A Convenient Method for the Estimation of the Multinomial Logit Model with Fixed Effects*

Xavier D'Haultfœuille† Alessandro Iaria‡

December 22, 2015

Abstract

The conditional maximum likelihood estimator of the fixed effect logit model suffers from a curse of dimensionality that may have severely limited its use in practice. As the number of alternatives and the number of choice situations per individual increase, the number of addends in the denominator of the fixed effect logit formula grows exponentially. We propose to by-pass this curse of dimensionality by exploiting a classic result by McFadden (1978) and to consistently estimate the fixed effect logit model on random samples of permutations of the observed choice sequences.

*We wish to thank Greg Crawford and an anonymous referee: the present version of the paper greatly benefited from their comments. We We gratefully acknowledge financial support from the research grant Labex Ecodec: Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

†CREST. Email address: xavier.dhaultfoeuille@ensae.fr

‡CREST. Email address: alessandro.iaria@ensae.fr

1 Introduction

As Chamberlain (1980) showed (see also Rasch, 1961, for an earlier reference), the fixed-effect (FE) logit model can be estimated consistently by a conditional maximum likelihood estimator (CMLE). However, it is well known that this estimator suffers from a "curse of dimensionality" that may have severely limited its use in practice (e.g., Arellano & Honoré, 2001, p. 3269, Baltagi, 2005, p. 235, and Greene, 2011, p. 723): as the number of alternatives and the number of choice situations per individual increase, the number of addends in the denominator of the FE logit formula grows exponentially.

Suppose there are $i = 1, \ldots, n$ individuals who are observed making discrete choices from $j = 1, \ldots, J$ alternatives over $t = 1, \ldots, T$ choice situations. Vector $\boldsymbol{Y}_{it} = (Y_{i1t}, \ldots, Y_{iJt})$, with elements $Y_{ijt} = \mathbb{1}\{i \text{ chooses } j \text{ in } t\}$, indicates i's choice in choice situation t. Individual i's sequence of choices is $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT})$. Then, the distribution of times i chose each of the J alternatives over the T choice situations is the Chamberlinian statistic $\sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i = (c_{i1}, \ldots, c_{iJ})$, which is sufficient for the individual fixed effects.

An impractical feature of the FE logit model is its growing "size" with respect to J and T. Indeed, i's probability of choosing any sequence of alternatives compatible with the distribution of choices $\boldsymbol{c}_i$ is a multinomial logit model with

$$\frac{T!}{c_{i1}! \cdots c_{ij}! \cdots c_{iJ}!} \tag{1}$$

addends in the denominator (see Chamberlain, 1980, p. 231). This number increases sharply with T and potentially with J. For example, with $T = 10$ and $J = 2$, the maximum number of addends would be 252, but with $J \geq 10$, it would be over 3.5 million.

Currently, statistical software such as STATA implements the estimation of the FE logit model by CMLE, thus computing and including in the denominator of the FE logit formula *all* the addends in (1).[1] In applications with large datasets, this may cause the practical computation of the CMLE to be at best extremely time consuming and, at worst, infeasible altogether.

As Chamberlain (1980, p. 231) shows, the FE logit model is a multinomial logit model with respect to the set of permutations of the observed sequence of choices $\boldsymbol{Y}_i = \boldsymbol{s}_i$. Using this idea and exploiting a classical result by McFadden (1978), we propose an alternative estimator that by-passes the "curse of dimensionality" of the CMLE. Specifically, we estimate the model by using random subsets of sequences consistent with $\sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i$.

In Section 2, we define the new estimator and derive its asymptotic properties. In Section 3,

[1] For example, STATA provides two such commands: *clogit* for the binary case ($J = 2$) and the recent *femlogit* for the general case (see Pforr, 2014).

we report the results of a Monte Carlo exercise that illustrates the practical usefulness of the proposed estimator.

2 Random Samples of Permutations of Observed Choice Sequences

Suppose the indirect utility of individual i for alternative j in choice situation t is $\mathscr{U}_{ijt} = \alpha_{ij} + \boldsymbol{X}'_{ijt}\boldsymbol{\beta} + \varepsilon_{ijt}$, where α_{ij} is i's fixed effect for alternative j, $\boldsymbol{X}_{ijt}$ a vector of time varying regressors potentially correlated with α_{ij}, $\boldsymbol{\beta}$ a vector of preference parameters, and ε_{ijt} are i.i.d. and follow a Gumbel distribution. For any observed sequence of choices $\boldsymbol{Y}_i = \boldsymbol{s}_i$ with $\sum_{t=1}^{T} \boldsymbol{s}_{it} = \boldsymbol{c}_i$, let $\mathcal{P}(\boldsymbol{c}_i)$ be the set of *all permutations* of i's observed sequence of choices: $\mathcal{P}(\boldsymbol{c}_i) = \left\{ \boldsymbol{k} = (\boldsymbol{k}_1, ..., \boldsymbol{k}_T) | \sum_{t=1}^{T} \boldsymbol{k}_t = \boldsymbol{c}_i \right\}$. Finally, for any sequence $\boldsymbol{k}$, let $\boldsymbol{Z}_i(\boldsymbol{k}) = \sum_{t=1}^{T} \boldsymbol{X}_{i\boldsymbol{k}_t t}$. Then, as shown by Chamberlain (1980, p. 231), i's probability of choosing any sequence $\boldsymbol{Y}_i = \boldsymbol{s}_i$ compatible with $\boldsymbol{c}_i$ is:

$$\begin{aligned}
\Pr\left[\boldsymbol{Y}_i = \boldsymbol{s}_i \middle| \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i, \boldsymbol{\beta} \right] &= \Pr\left[\boldsymbol{Y}_i = \boldsymbol{s}_i | \, \mathcal{P}(\boldsymbol{c}_i), \boldsymbol{\beta} \right] \\
&= \frac{\prod_{t=1}^{T} \exp\left(\alpha_{i\boldsymbol{s}_{it}} + \boldsymbol{X}'_{i\boldsymbol{s}_{it}t}\boldsymbol{\beta} \right)}{\sum_{\boldsymbol{k} \in \mathcal{P}(\boldsymbol{c}_i)} \prod_{t=1}^{T} \exp\left(\alpha_{i\boldsymbol{k}_t} + \boldsymbol{X}'_{i\boldsymbol{k}_t t}\boldsymbol{\beta} \right)} \\
&= \frac{\exp\left(\boldsymbol{Z}_i(\boldsymbol{s}_i)'\boldsymbol{\beta} \right)}{\sum_{\boldsymbol{k} \in \mathcal{P}(\boldsymbol{c}_i)} \exp\left(\boldsymbol{Z}_i(\boldsymbol{k})'\boldsymbol{\beta} \right)},
\end{aligned} \tag{2}$$

which is a multinomial logit model for the choice of sequences of alternatives over the set $\mathcal{P}(\boldsymbol{c}_i)$. The set $\mathcal{P}(\boldsymbol{c}_i)$ contains $T!\left(c_{i1}! \cdots c_{ij}! \cdots c_{iJ}!\right)^{-1}$ sequences of alternatives, which can result in an infeasibly large number of addends in the denominator of multinomial logit formula (2).

We now introduce our estimator. Following McFadden (1978, p. 87-91), our idea is to restrict $\mathcal{P}\left(\sum_{t=1}^{T} \boldsymbol{Y}_{it}\right)$ to a random subset D_i. Let $\pi\left[D_i = d_i | \, \boldsymbol{Y}_i = \boldsymbol{s}_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i \right]$ be the probability of drawing d_i. We focus on sampling schemes satisfying the *uniform conditioning property*, meaning that for any two sequences of alternatives $\boldsymbol{s}_i$, $\boldsymbol{k}_i$ in D_i, $\pi\left[D_i = d_i | \, \boldsymbol{Y}_i = \boldsymbol{s}_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i \right] = \pi\left[D_i = d_i | \, \boldsymbol{Y}_i = \boldsymbol{k}_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i \right]$.[2] Then $\boldsymbol{\beta}$ can be consistently estimated by maximizing

[2] Since the sampling scheme is devised by the econometrician, it is always possible to satisfy this property.

the log-likelihood function:

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \ln\left(\Pr\left[\boldsymbol{Y}_i = \boldsymbol{s}_i | \, D_i = d_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i, \boldsymbol{\beta}\right]\right) \\ &= \sum_{i=1}^{n} \ln\left(\frac{\pi\left[D_i = d_i | \, \boldsymbol{Y}_i = \boldsymbol{s}_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i\right] \Pr\left[\boldsymbol{Y}_i = \boldsymbol{s}_i | \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i, \boldsymbol{\beta}\right]}{\sum_{\boldsymbol{k} \in d_i} \pi\left[D_i = d_i | \, \boldsymbol{Y}_i = \boldsymbol{k}, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i\right] \Pr\left[\boldsymbol{Y}_i = \boldsymbol{k} | \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i, \boldsymbol{\beta}\right]}\right) \\ &= \sum_{i=1}^{n} \ln\left(\frac{\exp\left(\boldsymbol{Z}_i(\boldsymbol{s}_i)'\boldsymbol{\beta}\right)}{\sum_{\boldsymbol{k} \in d_i} \exp\left(\boldsymbol{Z}_i(\boldsymbol{k})'\boldsymbol{\beta}\right)}\right). \end{aligned} \tag{3}$$

The second equality follows from Bayes's rule and $\pi\left[D_i = d_i | \, \boldsymbol{Y}_i = \boldsymbol{k}_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i\right] = 0$ for any $\boldsymbol{k}_i \notin d_i$, and the third from the *uniform conditioning property*. Note how, for each i, the summation in the denominator of (3) is over d_i rather than $\mathcal{P}(\boldsymbol{c}_i)$.

An easily implementable random sampling scheme that satisfies the *uniform conditioning property* is to select D_i to be a set of $L+1$ sequences containing *(a)* the observed sequence of choices $\boldsymbol{Y}_i$ and *(b)* L other randomly drawn sequences from $\mathcal{P}\left(\sum_{t=1}^{T} \boldsymbol{Y}_{it}\right)$ without replacement. Denote by $R_i = T!\left(c_{i1}! \cdots c_{ij}! \cdots c_{iJ}!\right)^{-1}$ the number of elements in $\mathcal{P}(\boldsymbol{c}_i)$, then for any $\boldsymbol{s}_i \in d_i$:

$$\pi\left[D_i = d_i | \boldsymbol{Y}_i = \boldsymbol{s}_i, \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i\right] = \binom{R_i - 1}{L}^{-1} = \pi\left[D_i = d_i | \sum_{t=1}^{T} \boldsymbol{Y}_{it} = \boldsymbol{c}_i\right], \tag{4}$$

which implies that the *uniform conditioning property* indeed holds.

The next proposition provides the asymptotic properties of our estimator, $\widehat{\boldsymbol{\beta}}_{\text{sub}}$, and compares them with those of the CMLE, $\widehat{\boldsymbol{\beta}}_{CMLE}$. Let us introduce some additional notation. Because we consider i.i.d. samples, we can omit individual indices hereafter. Let $\boldsymbol{X} = (\boldsymbol{X}'_{11}, ..., \boldsymbol{X}'_{JT})'$, $P_{\boldsymbol{j}d} = \Pr(\boldsymbol{Y} = \boldsymbol{j} | D = d, \boldsymbol{X})$, $P_{\boldsymbol{j}} = \Pr(\boldsymbol{Y} = \boldsymbol{j} | \boldsymbol{X})$, $\overline{\boldsymbol{Z}}_d = \sum_{\boldsymbol{j} \in d} P_{\boldsymbol{j}d} \boldsymbol{Z}(\boldsymbol{j})$, $\mathcal{P} = \{0,1\}^{JT}$ and

$$\mathcal{I}(\widehat{\boldsymbol{\beta}}_{\text{sub}}) = \sum_{\boldsymbol{j} \in \mathcal{P}} E\left[P_{\boldsymbol{j}} \boldsymbol{Z}(\boldsymbol{j}) \boldsymbol{Z}(\boldsymbol{j})'\right] - E\left[\boldsymbol{Z}(\boldsymbol{j}) E\left[P_{\boldsymbol{j}D} \overline{\boldsymbol{Z}}'_D | \boldsymbol{X}\right]\right].$$

Proposition 1 *Suppose that* $(\boldsymbol{Y}_i, \boldsymbol{X}_i)_{i=1,...,n}$ *are i.i.d. and that* $\mathcal{I}(\widehat{\boldsymbol{\beta}}_{sub})$ *is nonsingular. Then:*

Intuition suggests that more elaborate sampling schemes may increase the efficiency of the estimator. For an early example of a sampling scheme that does not satisfy the *uniform conditioning property*, see Train, McFadden, and Ben-Akiva (1987).

1. $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{sub} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\widehat{\boldsymbol{\beta}}_{sub})^{-1}\right)$.

2. *$\widehat{\boldsymbol{\beta}}_{sub}$ is asymptotically less efficient than $\widehat{\boldsymbol{\beta}}_{CMLE}$: $\mathcal{I}(\widehat{\boldsymbol{\beta}}_{sub})^{-1}$ is larger than the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{CMLE}$.*

The proposition is proved in our supplement. Note that the asymptotic variance takes a similar form as that of the multinomial logit (see, e.g., Amemiya, 1985, pp. 288 and 296). Since $\widehat{\boldsymbol{\beta}}_{\text{sub}}$ is also a conditional maximum likelihood estimator, classical results (see, e.g., Andersen, 1970) imply that one can compute standard errors using the hessian of the objective function.

3 Monte Carlo Exercise

Each simulated dataset has 1000 individuals making choices over T choice situations. During each choice situation, individuals can choose from five alternatives $\{a, b, c, d, e\}$. The conditional indirect utility of individual i for alternative j in choice situation t is:

$$\mathscr{U}_{ijt} = \alpha_{ij} + \beta_1 x_{ijt1} + \beta_2 x_{ijt2} + \varepsilon_{ijt}, \quad \varepsilon_{ijt} \sim \text{Gumbel}. \tag{5}$$

Both regressors are independently and normally distributed with $x_{ijt1} \sim 3 \cdot \text{Normal}\,(\mu_j, 1)$ and $x_{ijt2} \sim \text{Normal}\,(\mu_j, 1)$, where $\mu_a = 0$, $\mu_b = 0.5$, $\mu_c = 1$, $\mu_d = 1.5$, and $\mu_e = 2$. Preferences take values $(\boldsymbol{\alpha}, \beta_1, \beta_2) = (\mathbf{0}, -2, 2)$. Since their effect on computation time is null, $\alpha_{ij} = 0$ for all (i, j).

We generate a hundred datasets for each of six data generating processes (DGP's) and then average the results. The six DGP's differ in the number of choice situations T each individual faces. For each of the simulated datasets, we estimate parameters β_1 and β_2 from the multinomial FE logit model implied by (5) using both the CMLE and our estimator. To compute our estimator, we use sets of 2,500, 5,000, and 10,000 permutations of the observed sequences of choices.

For each method, the average computation time required for estimation by a desktop computer with typical computing power is divided into two components: *(permutation time)* the amount of time required to enumerate all the addends in the denominator of the respective multinomial FE logit formula and, given the denominator, *(estimation time)* the actual time required to estimate β_1 and β_2.

Table 1 summarizes the results of the simulations. As anticipated, the time and memory costs implied by the CMLE are rapidly increasing in T: with a typical desktop computer, we are

Table 1: **Results of Monte Carlo Exercise**

Data		**CMLE**			**2,500 Sequences**			**5,000 Sequences**			**10,000 Sequences**		
Generating Process		**Bias**	**RMSE**	**SD**	**Bias**	**RMSE**	**SD**	**Bias**	**RMSE**	**SD**	**Bias**	**RMSE**	**SD**
$T = 7$	$\beta_1 = -2$	-0.001	0.097	0.097	0.004	0.11	0.11	-0.005	0.109	0.109	-0.004	0.109	0.109
	$\beta_2 = 2$	-0.003	0.11	0.11	-0.025	0.121	0.121	0.002	0.118	0.118	0	0.119	0.119
Permutation time			0.235 min.			0.25 min.			0.303 min.			0.368 min.	
Estimation time			0.023 min.			0.016 min.			0.021 min.			0.026 min.	
$T = 8$	$\beta_1 = -2$	-0.002	0.104	0.104	-0.007	0.113	0.112	-0.005	0.112	0.112	-0.002	0.105	0.105
	$\beta_2 = 2$	-0.014	0.109	0.108	-0.016	0.119	0.117	-0.011	0.119	0.118	-0.016	0.113	0.112
Permutation time			0.855 min.			0.597 min.			0.848 min.			1.206 min.	
Estimation time			0.074 min.			0.03 min.			0.049 min.			0.078 min.	
$T = 9$	$\beta_1 = -2$	-0.003	0.065	0.065	-0.109	0.102	0.102	-0.012	0.093	0.093	0.005	0.078	0.078
	$\beta_2 = 2$	-0.001	0.086	0.086	0.071	0.129	0.128	0.007	0.107	0.107	-0.006	0.099	0.099
Permutation time			3.362 min.			1.13 min.			1.853 min.			3.037 min.	
Estimation time			0.307 min.			0.039 min.			0.067 min.			0.116 min.	
$T = 10$	$\beta_1 = -2$	-0.014	0.073	0.072	-0.061	0.209	0.179	-0.018	0.124	0.122	-0.013	0.098	0.097
	$\beta_2 = 2$	-0.007	0.081	0.081	0.043	0.199	0.185	0.004	0.145	0.145	0.003	0.117	0.117
Permutation time			12.928 min.			1.689 min.			3.028 min.			5.427 min.	
Estimation time			5.684 min.			0.043 min.			0.075 min.			0.132 min.	
$T = 11$	$\beta_1 = -2$				-0.061	0.268	0.261	-0.004	0.153	0.153	-0.01	0.136	0.135
	$\beta_2 = 2$		OUT OF		0.043	0.252	0.248	-0.004	0.172	0.172	0.016	0.162	0.161
Permutation time			MEMORY			2.145 min.			4.055 min.			7.664 min.	
Estimation time						0.047 min.			0.084 min.			0.15 min.	
$T = 12$	$\beta_1 = -2$				-0.183	0.496	0.462	-0.123	0.315	0.29	-0.098	0.217	0.193
	$\beta_2 = 2$		OUT OF		0.165	0.544	0.519	0.134	0.356	0.33	0.112	0.288	0.266
Permutation time			MEMORY			2.534 min.			4.924 min.			9.568 min.	
Estimation time						0.049 min.			0.089 min.			0.162 min.	

Note: For each value of T, results are averaged over a hundred simulations with $n = 1,000$. For each estimator, the average computation time required for estimation by our desktop machine (Intel® Core™, i3-3220, 3.3 GHz, 8Gb RAM) is divided into two components: *(permutation time)* the amount of time required to enumerate all the addends in the denominator of the respective multinomial logit formula and, given the denominator, *(estimation time)* the actual time required to estimate β_1 and β_2.

unable to estimate the model with $T > 10$, since the computer runs out of memory.[3] By contrast, the proposed method is always able to estimate β_1 and β_2. Moreover, whenever we can compute the CMLE, our method only requires a small fraction of the CMLE time.

The original motivation of McFadden (1978) was to reduce what we call here estimation time. First, Table 1 confirms that McFadden (1978)'s method works well in that respect. Furthermore, in the context of the FE logit model, the proposed method is also effective in reducing the time required to compute the denominator of the multinomial logit probability. As T increases, the number of possible permutations of the observed sequences of choices rises exponentially, and so does the time required to enumerate them all when computing the CMLE. By contrast, the proposed method allows to cap the number of permutations to be enumerated. This enables to significantly reduce the permutation time in smaller examples, and to estimate the model at all in larger applications. Note that for any fixed number of permutations, the permutation time still increases with T. This happens because the observed sequences of choices become longer and each of their permutations takes more time to be executed.

The computational advantages of the proposed method do not come for free. As expected given Proposition 1, the CMLE is more precise than our estimator. Moreover, and probably intuitively, the larger the random set of permutations, the more accurate is our estimator.

Collectively, these results suggest that one should use the CMLE whenever possible but, otherwise, one should rely on the method proposed in this paper using as many sampled sequences as possible, so as to limit the efficiency loss.

[3] In these cases individuals had multinomial logit probabilities with millions of addends in the denominator.

References

[1] T. Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[2] E. B. Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32:pp. 283–301, 1970.

[3] M. Arellano and B. Honoré. Panel data models: some recent developments. *Handbook of econometrics*, 5:3229–3296, 2001.

[4] B. Baltagi. *Econometric analysis of panel data*. John Wiley & Sons, 2005.

[5] G. Chamberlain. Analysis of covariance with qualitative data. *Review of Economic Studies*, 47(1):225–238, 1980.

[6] W. Greene. *Econometric analysis*. Prentice Hall, 2011.

[7] D. McFadden. Modelling the choice of residential location. *Institute of Transportation Studies, University of California*, pages 75–96, 1978.

[8] K. Pforr. femlogit-implementation of the multinomial logit model with fixed effects. *Stata Journal*, 14(4):847–862, 2014.

[9] G. Rasch. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine*, pages 321–333, 1961.

[10] K. Train, D. McFadden, and M. Ben-Akiva. The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *RAND Journal of Economics*, 38(4):109–123, 1987.