# Consistent and unbiased random coefficients Logit estimator under sampling of alternatives

Qian Wang, Marcus Sundberg, Anders Karlström*

*KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden*

Random utility models (RUM) are widely applied in studying discrete choices, such as travel demand. When the choice set is large, as in destination choice and route choice, etc., it becomes impractical to consider all alternatives in the full choice set, and the computational time becomes prohibitive in estimation. Among other solutions, many researchers choose to sample alternatives, which can introduce a sampling bias in the maximum likelihood estimator. McFadden (1978) theoretically proved that in a multinomial logit model, sampling bias is canceled if alternatives are randomly sampled from a uniform distribution. Bias correction is necessary when using other sampling protocols, importance sampling for instance (see Ben-Akiva and Lerman (1985), Ben-Akiva and Bowman (1998) and Frejinger et al. (2009)). Later Guevara and Ben-Akiva (2013) extended the neat theoretical result of consistency, asymptotic normality and efficiency to multivariate extreme value (MEV) models[i] with sampling of alternatives. In many applications, the Mixed multinomial logit model (MXL) is preferable since the MEV family can neither manage random taste variation, nor handle panel data with correlation over time. Typically, the MXL models are estimated using a maximum simulated likelihood (MSL) estimator (Lerman and Manski (1981)).

However, to the best of our knowledge, there is no method to obtain consistent MXL estimates with sampling of alternatives. Bias correction via the log-likelihood function in the context of MXL models, with sampling of alternatives, is extremely difficult. The numerical effect of sampling of alternatives in MXL estimation has been analyzed. McConnell and Tseng (1999) claimed that sampling of alternatives in random parameters logit has no substantial effect on the parameter estimates. Nerella and R.Bhat (2004) concluded that the MSL estimator can acquire good numerical performance when sampling at least a quarter of the full choice set. A new sampling protocol by Lemp and Kockelman (2012) can gain significant computational efficiency, but sampling of 25% alternatives is still recommended. Additionally, the MSL estimator can not promise unbiased and consistent MXL estimates unless a sufficient number of random draws are simulated for each observation.

*Corresponding author: amail@kth.se; qianw@kth.se

[i] Also, and previously, known as generalized extreme value (GEV) models.

There is great interest to enable a consistent estimator for the MXL model with sampling of alternatives. We propose an unbiased and consistent estimator based on the principle of indirect inference (II) for the MXL model. We choose an MXL model with sampling of alternatives as an auxiliary model. Then we estimate it on both the observed data and pseudo-data which are simulated by the MXL model based on the full choice set (see Smith (1993) and Gouriroux et al. (1993)). The parameters of the full choice set model are calibrated so that the values of auxiliary parameters estimated on the corresponding pseudo-data and the observed data are as close as possible.

Consider an MXL model based on the full choice set with random coefficient $\beta_n$ following a distribution $f(\beta_n|\theta)$, and we are interested in estimating the structural parameter $\theta$. Commonly, the MSL estimator is used for MXL estimation. The MSL estimate on the full choice set, $\hat{\theta}_{MSL1}$, is a consistent and unbiased estimate if the number of random draws for each observation increases faster than the square root of the number of observations. When sampling subsets uniformly (Nerella and R.Bhat (2004)), the approximated MSL estimate $\hat{\theta}_{MSL_2}$ is biased.

Based on II, we choose an MXL model with choice set sampling as the auxiliary model. The auxiliary parameter estimate $\hat{\gamma}$ is estimated on the observed data with choice set sampling. Meanwhile, draw pseudo-values of the structural parameter $\theta_m$ $(m = 1, ..., M)$ from some domain, and simulate data sets. For each simulated data set, the corresponding auxiliary parameter estimate $\tilde{\gamma}_m(\theta_m)$ is estimated with choice set sampling. In line with the "smoothing" idea for discrete data Gouriroux et al. (1993), a smooth *binding function* $\tilde{\gamma}(\theta)$ is estimated given $M$ pairs of $\theta_m, \tilde{\gamma}_m(\theta_m)$, by local regression or ordinary least squares. Finally, provided the estimated *binding function* $\tilde{\gamma}(\theta)$, we are able to find the indirectly inferred estimate $\hat{\theta}_{II}$ of the true parameter $\theta$.

With the bias correction through the *binding function* $\tilde{\gamma}(\theta)$, $\hat{\theta}_{II}$ is theoretically a consistent and asymptotically normal MXL estimator. Two numerical experiments are designed to examine the ability of $\hat{\theta}_{II}$: sampling bias correction with decreasing number of samplings of alternatives; and simulation bias correction with fewer number of simulation draws. We prepare synthetic data of $J = 200$ alternatives for each of $N = 5000$ independent choices. Bias of estimators, $\hat{\theta}_{MSL_1}$, $\hat{\theta}_{MSL_2}$ and $\hat{\theta}_{II}$, are measured by $E[\hat{\theta}] - \theta$, where $E[\hat{\theta}]$ is approximated by the average value over 20 runs with different random seeds. Besides, the Root Mean Square Error (RMSE) of estimates are computed, measuring both bias and variance. Results are displayed in Table.1 and Table.2. As expected, $\hat{\theta}_{MSL_1}$ is unbiased when estimated with 200 Halton draws, and simulation bias become substantial when estimated with only five draws. Sampling biases and simulation biases are found in $\hat{\theta}_{MSL_2}$, and sampling biases grow with decreasing sampling size. In contrast, $\hat{\theta}_{II}$ has neither sampling bias nor simulation bias. $\hat{\theta}_{II}$ is unbiased estimator with choice set sampling and fewer draws. Bias due to sampling of alternatives and simulation is corrected by the II estimator. But the variance of $\hat{\theta}_{II}$ increases when using fewer draws and fewer alternatives.

| Bias($\hat{\theta}$) | | $\hat{\theta}_{MSL_1}$ | $\hat{\theta}_{MSL_2}$ | | | | $\hat{\theta}_{II}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Choice set size | | 200 | 50 | 25 | 10 | 5 | 50 | 25 | 10 | 5 |
| 200 draws | $\mu_1$ | 0.01 | -0.27 | -0.34 | -0.50 | -0.82 | -0.00 | -0.00 | -0.01 | -0.05 |
| | $\mu_2$ | 0.00 | -0.09 | -0.10 | -0.11 | -0.19 | -0.00 | -0.01 | -0.01 | -0.04 |
| | $\sigma_1$ | 0.00 | 0.23 | 0.27 | 0.39 | 0.62 | 0.01 | 0.01 | 0.01 | 0.04 |
| | $\sigma_2$ | 0.00 | 0.16 | 0.17 | 0.17 | 0.28 | 0.00 | 0.02 | 0.00 | 0.03 |
| 5 draws | $\mu_1$ | 0.74 | 0.72 | 0.76 | 0.80 | 0.86 | 0.01 | 0.04 | 0.02 | -0.02 |
| | $\mu_2$ | 0.35 | 0.37 | 0.40 | 0.46 | 0.52 | 0.00 | 0.01 | 0.00 | -0.02 |
| | $\sigma_1$ | -0.29 | -0.23 | -0.28 | -0.35 | -0.44 | 0.04 | 0.01 | 0.03 | 0.05 |
| | $\sigma_2$ | -0.26 | -0.22 | -0.28 | -0.40 | -0.49 | 0.02 | 0.01 | 0.02 | 0.05 |

Table 1: Bias of estimators

| RMSE($\hat{\theta}$) | | $\hat{\theta}_{MSL_1}$ | $\hat{\theta}_{MSL_2}$ | | | | $\hat{\theta}_{II}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Choice set size | | 200 | 50 | 25 | 10 | 5 | 50 | 25 | 10 | 5 |
| 200 draws | $\mu_1$ | 0.03 | 0.28 | 0.34 | 0.51 | 0.82 | 0.05 | 0.05 | 0.06 | 0.09 |
| | $\mu_2$ | 0.02 | 0.09 | 0.10 | 0.11 | 0.20 | 0.02 | 0.03 | 0.03 | 0.06 |
| | $\sigma_1$ | 0.03 | 0.23 | 0.28 | 0.39 | 0.63 | 0.05 | 0.05 | 0.06 | 0.08 |
| | $\sigma_2$ | 0.02 | 0.16 | 0.17 | 0.18 | 0.29 | 0.03 | 0.04 | 0.04 | 0.07 |
| 5 draws | $\mu_1$ | 0.74 | 0.72 | 0.76 | 0.80 | 0.86 | 0.09 | 0.10 | 0.14 | 0.20 |
| | $\mu_2$ | 0.35 | 0.37 | 0.40 | 0.46 | 0.52 | 0.04 | 0.05 | 0.09 | 0.13 |
| | $\sigma_1$ | 0.30 | 0.24 | 0.29 | 0.35 | 0.45 | 0.09 | 0.08 | 0.12 | 0.16 |
| | $\sigma_2$ | 0.26 | 0.23 | 0.28 | 0.40 | 0.49 | 0.05 | 0.05 | 0.10 | 0.14 |

Table 2: RMSE of estimators

# References

Ben-Akiva, M. and Bowman, J. (1998). Integration of an activity-based model system and a residential location model. *Urban Studies*, 35(7):11311153.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press, Cambridge, Massachusetts.

Frejinger, E., Bierlaire, M., and Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, 43(10):984–994.

Gouriroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118.

Guevara, C. A. and Ben-Akiva, M. E. (2013). Sampling of alternatives in multivariate extreme value (mev) models. *Transportation Research Part B: Methodological*, 48(0):31 – 52.

Lemp, J. D. and Kockelman, K. M. (2012). Strategic sampling for large choice sets in estimation and application. *Transportation Research Part A: Policy and Practice*, 46(3):602 – 613.

Lerman, S. R. and Manski, C. F. (1981). On the use of simulated frequencies to approximate choice probabilities. In Manski, C. F. and McFadden, D. L., editors, *tructural Analysis of Discrete Data and Econometric Applications*, pages 305–319. Cambridge: The MIT Press.

McConnell, K. E. and Tseng, W.-C. (1999). Some preliminary evidence on sampling of alternatives with the random parameters logit. *Marine Resource Economics*, 14(4):317–332.

McFadden, D. (1978). Modeling the choice of residential location. In A. Karlquist, e. a., editor, *Spatial Interaction Theory and Planning Models*, pages 531–552. Amsterdam, North-Holland Publishing Company.

Nerella, S. and R.Bhat, C. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models. *Transportation Research Record*, 1894:11–19.

Smith, A. A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8, Supplement: Special issue on Econometric Inference using Simulaiton Techniques:S63–S84.