Position Papers of the Federated Conference on Computer Science and Information Systems pp. 79–82
DOI: 10.15439/2019F354
ISSN 2300-5963 ACSIS, Vol. 19

Maximum Simulated Likelihood: Don't Stop 'Til You Get Enough?

Christopher Schrey*†, Tobias Schäffer*‡, Carsta Militzer-Horstmann*† and Nils Kossack*
*WIG2 Institute, Leipzig
†Leipzig University
‡Martin Luther University of Halle-Wittenberg

***Abstract*—Maximum simulated likelihood estimation can be employed in empirical health economics, amongst others, to tackle issues concerning endogenous treatment effects. While theory suggests that maximum simulated likelihood estimation is asymptotically consistent, efficient and equivalent to the maximum likelihood estimator when both the number of simulation draws *S* and sample size N $\rightarrow \infty$ and $\sqrt{N}/\ S \rightarrow 0$, there is no guidance on how large of an *S* to choose and even theory suggests to experiment. This piece of research reviews strategies of health economists that aim at dealing with this issue. Most pieces of applied research rely on experimentation until numerical stability is achieved, while some employ Monte-Carlo techniques to justify their choice of *S*. A more formal test was suggested, but seemed not to be employed yet. This lack of guidance induces a research problem that needs to be properly addressed.**

I. Introduction

ENDOGENEITY in non-linear regression models arising through self-selection into treatment is a problem very often encountered in, but not limited to, health economics. One prominent situation in which health economists face problems with endogenous regressors is when the effect of health insurance status on healthcare utilisation, such as visits to the doctor, is estimated. Different types of health insurance plans, such as deductibles and co-payments, are offered, to incentivise an economical utilisation of scarce medical resources. As participation in such insurance plans is non-random, selection bias complicates studies in which the effect of endogenous treatment (here: insurance choice) is estimated on a healthcare utilisation outcome, such as number of visits to the doctor. Self-selection occurs when optimising individuals possessing unobservable characteristics, such as awareness of future health states or risk preferences, select health insurance plans accordingly [1]. The same unobservable characteristics that affect insurance choice might then also affect future healthcare utilisation, thus leading to potential unobserved correlation between insurance choice decision and decision to consume health services [2].

One way of addressing this endogeneity issue is the endogenous treatment regression model by [1], that utilises a latent factor structure. Latent factors are incorporated into the treatment and outcome equations, thus allowing to make a distinction between selection on unobservables and selection on observables. As these latent factors cannot be observed, regular maximum likelihood estimation is not feasible. Yet, when assuming a distribution of the latent factors (e.g. standard normal), simulation-based estimation, i.e. *maximum simulated likelihood* (MSL) estimation, remains possible [3].

The properties of MSL estimates crucially depend on the number of simulation draws S (per observation) and sample size N. Given that $S, N \rightarrow \infty$ and $\sqrt{N}/S \rightarrow 0$, MSL is asymptotically consistent, efficient and equivalent to the maximum likelihood estimator [3]. Yet, this ratio does not provide guidance on what S should be for given N, it only describes the properties of MSL as N increases [4]. Consequently, researchers face a non-trivial problem when deciding how large of an S, given sample size N, to choose. On the one hand, the MSL-approach is computationally burdensome, as it makes extensive use of simulation techniques [5]. Generating random numbers requires a matrix of size $S \times N$, as there are S random draws for each of the N observations. As increasing N will also necessitate an increase in S, this will ultimately lead to non-trivial memory consumption, that is to say, to potentially prohibitively high computational cost [6]. On the other hand, consistency of the estimator requires $S, N \rightarrow \infty$ and $\sqrt{N}/S \rightarrow 0$. Some [7] recommend using S as large as computational reasonable, while others rely on experimentation with different sizes of S to achieve numerical stability of the estimator as their guide [8], [9], [4], [7]. Thus, the researcher needs to find a suitable trade-off between precision (favouring infinitely large S) on the one hand and computational cost (favouring fixed S) on the other hand. This lack of guidance with respect to choice of an appropriate amount of simulation draws imposes a serious challenge for applied research in two ways. First, having results at hands, the question to the researcher remains, whether or not a sufficient amount of simulation draws was employed [6]. Similarly, and equivalently important, the researcher's choice regarding S remains untraceable to the scientific community.

Consequently, the research problem of the underlying piece of research (work-in-progress) is to find guidance with respect to the choice of an appropriate amount of simulation draws to be employed in maximum simulated likelihood estimation within the endogenous treatment regression context. Establishing such guidance will be beneficial to the research community as it will make MSL-procedure more traceable. As a starting point in establishing such guidance, strategies of dealing with this issue in applied research are presented and discussed

within a preliminary literature review. Firstly, however, the MSL-approach and its peculiarities will be explained in more detail, before the relevant literature will be summarised. After the research problem is derived from the literature review, the intended future work to tackle this problem will be discussed.

II. Maximum Simulated Likelihood Estimation

To deal with the endogeneity of treatment (insurance choice) on healthcare utilisation, [1] introduced a latent factor structure into the treatment and outcome equations to account for selection on unobservables. These latent factors enter both treatment and outcome equation to allow for idiosyncratic influences on insurance status choice to affect healthcare utilisation, thus making a distinction between selection on unobservables and selection on observables possible [1]. These latent factors serve as proxies for unobservable characteristics and are interpreted as unobserved heterogeneity. Endogeneity arises, as the same latent factors, i.e. unobservable characteristics, determining insurance choice also affect the healthcare utilisation decision. As they cannot be observed, problems in estimation arise, as no closed-form solution to the respective integral exists [1]. Yet, when making assumptions with respect to the underlying distribution of the unobservable characteristics (e.g. standard normal distribution), maximum simulation likelihood estimation remains feasible. Here, simulation depends on the fact that integrating over a density is simply a form of averaging [10]. Thus, the effect of the unobservable latent factors can be integrated out, resulting in an unbiased (with respect to self-selection) estimate of the treatment effect. Among several possible ways of taking endogeneity into account (e.g. IV-approach, Difference-in-Difference, two-stage residual inclusion) the maximum simulated likelihood-procedure is the only approach that sufficiently addresses both endogeneity of treatment and non-linearity (count data) in the outcome [11].

When outcome y's (e.g. number of doctor visits) conditional density $f(y|\mathbf{x},\ \theta)$, where $\mathbf{x}$ may be individual i's observable characteristics, θ the parameters to be estimated and $\mathbf{u}$ unobservable characteristics, involves such an intractable integral, such that

$$f(y_i|\mathbf{x}_i,\theta) = \int h(y_i|\mathbf{x}_i,\theta,\mathbf{u}_i)g(\mathbf{u}_i)d(\mathbf{u}_i) \tag{1}$$

requires estimation [6]. Accordingly, one needs to approximate the intractable integral $h(y_i|\mathbf{x}_i,\theta,\mathbf{u}_i)$ with a subsimulator $\tilde{f}(y_i|\mathbf{x}_i,\theta,\mathbf{u}_i^s)$. To do so, S ($S = 1,\ldots,S$) random draws from the assumed distribution of $\mathbf{u}$ are drawn into the subsimulator. The average over S (denoted by $\mathbf{u}_{iS}$) of these subsimulators then provides the simulator $\hat{f}(y_i|\mathbf{x}_i,\theta,\mathbf{u}_{iS})$ such that [6]

$$\underbrace{\hat{f}(y_i|\mathbf{x}_i,\theta,\mathbf{u}_{iS})}_{Simulator} = \frac{1}{S}\sum_{s=1}^{S}\underbrace{\tilde{f}(y_i|\mathbf{x}_i,\theta,\mathbf{u}_i^s)}_{Subsimulator}. \tag{2}$$

While the usual maximum likelihood estimator maximises the log-likelihood $\ln \mathrm{L}_N(\theta) = \sum_{i=1}^{N}\ln f(y_i|\mathbf{x}_i,\ \theta)$, the maximum simulated likelihood estimator instead maximises the log-likelihood based on the simulated estimation of the density [6]

$$\ln \hat{L}_N(\theta) = \sum_{i=1}^{N}\ln \underbrace{\hat{f}(y_i|\mathbf{x}_i,\theta,\mathbf{u}_{iS})}_{Simulator}. \tag{3}$$

As the estimator is simulated rather than calculated precisely, simulation error is introduced [10]. This simulation error can be decomposed into three sources of error: simulation chatter, simulation noise and simulation bias [10]. Simulation chatter occurs, when different random draws are used at each likelihood iteration [10], [5]. While simulation chatter might render (simulated) likelihood maximisation infeasible, it can be easily encountered by using the same simulation draws per observation [10], [5]. Thus, simulation chatter does neither depend on the choice of S nor N. Deviations from each simulated value of its expectation lead to simulation noise [10]. As simulation noise cancels out over observations, it decreases with N, even if S is fixed [10]. Simulation bias occurs as the MSL simulator $\ln \hat{f}$ is biased for $\ln f$, even if the simulator $\hat{f}$ is unbiased for f, as a consequence of taking the natural logarithm [6]. An asymptotic bias-adjusted MSL-estimator, that makes use of a bias-adjusted log-likelihood function, is suggested by [3]. As this bias-adjustment assumes bias to be small, [6] adds, that the usefulness of this bias-reduction may vary from case to case, as the small bias-assumption may not always hold. After all, for the simulation bias to disappear, S and $N \to \infty$, while S must increase faster than $\sqrt{N}$, such that $\sqrt{N}/S \to 0$ [3], [10]. If the latter condition is met, MSL is asymptotically normal, efficient and equivalent to maximum likelihood estimation [3], [10]. However, this ratio does not state what S should be for given N, it only describes the properties of the MSL estimator as N increases [4].

III. Literature Review

Whether or not one has done a sufficient amount of simulations to tackle the simulation error issues is a difficult question to answer [6]. As no empirical guidance exists, theory suggest to experiment with different sizes of S until numerical stability of the estimator is achieved [4], [8], [6]. Consequently, [12], [1], [13], [14] report to have relied on such experimentation to find an appropriate S. From experience, [1] suggest, that estimating MSL in the context of endogenous treatment requires "considerably more" simulation draws than models involving seemingly unrelated errors. Yet, [1] do not further elaborate what this might imply in practical terms. Still, [1] state that their choice of S is based on other empirical studies that use MSL. Table I provides a summary of the choice of S, with respect to N, of empirical studies that employed the MSL-approach in the context of endogenous regressors in health economics. While there is no clear guidance on the quantity of simulation draws, consensus seems to exist regarding their quality. Quasi-random draws, such as the Halton-sequence, rather than pseudo-random draws,

TABLE I: OVERVIEW OF CHOICES REGARDING S IN APPLIED HEALTH ECONOMICS RESEARCH.

Reference	N	S	$\frac{\sqrt{N}}{S}$	Random variates
[14]	2,467	1,600	0.031	Halton
[1]	8,129	2,000	0.045	Halton
[12]	26,514	1,000	0.162	Halton
[7]	5,033	400	0.177	Halton
[16]	4,406	300	0.221	Antithetic
[13]	109,349	200	1.653	Halton
[21]	Did not report S			Halton
[11]	Did not report S			Halton
[2]	Did neither report S nor type of random variates			

are considered to greatly reduce the number of simulation draws required for a given amount of precision [10], [4], [5]. Halton-draws are more evenly distributed than pseudo-random draws, while also displaying lower variance, as they are negatively correlated [10]. Even though Halton-draws are rather deterministic than random, [8] add, that when it comes to simulation techniques, the randomness of draws is not as important as their uniform coverage over the domain of integration. Their desirable properties made the Halton-sequence the quasi-random variate of choice, as displayed in Table I. Also, consensus exists that MSL-estimation is, as suggested by theory [4], [6], a rather computationally burdensome approach, as also explicitly stated in several pieces of applied research [13], [12], [7], [15]. Even more so, [12] report to have used less simulation draws than desired (due to having relatively large N) to ensure convergence of their model, while [13] even report that one of their models did not converge after four days of CPU time. One notable deviation of the experimentation-strategy within applied research seems to be the approach by [16] who conducted a Monte-Carlo experiment prior to their empirical study to justify their choice regarding the number of simulation draws. Also, [17] are able to quantify simulation noise and simulation bias of their MSL approach, as their econometric model also offers an analytical solution, to which they can compare their MSL results. Similarly, [18] are able to quantify simulation error, as within their theoretical approach, they employ a simulated dataset, for which the true parameters are known. A different, more formal approach in choosing S is suggested by [19], who describes a diagnostic test, constructed from a Wald test statistic, that captures the magnitude of simulation bias and could be used to compute an amount of S that will produce an acceptable estimator. Even though some pieces of literature [6], [4] point out to this formal test, it was not employed in the reviewed literature. Yet, e.g. [20] employ this diagnostic test in the context of MSL-based dynamic probit models.

IV. DISCUSSION AND OUTLOOK

Within applied research, the question whether one has used enough simulation draws remains challenging. As no clear guidance exists, researchers rely on experimentation with different values of S to achieve numerical stability of the estimator. This procedure does not necessarily satisfy the reader's interest in transparency and traceability with respect to empirical research. One exception [16] in applied health economics research employed a Monte-Carlo study as a benchmark for their subsequent choice of S. Even though translating conclusions drawn from self-designed experimental data to "exogenous" real-world data might similarly raise doubts, it at least seems to be a somewhat more traceable way of justifying one's choice of S. Also, having an analytical, thus correct, solution, as a benchmark, might very much answer the question, whether or not one has used enough simulation draws. Yet, not having an analytical solution remains the motivation to employ MSL in the first place.

This piece of (emerging) research is tackling this overall lack of guidance with respect to choosing S by producing an empirical benchmark within the endogenous treatment context. This benchmark should not be solely based on self-designed experimental data, such as [16], as this type of data might not reflect real-world complexity, that is known to make the MSL-approach burdensome [6], [7]. Nevertheless, such a Monte-Carlo study might clearly be supplemental to reach this overall goal. Also, employing an econometric model on real-world data, for which an analytical solution is possible, does neither seem to be a desirable option, even though the true parameters would be known and could thus serve as a reference. Yet, as already stated, the lack of an analytical solution is the motivation to employ MSL in the first place.

In order to exploit real-world data, while also knowing the true parameter (with respect to selection on unobservables), the Oregon Health Insurance Experiment [22] will be employed. In 2008, within this experiment, a limited amount of Medicaid insurance coverage was allocated randomly to low-income individuals, while also recording healthcare utilisation behaviour of lottery winners and losers afterwards. Randomly assigned Medicaid insurance (i.e. treatment) can be considered exogenous with respect to healthcare utilisation. Thus, employing [1] endogenous treatment regression model, the effect of selection on unobservables on healthcare utilisation will be hypothesised to be zero (due to randomisation). Making use of the Oregon experiment will thus be beneficial to illustrate MSL-convergence behaviour in the context of endogenous treatment regression. These results can ultimately serve as a guide for other health economists to choose an appropriate amount of S, as the simulation error can be estimated quite well, as the true estimates (with respect to selection on unobservables) are known. Thus, MSL-convergence behaviour can be explicitly illustrated for different values of S. Additionally, the formal Wald-based test, suggested by [19], will be employed, to formally support (or reject) the findings. Also, the literature dealing with the MSL-procedure, especially in the realm of health economics, will be reviewed more extensively and intensively, with respect to strategies of choosing an appropriate S, while also promising alternatives to the MSL-approach, as suggested by [18], need to be studied closely. As a result, health econometricians will benefit from this ongoing piece of research, as it will provide them with some guidance whether or not they have chosen a sufficiently large S, when employing MSL estimation.

V. References

[1] P. Deb and P. K. Trivedi. "Specification and simulated likelihood estimation of a non–normal treatment–outcome model with selection: Application to health care utilization". In: *The Econometrics Journal* 9.2 (2006), pp. 307–331. DOI: 10.1111/j.1368-423X.2006.00187.x.

[2] D. Shane and P. K. Trivedi. "What drives differences in health care demand? The role of health insurance and selection bias". In: *Health, Econometrics and Data Group Working Papers* 12/09 (2012).

[3] C. Gouriéroux and A. Monfort. *Simulation-based econometric methods*. Oxford University Press, 1997. DOI: 10.1093/0198774753.001.0001.

[4] W. H. Greene. *Econometric analysis*. 6. ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

[5] A. Cameron and P. K. Trivedi. *Microeconometrics using Stata*. Rev. ed. A Stata Press publication. College Station, Tex.: Stata Press, 2010.

[6] A. Cameron and P. K. Trivedi. *Microeconometrics: Methods and applications*. New York, NY: Cambridge University Press, 2005.

[7] P. Deb and P. K. Trivedi. "Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment". In: *Stata Journal* 6.2 (2006), pp. 246–255.

[8] D. M. Drukker and R. Gates. "Generating Halton sequences using Mata". In: *Stata Journal* 6.2 (2006), 214–228(15).

[9] D. M. Drukker. "Maximum simulated likelihood: Introduction to a special issue". In: *Stata Journal* 6.2 (2006), 153–155(3).

[10] K. E. Train. *Discrete choice methods with simulation*. Cambridge: Cambridge University Press, 2009. DOI: 10.1017/CBO9780511805271.

[11] M. M. Garrido, P. Deb, J. F. Burgess, and J. D. Penrod. "Choosing models for health care cost analyses: issues of nonlinearity and endogeneity". In: *Health services research* 47.6 (2012), pp. 2377–2397. DOI: 10.1111/j.1475-6773.2012.01414.x.

[12] P. Deb, C. Li, P. K. Trivedi, and D. M. Zimmer. "The effect of managed care on use of health care services: results from two contemporaneous household surveys". In: *Health economics* 15.7 (2006), pp. 743–760. DOI: 10.1002/hec.1096.

[13] V. Atella and P. Deb. "Are primary care physicians, public and private sector specialists substitutes or complements? Evidence from a simultaneous equations model for count data". In: *Journal of health economics* 27.3 (2008), pp. 770–785. DOI: 10.1016/j.jhealeco.2007.10.006.

[14] M. Bratti and A. Miranda. "Endogenous treatment effects for count data models with endogenous participation or sample selection". In: *Health economics* 20.9 (2011), pp. 1090–1109. DOI: 10.1002/hec.1764.

[15] A. Geraci, D. Fabbri, and C. Monfardini. "Testing exogeneity of multinomial regressors in count data models: Does two-stage residual inclusion work?" In: *Journal of Econometric Methods* 7.1 (2018), p. 313. DOI: 10.1515/jem-2014-0019.

[16] M. K. Munkin and P. K. Trivedi. "Simulated maximum likelihood estimation of multivariate mixed–Poisson regression models, with application". In: *The Econometrics Journal* 2.1 (1999), pp. 29–48. DOI: 10.1111/1368-423X.00019.

[17] V. Atella and A. Holly. "Disentangling adverse selection, moral hazard and supply induced demand: An empirical analysis of the demand for healthcare services". In: *SSRN Electronic Journal* (2016). DOI: 10.2139/ssrn.2801679.

[18] C. R. Bhat, C. Varin, and N. Ferdous. "A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model". In: *Maximum simulated likelihood methods and applications*. Ed. by R. C. Hill and W. H. Greene. Vol. 26. Advances in Econometrics. Bingley, UK: Emerald, 2010, pp. 65–106. DOI: 10.1108/S0731-9053(2010)0000026007.

[19] V. A. Hajivassiliou. "Some practical issues in maximum simulated likelihood". In: *Simulation-based Inference in Econometrics*. Ed. by R. Mariano, T. Schuermann, and M. J. Weeks. Cambridge: Cambridge University Press, 2000, pp. 71–99.

[20] P. Contoyannis, A. M. Jones, and N. Rice. "Simulation-based inference in dynamic panel probit models: An application to health". In: *Empirical Economics* 29.1 (2004), pp. 49–77. DOI: 10.1007/s00181-003-0189-x.

[21] M. B. Buntin, C. H. Colla, P. Deb, N. Sood, and J. J. Escarce. "Medicare spending and outcomes after postacute care for stroke and hip fracture". In: *Medical care* 48.9 (2010), pp. 776–784. DOI: 10.1097/MLR.0b013e3181e359df.

[22] A. Finkelstein et al. *The Oregon health insurance experiment: Evidence from the first year*. Cambridge, MA, 2011. DOI: 10.3386/w17190.