb{R}$ where for $i \in I$, Y_i belongs to the one-parameter exponential family of probability measures valued in $\Lambda \subset \mathbb{R}$. In this setting, the log-likelihood $\log \mathcal{L}$ of the sample is

$$\log \mathcal{L}(\boldsymbol{\vartheta}, \phi \,|\, \boldsymbol{Y}) = \sum_{i=1}^{n} \frac{\lambda_i(\boldsymbol{\vartheta}) Y_i - b\left(\lambda_i(\boldsymbol{\vartheta})\right)}{a(\phi)} + \sum_{i=1}^{n} c(Y_i, \phi), \tag{1}$$

where $a : \mathbb{R} \to \mathbb{R}$, $b : \Lambda \to \mathbb{R}$ and $c : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ are fixed real-valued measurable functions and ϕ is the dispersion parameter, e.g. McCullagh & Nelder (1989, Section 2.2).

The parameters $\lambda_1, \dots, \lambda_n$ in Equation (1) depend on a p-dimensional parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$. For $i = 1, \dots, n$, denote $\mu_i = b'(\lambda_i(\boldsymbol{\vartheta}))$. Theoretical moments of Y_i are explicitly given as a function of a and derivatives of b

$$\mathbf{E}_{\boldsymbol{\vartheta}} Y_i = b'(\lambda_i(\boldsymbol{\vartheta})) = \mu_i \quad \text{and} \quad \mathbf{Var}_{\boldsymbol{\vartheta}} Y_i = b''(\lambda_i(\boldsymbol{\vartheta}))a(\phi) = V(\mu_i)a(\phi). \tag{2}$$

The function $V : \mu \mapsto V(\mu) = b'' \circ (b')^{-1}(\mu)$ is known as the variance function of the expectation μ. Using a twice continuously differentiable and bijective function g from $b'(\Lambda)$ to $\mathbb{R}$, GLMs are defined by assuming the following relation between the expectation $\mathbf{E}_{\boldsymbol{\vartheta}} Y_i$ and the predictor

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\vartheta} = \eta_i, \quad \text{for all } \boldsymbol{\vartheta} \in \Theta,$$

where η_i are the linear predictors. The parameter $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$ and the parameter $\phi > 0$ are unknown and are to be estimated. The function g is called the link function in the regression framework. In other words, the bijective function $\ell = (b')^{-1} \circ g^{-1}$ is settled; then we have $\lambda_i(\boldsymbol{\vartheta}) = \ell(\eta_i)$. We summarize with the following relations

$$X \times \Theta \longrightarrow D \underset{\ell}{\overset{\ell^{-1}}{\leftrightharpoons}} \Lambda,$$

where D is the space of linear predictor and X the possible set of value of $\boldsymbol{x}_i$ for $i \in I$. Here ℓ is chosen and, consecutively Θ, Λ and D must be set. We talk of canonical link function, when ℓ is the identity function, that is to say $g = (b')^{-1}$.

2.2 Score and Fisher information for GLMs

We introduce T as the part of the log-likelihood depending on $\boldsymbol{\vartheta}$ only in order to express the first-order condition verified by the maximum likelihood estimator (MLE)

$$T(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y}) = \sum_{i=1}^{n} Y_i \ell(\eta_i) - b\left(\ell(\eta_i)\right).$$

Hence, the log-likelihood (1) can be rewritten as

$$\log \mathcal{L}(\boldsymbol{\vartheta}, \phi \,|\, \boldsymbol{Y}) = \frac{T(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y})}{a(\phi)} + \sum_{i=1}^{n} c(Y_i, \phi).$$

The gradient of $\log \mathcal{L}$ is

$$\nabla_{(\boldsymbol{\vartheta},\phi)} \log \mathcal{L}(\boldsymbol{\vartheta}, \phi \,|\, \boldsymbol{Y}) = \begin{pmatrix} \nabla_{\boldsymbol{\vartheta}} T(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y})/a(\phi) \\ -\frac{a'(\phi)}{a(\phi)^2} T(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y}) + \sum_{i=1}^{n} \frac{\partial}{\partial \phi} c(Y_i, \phi) \end{pmatrix} =: \begin{pmatrix} U(\boldsymbol{\vartheta})/a(\phi) \\ \mathcal{V}(\boldsymbol{\vartheta}, \phi) \end{pmatrix}.$$

The MLE $(\widehat{\boldsymbol{\vartheta}}_n, \widehat{\phi}_n)$ for $(\boldsymbol{\vartheta}, \phi)$ satisfies

$$U(\widehat{\boldsymbol{\vartheta}}_n) = 0 \quad \text{and} \quad \mathcal{V}(\widehat{\boldsymbol{\vartheta}}_n, \widehat{\phi}_n) = 0. \tag{3}$$

Using (2), we define the weight matrix and the explanatory variables matrix by

$$W(\boldsymbol{\vartheta}) = \begin{pmatrix} \frac{1}{(g'(\mu_1))^2 V(\mu_1)} & & \\ & \ddots & \\ & & \frac{1}{(g'(\mu_n))^2 V(\mu_n)} \end{pmatrix}, \quad X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(p)} \\ & \vdots & \\ x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix}, \tag{4}$$

and Z the vector of mean deviations

$$Z(\boldsymbol{\vartheta}) = \begin{pmatrix} (y_1 - \mu_1)g'(\mu_1) \\ \vdots \\ (y_n - \mu_n)g'(\mu_n) \end{pmatrix}. \tag{5}$$

Note that all diagonal terms W are positive so that $W(\boldsymbol{\vartheta})$ is invertible. U can be rewritten as a matrix multiplication $U(\boldsymbol{\vartheta}) = X^T W(\boldsymbol{\vartheta}) Z(\boldsymbol{\vartheta})$, cf. (25) in Appendix A.2, where for a matrix M, M^T denotes the transposition of the matrix M. We have the following expression for the score vector and the Fisher information

$$\mathcal{S}_n(\boldsymbol{\vartheta}) = \frac{X^T W(\boldsymbol{\vartheta}) Z(\boldsymbol{\vartheta})}{a(\phi)}, \quad \mathcal{I}_n(\boldsymbol{\vartheta}) = \frac{X^T W(\boldsymbol{\vartheta}) X}{a(\phi)}.$$

Since Fahrmeir & Kaufmann (1985), under regularity conditions, the MLE $\widehat{\boldsymbol{\vartheta}}_n$ of $\boldsymbol{\vartheta}$ asymptotically exists and as soon as the MLE is unique, that is to say there is no over-parametrization in the model, we have

$$\mathcal{I}_n^{T/2}(\boldsymbol{\vartheta})(\widehat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_p\left(\mathbf{0}_p, I_p\right),$$

with $\mathcal{I}_n^{1/2}\mathcal{I}_n^{T/2} = \mathcal{I}_n$ where I_p is the identity matrix of $\mathbb{R}^{p\times p}$.

3 Estimation of GLMs with categorical variables

Consider the case where all m explanatory variables are categorical, that is for $j = 1, \ldots, m$ every observations $(x_i^{(j+1)})_i$ take values in a finite set $\{v_{j,1}, \ldots, v_{j,d_j}\}$ and $x_i^{(1)} = 1$ is the intercept. Assuming values are unordered, $x_i^{(j+1)}$ needs to be encoded using binary dummies as

$$x_i^{(j+1),k} = 1_{\{x_i^{(j+1)} = v_{j,k}\}}, \quad k \in \{1, \ldots, d_j\}.$$

These binary dummies can be used both in single-effect models or with cross-effect models.

In the following, we introduce the notion of restricted parameter which is new from our previous studies Brouste, Dutang & Rohmer (2020), Brouste et al. (2022) and recall the asymptotic properties of the MLE of this parameter in Section 3.1. The properties of the closed-form estimator (CFE) are given in Section 3.2.

It is worth mentioning that, on the one hand, the MLE will be shown to be asymptotically efficient but time-consuming in the general setting. On the other hand, the CFE is fast to be computed but is not asymptotically efficient.

3.1 Notion of restricted parameter

For the sake of presentation, we firstly consider the single-effect model and secondly present the general case.

3.1.1 Model with single effects only

First let consider the GLMs with single effect only

$$g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) = \vartheta_1 + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)}. \tag{6}$$

For $j = 2, \ldots, m+1$, denote $\boldsymbol{\vartheta}^{(j)} = (\vartheta_k^{(j)})_{k=1,\ldots,d_j}$. Hence the parameter vector is written as $\boldsymbol{\vartheta} = (\vartheta_1, \boldsymbol{\vartheta}^{(2)}, \ldots, \boldsymbol{\vartheta}^{(m+1)})^T$. The linear predictor vector $\boldsymbol{\eta} = (\eta_i)_{i=1\ldots,n}$ can be rewritten as $\boldsymbol{\eta} = X\boldsymbol{\vartheta}$, with

$$X = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{x}^{(2)} & \ldots & \boldsymbol{x}^{(m+1)} \end{pmatrix}, \ \boldsymbol{x}^{(j)} = \begin{pmatrix} x_1^{(j),1} & \ldots & x_1^{(j),d_j} \\ \vdots & \vdots & \\ x_n^{(j),1} & \ldots & x_n^{(j),d_j} \end{pmatrix}.$$

Because the redundancies of the matrix X going from $\sum_{k=1}^{d_j} x_i^{(j),k} = 1$ for all i, j, at least m linear conditions are needed to be imposed to get the identifiability of the model. Let contrast vectors R_j such that

$$R_j^T \boldsymbol{\vartheta}^{(j)} = 0, \quad j = 1, \ldots, m.$$

This is equivalent to $R\boldsymbol{\vartheta} = 0$ with

$$R = \begin{pmatrix} 0 & R_1^T & \ldots & \mathbf{0} \\ \vdots & & \ddots & \\ 0 & \mathbf{0} & \ldots & R_m^T \end{pmatrix}.$$

Note that the matrix R is a particular case of the constrast matrix considered in Brouste et al. (2022). In other words, the identifiability conditions assume that one component of $\boldsymbol{\vartheta}^{(j)}$ can be rewritten as a linear combination of the others. Therefore, we define a restricted parameter $\tilde{\boldsymbol{\vartheta}}^{(j+1)}$ of size $d_j - 1$ and a matrix B_j of size $d_j \times (d_j - 1)$ such that

$$\boldsymbol{\vartheta}^{(j+1)} = B_j \tilde{\boldsymbol{\vartheta}}^{(j+1)}. \tag{7}$$

In Table 1, two examples of such vectors R_j are given as well as the corresponding restricted parameter and the corresponding coding matrix B_j. Note that for a given R_j, the choice of the restricted parameter is not necessarily unique. We refer to Venables (2023) for further examples of coding matrices B_j and associated contrasts matrices in the R statistical software (R Core Team 2023).

Name	R_j	B_j	implication	$\tilde{\boldsymbol{\vartheta}}^{(j+1)}$	R code
zero-sum	$(1, \ldots, 1)$	$\begin{pmatrix} I_{d_j-1} \\ -\mathbf{1}_{d_j-1}^T \end{pmatrix}$	$\vartheta_{d_j}^{(j+1)} = -\sum_{k=1}^{d_j-1} \vartheta_k^{(j+1)}$	$\begin{pmatrix} \vartheta_1^{(j+1)} \\ \vdots \\ \vartheta_{d_j-1}^{(j+1)} \end{pmatrix}$	`contr.sum`
ref. category	$(1, 0, \ldots, 0)$	$\begin{pmatrix} \mathbf{0}_{d_j-1}^T \\ I_{d_j-1} \end{pmatrix}$	$\vartheta_1^{(j+1)} = 0$	$\begin{pmatrix} \vartheta_2^{(j+1)} \\ \vdots \\ \vartheta_{d_j}^{(j+1)} \end{pmatrix}$	`contr.treatment`

Table 1: Contrast examples and restricted parameters

Using the restricted parameters and the matrices $B_1, \ldots, B_m$, the linear predictor can be rewritten as

$$\boldsymbol{\eta} = \tilde{X}\tilde{\boldsymbol{\vartheta}}, \tag{8}$$

with

$$\tilde{X} = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{x}^{(2)}B_1 & \ldots & \boldsymbol{x}^{(m+1)}B_m \end{pmatrix}, \quad \tilde{\boldsymbol{\vartheta}} = (\vartheta_1, \tilde{\boldsymbol{\vartheta}}^{(2)}, \ldots, \tilde{\boldsymbol{\vartheta}}^{(m+1)}).$$

Using Appendix A.3, the score vector and the Fisher information write in an analogous way as the previous section

$$\tilde{\mathcal{S}}(\boldsymbol{\vartheta}) = \frac{\tilde{X}^T W(\boldsymbol{\vartheta}) Z(\boldsymbol{\vartheta})}{a(\phi)}, \quad \tilde{\mathcal{I}}_n(\boldsymbol{\vartheta}) = \frac{\tilde{X}^T W(\boldsymbol{\vartheta}) \tilde{X}}{a(\phi)}. \tag{9}$$

Note that the matrix W can be big for large datasets, and the Information matrix can be time-consuming. To reduce the computing time, note that, it can be rewritten as $\tilde{\mathcal{I}}_n(\boldsymbol{\vartheta}) = (\tilde{X} \odot S)^T \tilde{X}/a(\phi)$, where $\odot$ is Hadamard's product and S is the vector constituted with the diagonal elements of W.

3.1.2 General case

To take all possible GLM settings into account, we consider a GLM with predictor defined as

$$\begin{aligned}
g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) =& \vartheta_1 + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)} && \text{Intercept and single effect} \\
&+ \sum_{j_2<j_3} \sum_{k_2,k_3} x_i^{(j_2),k_2} x_i^{(j_3),k_3} \vartheta_{k_2,k_3}^{(j_2,j_3)} && \text{Double effect} \\
&+ \sum_{j_2<j_3<j_4} \sum_{k_2,k_3,k_4} x_i^{(j_2),k_2} x_i^{(j_3),k_3} x_i^{(j_4),k_4} \vartheta_{k_2,k_3,k_4}^{(j_2,j_3,j_4)} && \text{Triple effect} \\
&+ \dots && \dots \\
&+ \sum_{k_2,\dots,k_{m+1}} x_i^{(2),k_2} \dots x_i^{(m+1),k_{m+1}} \vartheta_{k_2,\dots,k_{m+1}}^{(2,\dots,m+1)}, && \text{All crossed effect}
\end{aligned} \tag{10}$$

where g is the link function and indexes j_i are in $\{2, \dots, m+1\}$ and k_j are in $\{1, \dots, d_j\}$ for $j = 2, \dots, m+1$. The linear predictor $\boldsymbol{\eta} = (\eta_i)_{i=1\dots,n}$ can be rewritten as $\boldsymbol{\eta} = X\boldsymbol{\vartheta}$, with

$$X = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{x}^{(2)} & \dots & \boldsymbol{x}^{(m+1)} & \boldsymbol{x}^{(2,3)} & \dots & \boldsymbol{x}^{(m,m+1)} & \boldsymbol{x}^{(2,3,4)} \dots & \boldsymbol{x}^{(m-1,m,m+1)} & \dots \boldsymbol{x}^{(2,\dots,m+1)} \end{pmatrix},$$

and with for $2 \leq j_2 < j_3 < \dots$,

$$\boldsymbol{x}^{(j_2,j_3,\dots)} = \begin{pmatrix} x_1^{(j_2,j_3,\dots),1} & \dots & x_1^{(j_2,j_3,\dots),d_{j_2} d_{j_2} \dots} \\ \vdots & & \vdots \\ x_n^{(j_2,j_3,\dots),1} & \dots & x_n^{(j_2,j_3,\dots),d_{j_2} d_{j_2} \dots} \end{pmatrix}, \quad x_i^{(j_2,j_3,\dots),k_1 k_2,\dots} = x_i^{(j_2),k_1} x_i^{(j_3),k_2} \dots$$

The unknown parameter vector is

$$\boldsymbol{\vartheta} = \left(\vartheta_1, (\boldsymbol{\vartheta}^{(j)})_j, (\boldsymbol{\vartheta}^{(j_2,j_3)})_{j_2<j_3}, (\boldsymbol{\vartheta}^{(j_2,j_3,j_4)})_{j_2<j_3<j_4}, \dots, (\boldsymbol{\vartheta}^{(2,\dots,m+1)})\right)^T,$$

with $\boldsymbol{\vartheta}^{(j)} = (\vartheta_k^{(j)})_k$, $\boldsymbol{\vartheta}^{(j_2,j_3)} = (\vartheta_{k_2,k_3}^{(j_2,j_3)})_{k_2,k_3}, \dots$, $\boldsymbol{\vartheta}^{(2,\dots,m+1)} = (\vartheta_{k_2,\dots,k_{m+1}}^{(2,\dots,m+1)})_{k_2,\dots,k_{m+1}}$). The total number of parameters of the model in (10) is

$$p = 1 + \sum_{j=2}^{m+1} d_j + \sum_{j_2<j_3} d_{j_2} d_{j_3} + \dots + \prod_{j=2}^{m+1} d_j.$$

Because the model in (10) is not identifiable, we need to impose at least $p - \prod_{j=2}^{m+1} d_j$ linear conditions. Here, we propose to reduce the dimension of the parameters.

name	parameter	size of the restricted param.	contrast number
Intercept	ϑ_1	1	0
single effect	$\boldsymbol{\vartheta}^{(j)}$	$d_j - 1$	$d_j - (d_j - 1)$
double effect	$\boldsymbol{\vartheta}^{(j_2,j_3)}$	$(d_{j_3} - 1)(d_{j_2} - 1)$	$d_{j_3}d_{j_2} - (d_{j_3} - 1)(d_{j_2} - 1)$
triple effect	$\boldsymbol{\vartheta}^{(j_2,j_3,j_4)}$	$(d_{j_4} - 1)(d_{j_3} - 1)(d_{j_2} - 1)$	$d_{j_4}d_{j_3}d_{j_2} - (d_{j_4} - 1)(d_{j_3} - 1)(d_{j_2} - 1)$
$\vdots$		$\vdots$	$\vdots$
all crossed-effect	$\boldsymbol{\vartheta}^{(2,\ldots,m+1)}$	$\prod_{j=2}^{m+1}(d_j - 1)$	$\prod_{j=2}^{m+1} d_j - \prod_{j=2}^{m+1}(d_j - 1)$

Table 2: size of the restricted parameters and contrast numbers

In Table 2, the number of considered linear contrasts and the size of the restricted parameter $\tilde{\boldsymbol{\vartheta}}^{(j)}$ are given for each elements of the parameter vector. Note that this decomposition is not unique. For example in the single-effect part, one could impose a null intercept or a zero-condition on one of the parameter.

Let $R_j, R_{j_2,j_3}, \ldots, R_{j_2,j_3,\ldots,j_{m+1}}$, $2 \leq j_2 < j_3 < \ldots < j_{m+1} \leq m+1$, matrices such that

$$R_j^T \boldsymbol{\vartheta}^{(j)} = 0, \quad R_{j_2,j_3}^T \boldsymbol{\vartheta}^{(j_2,j_3)}, \quad \ldots, \quad R_{2,\ldots,m+1}^T \boldsymbol{\vartheta}^{(2,\ldots,m+1)} = 0.$$

That is to say, $R\boldsymbol{\vartheta}_{-1} = 0$, with R the diagonal block matrix

$$R = \operatorname{diag}\left((R_j)_j, (R_{j_2,j_3})_{j_2,j_3}, \ldots, R_{2,\ldots,m+1}\right),$$

and $\boldsymbol{\vartheta}_{-1}$ the parameter vector without the intercept ϑ_1.

For all m-effect parts of the model, we can define a restricted parameter $\tilde{\boldsymbol{\vartheta}}^{(j)}$, $\tilde{\boldsymbol{\vartheta}}^{(j_2,j_3)}$, …, $\tilde{\boldsymbol{\vartheta}}^{(2,\ldots,m+1)}$ of respective size $\tilde{d}_j = d_j - 1$, $\tilde{d}_{j_2,j_3} = (d_{j_3} - 1)(d_{j_2} - 1)$, …, $\tilde{d}_{2,\ldots,m+1} = \prod_{j=2}^{m+1}(d_j - 1)$ and the corresponding coding matrix B_j, B_{j_2,j_3}, …, $B_{2,\ldots,m+1}$ respectively of size $d_j \times \tilde{d}_j$, $d_{j_2}d_{j_3} \times \tilde{d}_{j_2,j_3}$, …, $\prod_{j=2}^{m+1} d_j \times \tilde{d}_{2,\ldots,m+1}$ such that

$$\boldsymbol{\vartheta}^{(j)} = B_j \tilde{\boldsymbol{\vartheta}}^{(j)}, \quad \boldsymbol{\vartheta}^{(j_2,j_3)} = B_{j_2 j_3} \tilde{\boldsymbol{\vartheta}}^{(j_2,j_3)}, \quad \ldots, \quad \boldsymbol{\vartheta}^{(2,\ldots,m+1)} = B_{2,\ldots,m+1} \tilde{\boldsymbol{\vartheta}}^{(2,\ldots,m+1)}. \tag{11}$$

The coding matrices can be rewritten in terms of the Kronecker product. Consider B_2, B_3 and B_4 of respective dimension $d_2 \times (d_2 - 1)$, $d_3 \times (d_3 - 1)$ and $d_4 \times (d_4 - 1)$. Then the Kronecker product $B_{2,3} = B_3 \otimes B_2$ has dimension $d_3 d_2 \times (d_3 - 1)(d_2 - 1)$. In the same way, $B_{2,3,4} = B_4 \otimes B_{2,3}$ has dimension $d_4 d_3 d_2 \times (d_4 - 1)(d_3 - 1)(d_2 - 1)$. More generally, for B_j of dimension $d_j \times (d_j - 1)$, $\bigotimes_{j=m+1}^{2} B_j$ has dimension $\prod_{j=m+1}^{2} d_j \times \prod_{j=m+1}^{2}(d_j - 1)$.

Using the restricted parameters and the coding matrices B_j, $B_{j_2 j_3}, \ldots, B_{2,\ldots,m}$ the linear predictor can be rewritten as

$$\boldsymbol{\eta} = \tilde{X}\tilde{\boldsymbol{\vartheta}}, \tag{12}$$

with the new regressor matrix

$$\tilde{X} = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{x}^{(2)}B_1 & \ldots & \boldsymbol{x}^{(m+1)}B_m & \boldsymbol{x}^{(2,3)}B_{2,3} & \ldots & \boldsymbol{x}^{(m,m+1)}B_{m,m+1} & \ldots & \boldsymbol{x}^{(2,\ldots,m+1)}B_{2,\ldots,m+1} \end{pmatrix}$$

Note that the linear predictor η_i in (12) is identical whether we use $\boldsymbol{\vartheta}$ or $\tilde{\boldsymbol{\vartheta}}$, leading to the same expectation μ_i, henceforth the same matrices $W(.)$ and $Z(.)$.

3.2 Asymptotic properties of the MLE and the closed-form estimator

Since the covariates are supposed to be categorical in this paper, the vector of linear predictors $\boldsymbol{\eta}$ defined in (12) takes $d = \prod_j d_j$ distinct values namely $h_1, \ldots, h_d$. We consider the unique $d \times p$ matrix Q composed of binary dummies such that

$$\boldsymbol{\eta}^\star = Q\boldsymbol{\vartheta}, \quad \boldsymbol{\eta}^\star = (h_j)_{j=1,\ldots,d}.$$

For example, in the single-effect model (6), $Q = (M_m^{()}, M_m^{(2)}, \ldots, M_m^{(m+1)})$ with $M_m^{()} = 1_{d_{m+1}\ldots d_2}$, and for $2 < j < m+1$

$$M_m^{(2)} = \mathbf{1}_{d_{m+1}\ldots d_3} \otimes I_{d_2}, \quad M_m^{(m+1)} = I_{d_{m+1}} \otimes \mathbf{1}_{d_m \ldots d_2}, \quad M_m^{(j)} = \mathbf{1}_{d_{m+1}} \otimes I_{d_j} \otimes \mathbf{1}_{d_{j-1}\ldots d_2}.$$

The general form of Q for the model (10) is given in Brouste et al. (2022).

Using the same restricted parameter as (11), and considering vectors R_j of size d_{j+1} such that $R_j\boldsymbol{\vartheta}^{(j+1)} = 0$, the restricted linear predictors $\boldsymbol{\eta}^\star$ can be rewritten as previously as

$$\boldsymbol{\eta}^\star = \tilde{Q}\tilde{\boldsymbol{\vartheta}}.$$

For example in the single-effect model

$$\tilde{Q} = \begin{pmatrix} M_m^{()} & M_m^{(2)}B_1 & \ldots & M_m^{(m+1)}B_m \end{pmatrix}, \quad \tilde{\boldsymbol{\vartheta}} = (\vartheta_1, \tilde{\boldsymbol{\vartheta}}^{(2)}, \ldots, \tilde{\boldsymbol{\vartheta}}^{(m+1)}).$$

Note the the score $\tilde{\mathcal{S}}$ and the information matrix $\tilde{\mathcal{I}}_n(\boldsymbol{\vartheta})$ can be rewritten in term of the $\tilde{Q}$ matrix as following

$$a(\phi)\tilde{\mathcal{S}}(\boldsymbol{\vartheta}) = n\tilde{Q}^T\Sigma_n(\boldsymbol{\vartheta})Z^\star(\boldsymbol{\vartheta}), \quad a(\phi)\tilde{\mathcal{I}}_n(\boldsymbol{\vartheta}) = n\tilde{Q}^T\Sigma_n(\boldsymbol{\vartheta})\tilde{Q},$$

with Σ_n is the diagonal matrix whose the diagonal elements are

$$\Sigma_{n,j,j}(\boldsymbol{\vartheta}) = \frac{m_j}{n}\left((g'(\mu_j^\star))^2 V(\mu_j^\star)\right)^{-1}, \quad \mu_j^\star = g^{-1}(h_j), j = 1, \ldots, d, \tag{13}$$

and

$$Z^\star(\boldsymbol{\vartheta}) = ((\bar{y}_n^{(j)} - \mu_j^\star)g'(\mu_j^\star))_{j=1,\ldots,d}, \tag{14}$$

see Appendix A.3. Hence, when p_j satisfies $\frac{m_j}{n} \to p_j$ as $n \to \infty$, the asymptotic distribution of the MLE converges in distribution

$$\sqrt{n}(\widehat{\tilde{\boldsymbol{\vartheta}}}_n - \tilde{\boldsymbol{\vartheta}}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, \tilde{\mathcal{I}}^{-1}(\boldsymbol{\vartheta})\right), \tag{15}$$

with $p^\star$ the dimension of the restricted parameter and the information matrix expressed as

$$\tilde{\mathcal{I}}(\boldsymbol{\vartheta}) = \frac{\tilde{Q}\Sigma\tilde{Q}^T}{a(\phi)}, \quad \Sigma(\boldsymbol{\vartheta}) = \mathrm{diag}(v_1, \ldots, v_d), \tag{16}$$

with diagonal elements are asymptotic variances $v_j = p_j((g'(\mu_j^\star))^2 V(\mu_j^\star))^{-1}$.

Because the MLE can be time-consuming for large datasets or for a large number of explanatory variables or modalities for each variables, as soon as $Q^TQ + R^TR$ is definite positive, an alternative closed-form estimator (CFE) which avoids the using of a time-consuming IWLS algorithm has been defined in Brouste et al. (2022).

The closed-form estimator of the restricted parameter is

$$\tilde{\boldsymbol{\vartheta}}_n^{CFE} = (\tilde{Q}^T\tilde{Q})^{-1}\tilde{Q}^T g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}),$$

where

$$g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}) = \begin{pmatrix} g(\overline{Y}_n^1) \\ \vdots \\ g(\overline{Y}_n^k) \\ \vdots \\ g(\overline{Y}_n^d) \end{pmatrix}, \quad \overline{Y}_n^k = \frac{\sum_{i=1;\eta_i=h_k}^{n} Y_i}{m_k}, m_k = \#\{i \in \{1,\dots,n\}; \eta_i = h_k\}. \tag{17}$$

From Theorem 1 in Brouste et al. (2022), we know the asymptotic distribution of $\tilde{\boldsymbol{\vartheta}}_n^{CFE}$:

$$\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \tilde{\boldsymbol{\vartheta}}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, a(\phi)(\tilde{Q}^T\tilde{Q})^{-1}\tilde{Q}^T\Sigma^{-1}(\tilde{\boldsymbol{\vartheta}})\tilde{Q}(\tilde{Q}^T\tilde{Q})^{-1}\right), \tag{18}$$

with $\Sigma(\boldsymbol{\vartheta})^{-1}$ the diagonal matrix whose diagonal elements are $1/v_j$ and p_j satisfies $\frac{m_j}{n} \to p_j$ as $n \to \infty$.

As mentioned in Brouste et al. (2022), as soon as the closed-form estimator coincides with the MLE (in this case the $\tilde{Q}$ is a square matrix), $\tilde{\boldsymbol{\vartheta}}_n^{CFE} = \tilde{Q}^{-1}g(\bar{Y}_n)$, and consequently using the delta method we get

$$\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \tilde{\boldsymbol{\vartheta}}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, a(\phi)\tilde{Q}^{-1}\Sigma^{-1}(\boldsymbol{\vartheta})(\tilde{Q}^{-1})^T\right). \tag{19}$$

That is the asymptotic variance of $\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \tilde{\boldsymbol{\vartheta}})$ coincides with $\tilde{\mathcal{I}}^{-1}(\boldsymbol{\vartheta})$ defined in (15).

In general, for instance in single effect models, the CFE is not the MLE and is not asymptotically efficient. For this reason, it is desirable to add a single step of the gradient descent method in order to reach asymptotic efficiency. Hence we consider a one-step version of $\tilde{\boldsymbol{\vartheta}}_n^{CFE}$ in the next subsection.

3.3 One-Step Closed-form Estimator

The One-Step Closed-form Estimator (OS-CFE) of $\tilde{\boldsymbol{\vartheta}}$ is defined as

$$\begin{aligned}\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} &= \tilde{\boldsymbol{\vartheta}}_n^{CFE} + \tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})^{-1}\tilde{\mathcal{S}}(\tilde{\boldsymbol{\vartheta}}_n^{CFE}) \\ &= (\tilde{Q}'\tilde{Q})^{-1}\tilde{Q}^T g(\overline{\boldsymbol{Y}}_{\boldsymbol{n}}) + (\tilde{Q}^T\Sigma_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})\tilde{Q})^{-1}\tilde{Q}^T\Sigma_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})Z^\star(\tilde{\boldsymbol{\vartheta}}_n^{CFE}),\end{aligned}$$

where $Z^\star$ is defined in (14). It is worth emphasizing that the OS-CFE of $\tilde{\boldsymbol{\vartheta}}$ does not depend on the dispersion parameter ϕ by simplification. The main result is that the OS-CFE of the restricted parameter $\tilde{\boldsymbol{\vartheta}}$ is asymptotically equivalent in probability to the MLE. The proof of Theorem 1 is postponed in Appendix B.

Theorem 1. *Under some regular conditions, as soon as for all $j = 1,\dots,d$ the frequencies $\frac{m_j}{n} \to p_j$ as $n \to \infty$,*

$$\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} - \hat{\tilde{\boldsymbol{\vartheta}}}_n) \xrightarrow[n\to+\infty]{P} 0.$$

It also means that the OS-CFE is asymptotically normal with an optimal asymptotic variance, i.e.

$$\sqrt{n}(\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} - \tilde{\boldsymbol{\vartheta}}) \xrightarrow[n\to+\infty]{L} \mathcal{N}_{p^\star}\left(\mathbf{0}_{p^\star}, \tilde{\mathcal{I}}^{-1}(\boldsymbol{\vartheta})\right) \tag{20}$$

where $\tilde{\mathcal{I}}(\boldsymbol{\vartheta})$ is defined in (16).

4 Monte Carlo illustrations

The performances on finite size samples of the aforementioned estimators (MLE, CFE, OS-CFE), in terms of computation times and asymptotic variances, are assessed with numerical examples. Monte Carlo simulations of samples for Poisson and Gamma GLMs are done. All computations are carried out with the R statistical software (R Core Team 2023).

More precisely, for the Poisson GLMs, the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is composed of independent Poisson-distributed random variables with respective distributions

$$f(y_i, \mu_i) = \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i y_i), \quad \mu_i > 0, \quad y_i \in \mathbb{N}, \quad i = 1, \ldots, n.$$

In other words, Y_i is characterized in Equation (1) by

$$\lambda_i = \mu_i, \quad a(\phi) = 1, \quad b(\lambda) = \exp(\lambda) \quad \text{and} \quad c(y_i, \phi) = -\log(y_i!).$$

For the Gamma GLMs, the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is composed of independent Gamma-distributed random variables with respective distributions

$$f(y_i, (\alpha, \beta_i)) = \frac{y_i^{\alpha-1}}{\Gamma(\alpha)} \beta_i^{\alpha} \exp(-\beta_i y_i), \quad \alpha > 0, \quad \beta_i > 0, \quad y_i > 0, \quad i = 1, \ldots, n.$$

In other words, Y_i is characterized in Equation (1) by

$$\lambda_i = -\frac{\beta_i}{\alpha},\ a(\phi) = \phi = \frac{1}{\alpha},\ b(\lambda) = -\log(-\lambda) \text{ and } c(y_i, \phi) = \left(\frac{1}{\phi} - 1\right) \log(y_i) - \log \Gamma\left(\frac{1}{\phi}\right) + \frac{1}{\phi} \log \frac{1}{\phi}.$$

In this numerical example the canonical setting is used (ℓ is the identity function) leading to a log link function for the Poisson distribution ($g(x) = \log(x)$) and the inverse link function for the Gamma distribution ($g(x) = 1/x$).

4.1 Simulations with a fixed sample size and a modality number

In our simulations, we consider two explanatory variables $x_i^{(2)}$, $x_i^{(3)}$ and d_2, d_3 modalities. As a starting point d_2 and d_3 have respectively been taken equal to 2 and 3, and the true parameters are chosen arbitrarily for each case as represented in Table 3. For the Gamma distribution, the dispersion parameter ϕ is also chosen arbitrary: $\phi = 8$. We consider $B = 10^4$ Monte Carlo simulations of samples with the size of $n = 10^4$.

Distribution	Link function	intercept	Variable 2		Variable 3		
	$g(x)$	ϑ_1	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta_1^{(3)}$	$\vartheta_2^{(3)}$	$\vartheta_3^{(3)}$
Gamma	$1/x$	10	1	-1	2	3	-5
Poisson	$\log(x)$	0.05	1	-1	0.5	0.5	-1

Table 3: True parameter values

Further, for the sake of ease, histograms for Poisson and Gamma GLMs are drawn for only ϑ_1 and $\vartheta_2^{(2)}$ parameters. The sequence of OS-CFE naturaly overcomes the performance of CFE in terms of asymptotic variance (see Figures 1 and 2). According to the other comparison in terms of computation time which is highlighted in Table 4, OS-CFE is almost 50 times faster than MLE to be computed for the dataset with the size of $n = 10^4$.

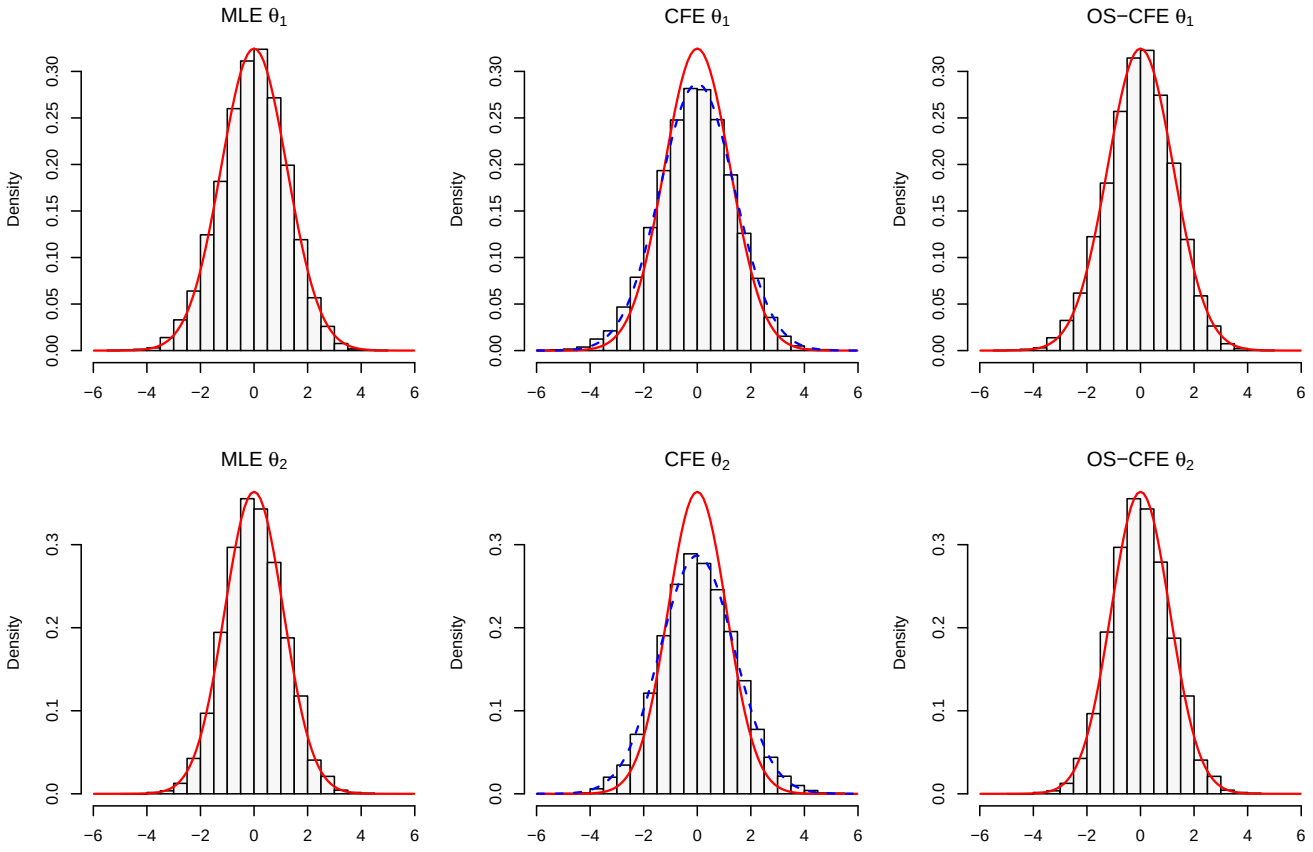

Figure 1: Histograms for the $B = 10^4$ simulations of the renormalized statistical errors of MLE, CFE, OS-CFE for the Poisson distribution with 2 categorical variables with $d_2 = 2$, $d_3 = 3$ for $\theta_1 = \vartheta_1$ and $\theta_2 = \vartheta_2^{(2)}$. Red and blue lines are the theoretical Gaussian asymptotic densities respectively of the MLE (in red) and CFE (in blue).

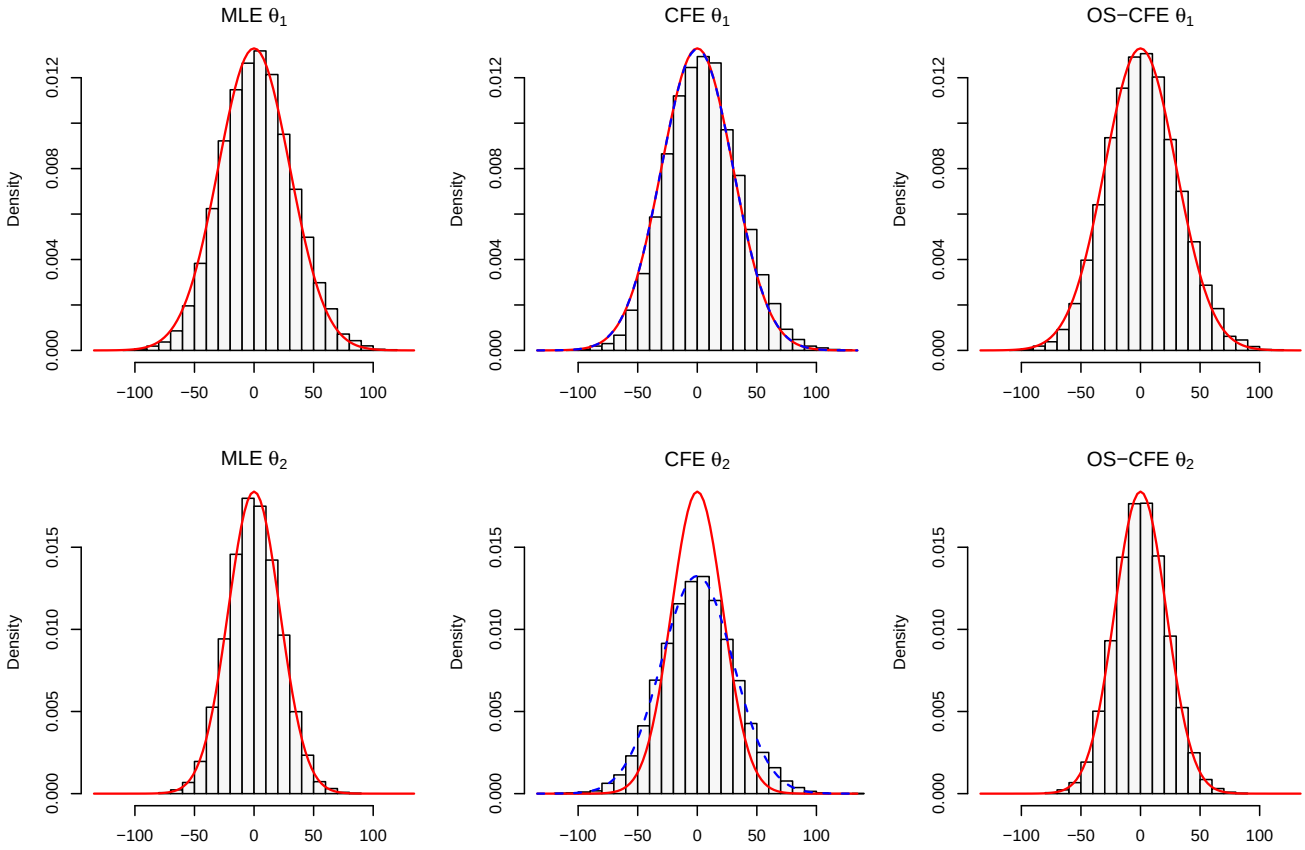

Figure 2: Histograms for the $B = 10^4$ Monte Carlo simulations of the renormalized statistical errors of MLE, CFE, OS-CFE for the Gamma distribution (canonical link) with 2 categorical variables with $d_2 = 2$, $d_3 = 3$ for $\theta_1 = \vartheta_1$ and $\theta_2 = \vartheta_2^{(2)}$ and fixed $\phi = 8$. Red and blue lines are the theoretical Gaussian asymptotic densities respectively of the MLE (in red) and CFE (in blue).

Computation time	MLE	CFE	OS-CFE
Poisson	848.07	9.05	17.73
Gamma	1601.44	10.65	31.61

Table 4: Total computation time (s) based on $B = 10^4$ runs for Poisson and Gamma distributions

4.2 Simulations with an increasing sample size

In this section for both Poisson and Gamma GLMs keeping the number of modalities constant ($d_2 = 2$ and $d_3 = 3$), the sample sizes are increasing from $n = 10^3$ to $n = 10^7$. The number of Monte Carlo simulations are set to $B = 100$. The total computation time of the three estimators are computed. The comparative gain of the OS-CFE over the MLE is increasing with the sample size for both distributions, from 4 to 80 times faster as shown in Tables 5 and 6.

Dataset size	10^3	10^4	10^5	10^6	10^7
MLE	0.33	2.35	24.86	257.15	2678.38
CFE	0.07	0.06	0.26	2.50	26.77
OS-CFE	0.08	0.11	0.50	5.14	56.10

Table 5: Total computation time (s) based on $B = 100$ runs for Poisson distribution.

Dataset size	10^3	10^4	10^5	10^6	10^7
MLE	0.41	3.64	30.63	389.88	5120.97
CFE	0.06	0.06	0.29	2.96	38.29
OS-CFE	0.07	0.11	0.52	5.30	64.98

Table 6: Total computation time (s) based on $B = 100$ runs for Gamma distribution.

4.3 Simulations with an increasing number of modalities

Table 7 summarizes the results of the Monte Carlo simulations with an increasing number of modalities when the number of simulations are fixed to $B = 100$ and sample size is $n = 10^5$. Simulations have been done for 2 categorical variables with equal number of modalities $d_2 = d_3 = d$ varying from $d = 5$ to 40.

In this particular case, the true parameter value for Gamma distribution are chosen by simply taking the $\vartheta_1 = 3d + 1$, the $\vartheta_k^{(j)}$ are equal to k/d for $k = 1, \ldots, d-1$, and the $\vartheta_d^{(j)} = -\frac{(d-1)}{2}$ for the two variables $j = 1, 2$.

In addition, for the Poisson distribution case, ϑ_1 has been chosen to be equal to 0.5, $\vartheta_1^{(j)}$ to $\vartheta_{d-1}^{(j)}$ are $1, 2, \ldots, 1$ or 2 each divided by the sum of all 1 and 2's, and the $\vartheta_d^{(j)}$ is the minus sum of all modalities for each of the two variables.

The data presented in Table 7 demonstrates that as the number of modalities increases, the computation time for all estimators also increases for both Poisson and Gamma GLMs. The OS-

CFE method demonstrates faster computation times when compared to the MLE method, even at the highest number of modalities.

d	Poisson			Gamma		
	MLE	CFE	OS-CFE	MLE	CFE	OS-CFE
5	36.49	0.31	0.71	50.59	0.41	0.67
10	71.95	0.58	0.92	109.39	0.60	0.91
15	135.12	0.97	1.56	175.90	0.79	1.67
20	150.19	1.15	2.42	181.79	0.92	1.92
25	173.01	1.42	3.97	346.64	2.23	5.50
30	264.51	2.61	10.27	429.85	2.51	9.58
35	254.86	3.23	14.10	618.42	3.97	19.96
40	343.16	3.85	25.74	675.58	4.60	28.12

Table 7: Total computation times based on $B = 100$ runs for Poisson (canonical link) and Gamma (canonical link) GLMs.

5 Application to claim amounts in car insurance

The Covea Affinity dataset under study is composed of 76,446 claim amounts ranging from 4 to 33,531 EUR. Three covariates have been selected from the 124 available for the pricing of the guarantee

- vehicle brand with $d_2 = 2$ modalities,
- pricing segment with $d_3 = 6$ modalities,
- age class with $d_4 = 8$ modalities.

For confidentiality reasons, the modality values are not revealed.

The single effect models

$$g\left(\mathbf{E}_{\boldsymbol{\vartheta}} Y_i\right) = \vartheta_1 + \sum_{j=2}^{m+1} \sum_{k=1}^{d_j} x_i^{(j),k} \vartheta_k^{(j)},$$

is generally used by the insurers to model the claim amounts with (non-canonical) Gamma GLMs with a log link function ($g(x) = \log(x)$). The "reference category" linear contrast has been used. It is worth recalling that in this setting the MLE has no closed-form and the closed-form estimator is not efficient. In order to compute the log-likelihood, we use Equation (3) to fit the dispersion parameter.

The one-step estimator has been applied to the Covea dataset: Table 8 give parameter estimates for MLE, CFE and OS-CFE. The CFE and OS-CFE were almost 30 times faster to obtain than the MLE, with similar estimate and similar fitted log-likelihood.

	CFE	OS-CFE	MLE
ϑ_1	6.23	6.04	6.03
$\vartheta_2^{(2)}$	0.24	0.08	0.03
$\vartheta_2^{(3)}$	0.18	0.22	0.22
$\vartheta_3^{(3)}$	-0.48	0.04	-0.01
$\vartheta_4^{(3)}$	-0.07	0.08	0.09
$\vartheta_5^{(3)}$	0.06	0.18	0.19
$\vartheta_6^{(3)}$	0.20	0.21	0.22
$\vartheta_2^{(4)}$	-0.07	0.00	0.01
$\vartheta_3^{(4)}$	0.06	0.16	0.16
$\vartheta_4^{(4)}$	0.17	0.18	0.18
$\vartheta_5^{(4)}$	0.34	0.41	0.40
$\vartheta_6^{(4)}$	0.11	0.44	0.42
$\vartheta_7^{(4)}$	0.16	0.25	0.26
$\vartheta_8^{(4)}$	-0.01	0.34	0.33
$\log \mathcal{L}$	-554,868	-553,708	-553,685
Time (s)	0.01	0.01	0.30

Table 8: Values of $\widehat{\boldsymbol{\vartheta}}_n$, log-likelihood and total computation time (s) for CFE, OS-CFE and MLE

6 Conclusion

Generalized linear models with single effects and sole categorical explanatory variables are widely used in different applications (e.g., insurance, agriculture, biology). On the one hand, the classical iteratively re-weighted least-square calibration algorithm is asymptotically efficient but can be time consuming for large datasets and/or large number of variables. On the other hand, closed-form estimators proposed in Brouste et al. (2022) are fast to be computed but are not asymptotically efficient.

In this paper, we proposed a fast and asymptotically efficient method for the calibration of GLMs with categorical explanatory variables based on the one-step procedure that can also be applied to single effect models. It is 30 times faster than the classical method on both simulated and real datasets.

The proposed estimator could be further used for variable and/or model selection. For instance, the model selection with fastly computable Akaike Information Criterion could be treated in a further work. The one-step procedure can be extended in many directions: e.g., for multivariate regression models.

Acknowledgments We would like to thank Héloïse Bertrand and Nicolas Villette for fruitful discussions on the topic. This research benefited from the support of the ANR project 'Efficient inference for large and high-frequency data' (ANR-21-CE40-0021), the 'Chair Risques Émergents ou Atypiques en Assurance', under the aegis of Fondation du Risque, a joint initiative by Le Mans Université and MMA company, member of Covea group and the 'Chair Impact de la Transition Climatique en Assurance', under the aegis of Fondation du Risque, a joint initiative by Le Mans Université and Groupama Centre-Manche company, member of Groupama group. This preprint

has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this article is published in Statistics and Computing, and is available online at https://10.1007/s11222-023-10313-4.

References

Brouste, A., Dutang, C. & Ntousa Meniedou, D. (2021), 'Onestep - Le Cam's onestep estimation procedure', *R Journal* **13**(1), 366–377.

Brouste, A., Dutang, C. & Rohmer, T. (2020), 'Closed-form maximum likelihood estimator for generalized linear models in the case of categorical explanatory variables: application to insurance loss modeling', *Computational Statistics* **35**(2), 689–724.

Brouste, A., Dutang, C. & Rohmer, T. (2022), 'A closed-form alternative estimator for GLM with categorical explanatory variables', *Communications in Statistics-Simulation and Computation* pp. 1–17.

Brouste, A. & Masuda, H. (2018), 'Efficient estimation of stable Lévy process with symmetric jumps', *Statistical Inference for Stochastic Processes* **21**, 289–307.

Brouste, A., Soltane, M. & Votsi, E. (2020), 'Onestep estimation for the fractional gaussian noise model at high-frequency', *ESAIM Probability and Statistics* **24**, 827–841.

Dabye, A., Gounoung, A. & Kutoyants, Y. (2018), 'Method of moments estimators and multi-step mle for poisson processes', *Journal of Contemporary Mathematical Analysis* **53**(4), 187–196.

Denuit, M., Hainaut, D. & Trufin, J. (2020), *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*, Springer Nature.

Fahrmeir, L. & Kaufmann, H. (1985), 'Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models', *The Annals of Statistics* pp. 342–368.

Gloter, A. & Yoshida, N. (2021), 'Adaptive estimation for degenerate diffusion processes', *Electronic Journal of Statistics* **15**(1), 1424–1472.

Kadarmideen, H., Thompson, R. & Simm, G. (2000), 'Linear and threshold model genetic parameters for disease, fertility and milk production in dairy cattle', *Animal Science* **71**, 411–419.

Kamatani, K. & Uchida, M. (2015), 'Hybrid multi-step estimators for stochastic differential equations based on sampled data', *Statistical Inference for Stochastic Processes* **18**(177–204).

Kutoyants, Y. & Motrunich, A. (2016), 'On multi-step MLE-process for markov sequences', *Metrika* **79**(705–724).

Le Cam, L. (1956), 'On the asymptotic theory of estimation and testing hypothesis,', *Proceedings of the 3rd Berkeley Symposium* **1**, 355–368.

Lindsey, J. (1997), *Applying Generalized Linear Models*, Springer Texts in Statistics.

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, CRC press.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Venables, B. (2023), *codingMatrices: Alternative Factor Coding Matrices for Linear Model Formulae*. R package version 0.4.0.
URL: *https://CRAN.R-project.org/package=codingMatrices*

Wuethrich, M. & Merz, M. (2021), Statistical foundations of actuarial learning and its applications. SSRN papers.
URL: *https://ssrn.com/abstract=3822407*

A Gradient and Hessian of the log-likelihood

Let $\tilde{b} = (b')^{-1}$. We recall that

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(p)} \\ & \vdots & \\ x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

A.1 Gradient and Hessian of a single observation

For ease of notation, we consider a single observation in (1). We recall that $\mu_i = g^{-1}(\eta_i)$ and $V(\mu_i) = b''(\lambda_i(\boldsymbol{\vartheta}))$.

The ith contribution is

$$l_i(\boldsymbol{\vartheta}) = \frac{\lambda_i(\boldsymbol{\vartheta}) y_i - b\left(\lambda_i(\boldsymbol{\vartheta})\right)}{a(\phi)} + c(y_i, \phi). \tag{21}$$

To derive the first and second order derivatives, we first compute

$$\frac{\partial \lambda_i}{\partial \vartheta_j} = (\tilde{b})'(g^{-1}(\eta_i)) \times (g^{-1})'(\eta_i) \times x_i^{(j)} = \frac{1}{b''(\tilde{b}(g^{-1}(\eta_i))} \times \frac{x_i^{(j)}}{g'(g^{-1}(\eta_i))} = \frac{x_i^{(j)}}{g'(\mu_i)V(\mu_i)}.$$

Hence

$$\frac{\partial l_i(\boldsymbol{\vartheta})}{\partial \vartheta_j} = \frac{y_i - b'(\lambda_i(\boldsymbol{\vartheta}))}{a(\phi)} \frac{\partial \lambda_i}{\partial \vartheta_j} = x_i^{(j)} \frac{y_i - \mu_i}{a(\phi) g'(\mu_i) V(\mu_i)}. \tag{22}$$

Furthermore,

$$\begin{aligned} \frac{\partial^2 l_i(\vartheta)}{\partial \vartheta_j \partial \vartheta_l} &= \frac{\partial}{\partial \vartheta_l} \left(\frac{y_i - b'(\lambda_i)}{a(\phi_i)} \right) \frac{1}{b''(\lambda_i)} \frac{x_i^{(j)}}{g'(\mu_i)} + \frac{\partial}{\partial \vartheta_l} \left(\frac{1}{b''(\lambda_i)} \right) \frac{y_i - b'(\lambda_i)}{a(\phi)} \frac{x_i^{(j)}}{g'(\mu_i)} \\ &\quad + \frac{\partial}{\partial \vartheta_l} \left(\frac{x_i^{(j)}}{g'(\mu_i)} \right) \frac{y_i - b'(\lambda_i)}{a(\phi)} \frac{1}{b''(\lambda_i)}. \end{aligned}$$

The first derivative term is

$$\frac{\partial}{\partial \vartheta_l} \left(\frac{y_i - b'(\lambda_i)}{a(\phi)} \right) = \frac{-b''(\lambda_i)}{a(\phi)} \times \frac{\partial \lambda_i}{\partial \vartheta_l} = \frac{-b''(\lambda_i)}{a(\phi)} \times \frac{1}{b''(\lambda_i)} \frac{x_i^{(l)}}{g'(\mu_i)} = \frac{-x_i^{(l)}}{a(\phi) g'(\mu_i)}.$$

The second derivative term is

$$\frac{\partial}{\partial\vartheta_l}\left(\frac{1}{b''(\lambda_i)}\right) = -\frac{b'''(\lambda_i)}{(b''(\lambda_i))^2}\frac{\partial\lambda_i}{\partial\vartheta_l} = -\frac{b'''(\lambda_i)}{(b''(\lambda_i))^3}\frac{x_i^{(l)}}{g'(\mu_i)} = -\frac{b'''(\lambda_i)}{(V(\mu_i))^3}\frac{x_i^{(l)}}{g'(\mu_i)}.$$

The third derivative term is

$$\frac{\partial}{\partial\vartheta_l}\left(\frac{x_i^{(j)}}{g'(\mu_i)}\right) = -\frac{x_i^{(j)}}{(g'(\mu_i))^2}\frac{\partial(g'(\mu_i))}{\partial\vartheta_l} = -\frac{x_i^{(j)}g''(\mu_i)}{(g'(\mu_i))^2}\frac{\partial\mu_i}{\partial\vartheta_l} = -\frac{x_i^{(j)}x_i^{(l)}g''(\mu_i)}{(g'(\mu_i))^3},$$

since

$$\frac{\partial\mu_i}{\partial\vartheta_l} = \frac{\partial(b'(\lambda_i))}{\partial\vartheta_l} = b''(\lambda_i)\frac{\partial\lambda_i}{\partial\vartheta_l} = b''(\lambda_i)\times\frac{1}{b''(\lambda_i)}\frac{x_i^{(l)}}{g'(\mu_i)} = \frac{x_i^{(l)}}{g'(\mu_i)}.$$

This leads to

$$\frac{\partial^2 l_i(\vartheta)}{\partial\vartheta_j\partial\vartheta_l} = -\frac{x_i^{(l)}x_i^{(j)}}{a(\phi_i)V(\mu_i)g'(\mu_i)^2} - \frac{y_i-b'(\lambda_i)}{a(\phi_i)}\frac{x_i^{(j)}x_i^{(l)}b'''(\lambda_i)}{g'(\mu_i)^2(V(\mu_i))^3} - \frac{y_i-b'(\lambda_i)}{a(\phi_i)}\frac{x_i^{(j)}x_i^{(l)}g''(\mu_i)}{(g'(\mu_i))^3V(\mu_i)}. \quad (23)$$

A.2 Score and Fisher information

The score component is obtained by summing (22) over observations

$$\mathcal{S}_j(\boldsymbol{\vartheta}) = \frac{\partial l_i(\boldsymbol{\vartheta})}{\partial\vartheta_j} = \frac{1}{a(\phi)}\sum_{i=1}^{n} x_i^{(j)}\frac{y_i-\mu_i}{g'(\mu_i)V(\mu_i)}. \quad (24)$$

Using the matrices of weights, covariables (4), and the mean deviation vector (5), the score is obtained by

$$\mathcal{S}(\boldsymbol{\vartheta}) = \frac{1}{a(\phi)}X^T W(\boldsymbol{\vartheta})Z(\boldsymbol{\vartheta}). \quad (25)$$

Therefore, the information matrix is obtained by summing (23) over observations and taking the expectation with respect to Y_i (using $\mathbf{E}(Y_i)-\mu_i=0$)

$$\mathcal{I}_{n,l,j}(\boldsymbol{\vartheta}) = -\mathbf{E}\left(\sum_i \frac{\partial^2 l_i(\vartheta)}{\partial\vartheta_j\partial\vartheta_l}\right) = \sum_{i=1}^{n}\frac{x_i^{(l)}x_i^{(j)}}{a(\phi)V(\mu_i)g'(\mu_i)^2}. \quad (26)$$

So that the Fisher information matrix is

$$\mathcal{I}_n(\boldsymbol{\vartheta}) = (\mathcal{I}_{n,l,j}(\boldsymbol{\vartheta}))_{l,j} = \frac{X^T W(\boldsymbol{\vartheta})X}{a(\phi)}. \quad (27)$$

A.3 Information matrix and score in term of $\tilde{Q}$

Define $\tilde{q}_j^{(k)}$ the element of the jth row and kth column of $\tilde{Q}$. Note that the ith row of $\tilde{X}$ is equal to the jth row of $\tilde{Q}$ for all i such that $\eta_i = h_j$. Using (26), the information matrix is $\tilde{\mathcal{I}}_n(\boldsymbol{\vartheta}) = (\tilde{\mathcal{I}}_{n,k,l}(\boldsymbol{\vartheta}))_{k,l=1,\dots,p^\star}$ with

$$\begin{aligned}\tilde{\mathcal{I}}_{n,k,l}(\boldsymbol{\vartheta}) &= \frac{1}{a(\phi)}\sum_{i=1}^{n}\frac{\tilde{x}_i^{(k)}\tilde{x}_i^{(l)}}{V(\mu_i)(g'(\mu_i))^2} = \frac{1}{a(\phi)}\sum_{j=1}^{d}\sum_{i=1;\eta_i=h_j}^{n}\frac{\tilde{x}_i^{(k)}\tilde{x}_i^{(l)}}{V(\mu_i)(g'(\mu_i))^2}\\ &= \frac{1}{a(\phi)}\sum_{j=1}^{d}\frac{1}{V(\mu_j^\star)(g'(\mu_j^\star))^2}\sum_{i=1;\eta_i=h_j}^{n}\tilde{x}_i^{(k)}\tilde{x}_i^{(l)} = \frac{1}{a(\phi)}\sum_{j=1}^{d}\frac{m_j}{V(\mu_j^\star)(g'(\mu_j^\star))^2}\tilde{q}_j^{(k)}\tilde{q}_j^{(l)}.\end{aligned}$$

Hence the information matrix rewrites similarly to the non-restricted case

$$\tilde{\mathcal{I}}_n(\boldsymbol{\vartheta}) = \frac{n\tilde{Q}^T\Sigma_n(\boldsymbol{\vartheta})\tilde{Q}}{a(\phi)}.$$

Now, using (22) the score component is

$$\begin{aligned}\tilde{U}_k(\boldsymbol{\vartheta}) &= \sum_{i=1}^{n} \frac{\tilde{x}_i^{(k)}(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)} \\ &= \sum_{j=1}^{d} \frac{1}{V(\mu_j^\star)g'(\mu_j^\star)} \sum_{i=1;\eta_i=h_j}^{n} \tilde{q}_j^{(k)}(y_i-\mu_j^\star) \\ &= \sum_{j=1}^{d} \frac{m_j\tilde{q}_j^{(k)}(\bar{y}_n^{(j)}-\mu_j^\star)}{V(\mu_j^\star)g'(\mu_j^\star)}.\end{aligned}$$

This yields to

$$\tilde{U}(\boldsymbol{\vartheta}) = n\tilde{Q}^T\Sigma_n(\boldsymbol{\vartheta})Z^\star(\boldsymbol{\vartheta}).$$

B Proof of Theorem 1

Let us recall that the One-Step Closed-form Estimator (OS-CFE) is defined as

$$\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} = \tilde{\boldsymbol{\vartheta}}_n^{CFE} + \tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})^{-1}\tilde{S}(\tilde{\boldsymbol{\vartheta}}_n^{CFE}). \tag{28}$$

Let $\ell_n(\boldsymbol{\vartheta}) = \log \mathcal{L}(\boldsymbol{\vartheta} \,|\, \boldsymbol{Y})$. The mean-value theorem gives, for the initial sequence of guess estimators $(\tilde{\boldsymbol{\vartheta}}_n^{CFE}, n \geq 1)$,

$$\begin{aligned}\frac{\partial}{\partial\vartheta}\ell_n\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE}\right) &= \frac{\partial}{\partial\vartheta}\ell_n\left(\widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) + \int_0^1 \frac{\partial^2}{\partial\vartheta^2}\ell_n\left(\widehat{\tilde{\boldsymbol{\vartheta}}}_n + v\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)\right)dv \cdot \left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) \\ &= \int_0^1 \frac{\partial^2}{\partial\vartheta^2}\ell_n\left(\widehat{\tilde{\boldsymbol{\vartheta}}}_n + v\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)\right)dv \cdot \left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) \qquad (29)\end{aligned}$$

since $\frac{\partial}{\partial\vartheta}\ell_n\left(\widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) = 0$ by definition. From (28), we have

$$\left(\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) = \left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) + \tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})^{-1} \cdot \frac{\partial}{\partial\vartheta}\ell_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})$$

and

$$\left(\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) = \left(I_p + \tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})^{-1}\int_0^1 \frac{\partial^2}{\partial\vartheta^2}\ell_n\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} + v\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)\right)dv\right)\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)$$

where I_p is the $p \times p$ identity matrix.

In our setting, since the sequence of initial guess estimators is $\sqrt{n}$-consistent (see Equation (19)), we get the asymptotic equivalence by showing that the quantity

$$\sqrt{n}\left(\tilde{\boldsymbol{\vartheta}}_n^{OS-CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)$$
$$= \underbrace{\left(I_p + \left(\frac{\tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}}_n^{CFE})}{n}\right)^{-1} \int_0^1 \frac{1}{n}\frac{\partial^2}{\partial\vartheta^2}\ell_n\left(\widehat{\tilde{\boldsymbol{\vartheta}}}_n + v\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)\right) dv\right)}_{(A)} \cdot \underbrace{\sqrt{n}\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right)}_{(B)} \quad (30)$$

converges to zero as $n \to \infty$.

- For the quantity (B), we get from (15) and (18)

  $$\sqrt{n}\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \widehat{\tilde{\boldsymbol{\vartheta}}}_n\right) = \sqrt{n}\left(\tilde{\boldsymbol{\vartheta}}_n^{CFE} - \boldsymbol{\vartheta}\right) - \sqrt{n}\left(\widehat{\tilde{\boldsymbol{\vartheta}}}_n - \boldsymbol{\vartheta}\right)$$

  is bounded in probability.

- For the quantity (A), by Markov's law of large number and Equation (23) we have

  $$\frac{1}{n}\frac{\partial^2}{\partial\vartheta^2}\ell_n(\tilde{\boldsymbol{\vartheta}}) + \frac{\tilde{\mathcal{I}}_n(\tilde{\boldsymbol{\vartheta}})}{n} \xrightarrow[n\to+\infty]{a.s.} 0$$

  Using the consistency of the initial guess estimator and the uniform continuity of the Fisher information matrix (16), we get the convergence to zero of the quantity (A).