Simulation error in maximum likelihood estimation of discrete choice models

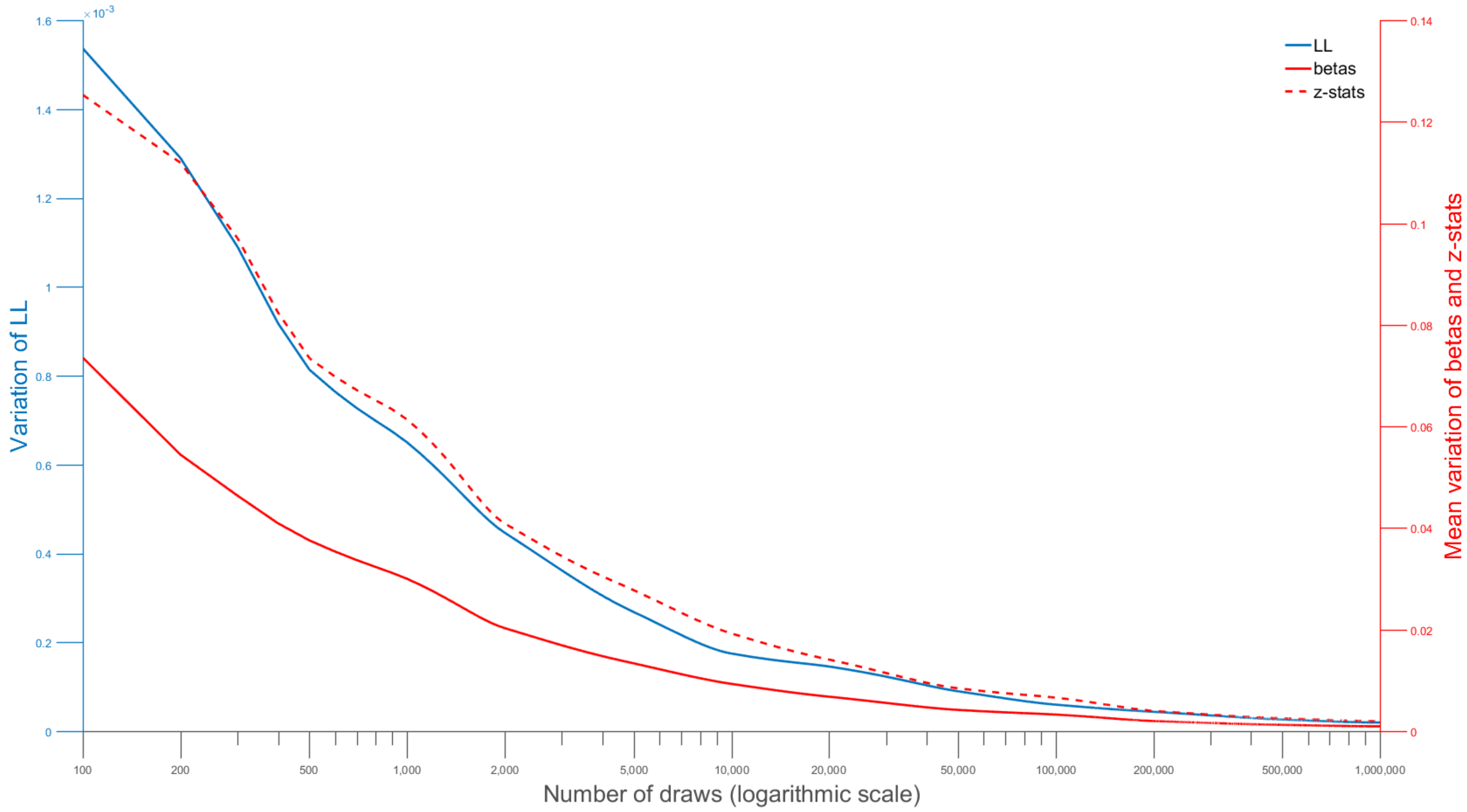
Mikołaj Czajkowski, Wiktor Budziński

czaj.org

Simulation error

- Discrete choice data
  - Mixed (random parameters) logit models
    - Estimation via simulated maximum likelihood method
- Simulating the value of the log-likelihood function
  - Necessarily associated with simulation error
    - Depends on the number and type of draws
- A different set of draws = somewhat different estimation results

Simulation error vs. the number of draws

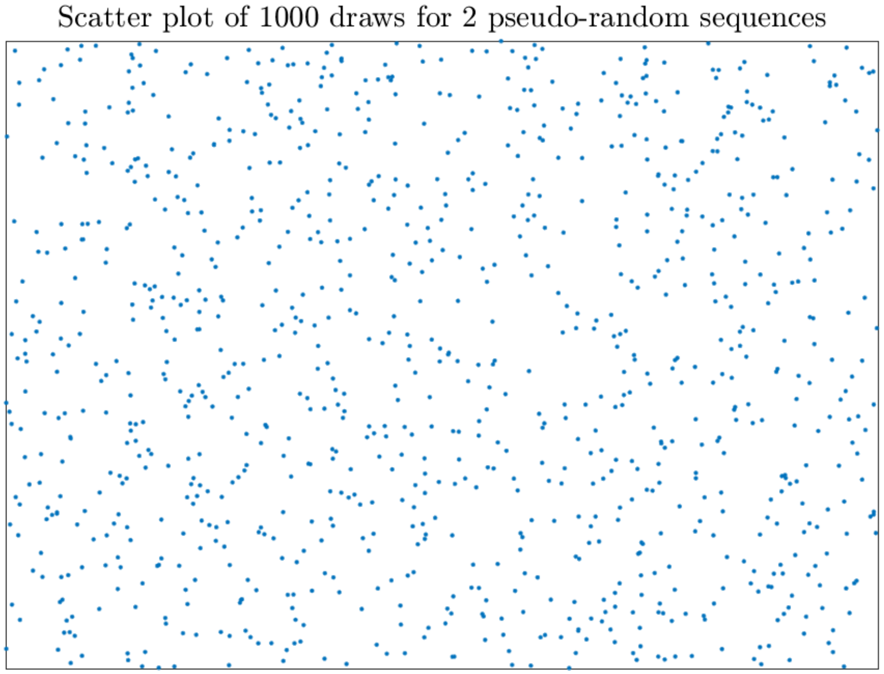


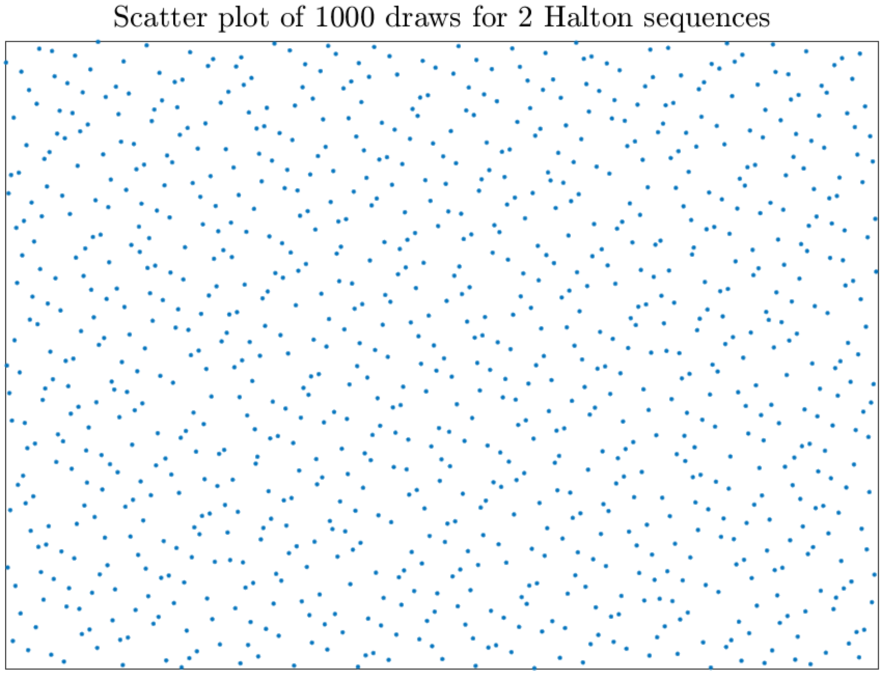
Quasi Monte Carlo methods

- Quasi Monte Carlo methods reduce simulation-driven variation
  - Halton sequence (Train 2000, Bhat 2001),
  - Sobol sequence (Garrido 2003)
  - Randomized (t,m,s)-nets (Sándor and Train 2004)
  - Modified Latin Hypercube (Hess, Train and Polak 2006)
  - Lattice rules (Munger *et al.* 2012)
  - Generalized antithetic draws with double base shuffling (Sidharthan and Srinivasan 2010)

  - Shuffling, scrambling sequences (Bhat 2003, Hess, Polak and Daly 2003, Hess and Polak 2003, Wang and Kockelman 2008)

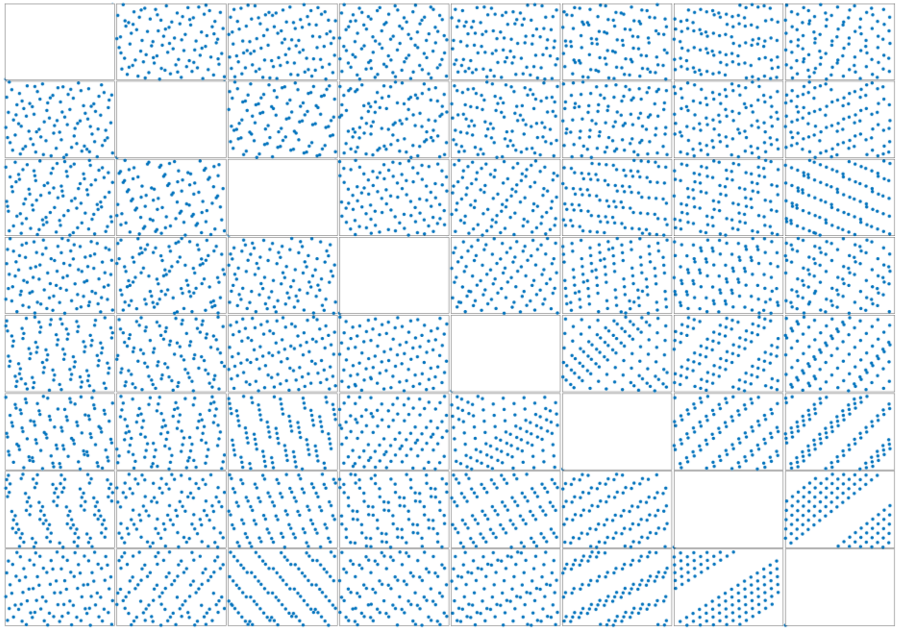
Pseudo-random vs. Halton sequence



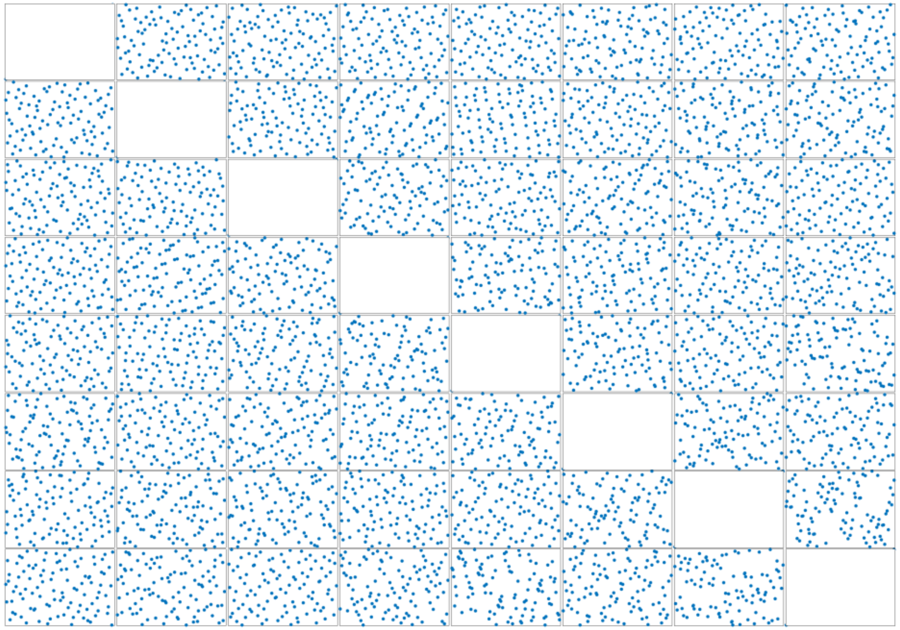


Halton vs. scrambled Halton sequence

Scatter plot matrix of 100 draws for 8 Halton sequences



Scatter plot matrix of 100 draws for 8 scrambled Halton sequences



Gaps in existing evidence

- What is the extent of the simulation bias resulting from using different numbers of different types of draws in various conditions (datasets)?
  - Shortcoming of the existing studies:
    - Low numbers of QMC draws (≤ 200)
    - Low number of repetitions for each type and number of draws (≤ 10)
    - Results likely to depend on the number of observations (individuals, choice tasks per individual)
  - Examples of 100 Halton draws leading to smaller bias than 1,000 pseudo-random draws (e.g., Bhat, 2001) have led some to actually use very few draws for simulations
- Using too few draws can lead to spurious convergence of models that are theoretically or empirically unidentified (Chiou and Walker 2007)
- Our study aims at filling these gaps

Design of our simulation study – Choice task setting and explanatory variables

Explanatory variables (choice attributes)	Assumed parameter distribution	Possible values of the explanatory variables		
		Alternative 1 (status quo / opt-out)	Alternative 2	Alternative 3
X_1 (alternative specific constant)	$N(-1.0, 0.5)$	$X_1 = 1$	$X_1 = 0$	$X_1 = 0$
X_2 (dummy)	$N(1.0, 0.5)$	$X_2 = 0$	$X_2 \in \{0,1\}$	$X_2 \in \{0,1\}$
X_3 (dummy)	$N(1.0, 0.5)$	$X_3 = 0$	$X_3 \in \{0,1\}$	$X_3 \in \{0,1\}$
X_4 (dummy)	$N(1.0, 0.5)$	$X_4 = 0$	$X_4 \in \{0,1\}$	$X_4 \in \{0,1\}$
X_5 (discrete)	$N(-1.0, 0.5)$	$X_5 = 0$	$X_5 \in \{1,2,3,4\}$	$X_5 \in \{1,2,3,4\}$

Design of our simulation study – Choice task setting and explanatory variables

	Draws		Datasets		
Repetitions	Types of draws	Number of draws	Number of choice tasks per individual	Number of individuals	Experimental designs
1,000	*pseudo-random* *MLHS* *Halton* *Sobol*	100 200 500 1,000 2,000 5,000 10,000 20,000* 50,000* 100,000* 200,000* 500,000* 1,000,000*	4 8 12	400 800 1,200	OOD-design MNL-design MXL-design

*Selected settings only.

Methodology of comparisons

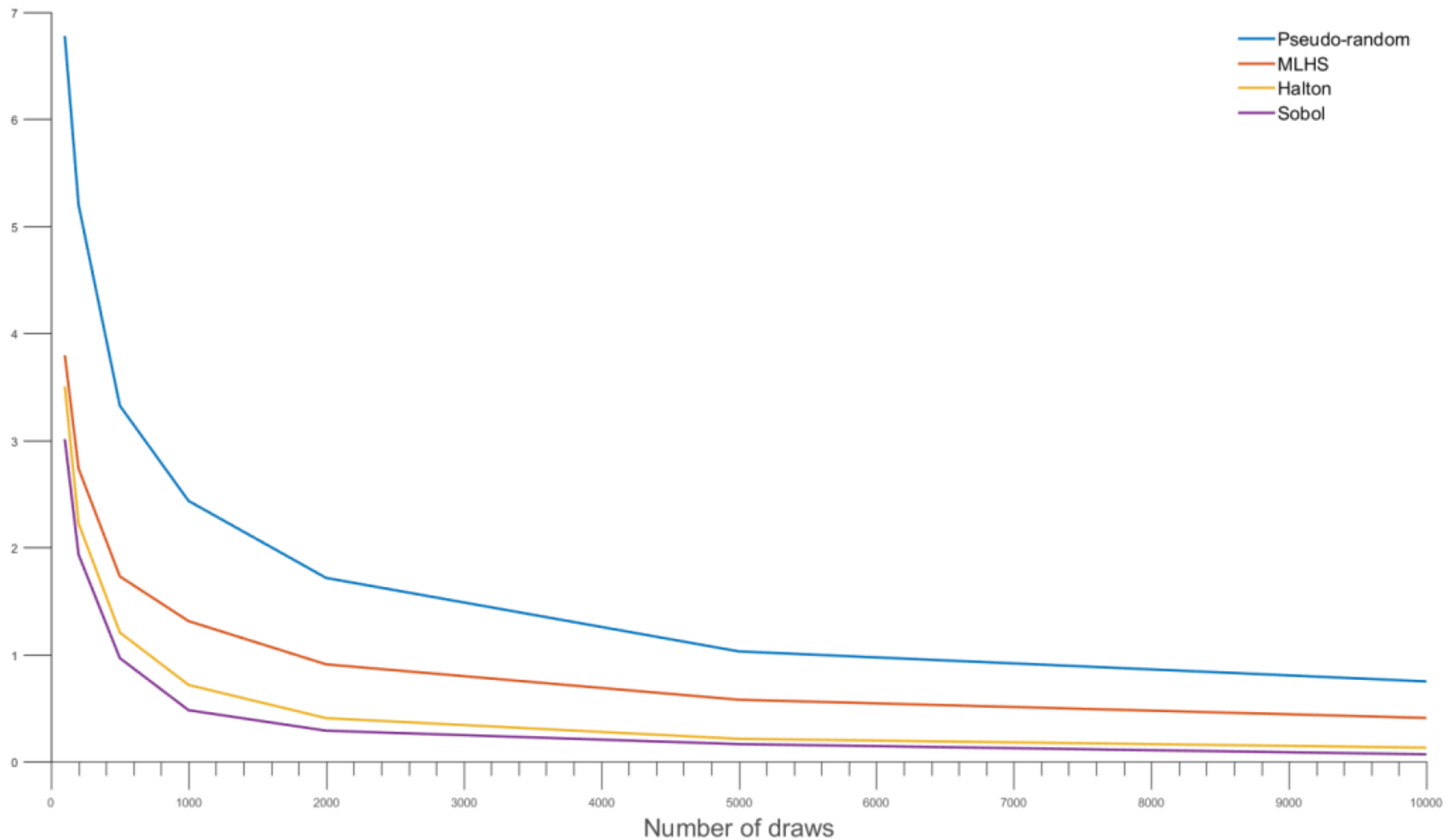
- We need a measure that takes expected values into account but also penalizes variance
  - For typical equality tests – the larger the variance, the more difficult to reject the equality hypothesis
- Testing equivalence instead of equality
  - Reverse the null and the alternative hypotheses
  - Test if the absolute difference is higher than a priori defined ‘acceptable’ level
- Minimum Tolerance Level (MTL)
  - What is the minimum ‘acceptable’ difference that allows to conclude that two values are equivalent at the required significance level
  - How many draws of type A are required, so that with 95% probability the difference in LL / estimates / s.e. / z-stats is not going to be statistically different than:
    - The critical value of the LR-test
    - If the model was estimated using *n* draws of type B

Example – using MTL for the values of the LL function

- Re-estimating the model using a different set of draws is likely to result in a somewhat different value of the LL function
- If LL is used for inference (e.g., LR-test), it is possible to conclude that one specification is superior to another only because one was more ‘lucky’ with the draws
- By using the MTL approach we are able to evaluate the probability of such an outcome
  - Assume α = 0.05, the interpretation of $MTL_{0.05}$ is that with 95% probability using a different set of draws would not cause the difference in LL values to be higher than $MTL_{0.05}$
  - We can provide recommendations for the minimum number of draws that would result in $MTL_{0.05}$ lower than the specified level
    - E.g., the critical value of the LR-test – probability of erroneously concluding that one model is preferred to another (because of simulation error) is lower than a desired significance level, e.g., 0.05

Results – relative performance of types of draws

– Example: $MTL_{0.05}$ of LL for MXL-design, 400 x 4:



Percentage of times each type of draws resulted in the lowest simulation error ($MTL_{0.05}$) for the log-likelihood function value

Number of draws used	*Pseudo-random*	*MLHS*	*Halton*	*Sobol*
100	0.00%	0.00%	18.52%	81.48%
200	0.00%	0.00%	29.63%	70.37%
500	0.00%	0.00%	22.22%	77.78%
1,000	0.00%	0.00%	25.93%	74.07%
2,000	0.00%	0.00%	0.00%	92.59%
5,000	0.00%	0.00%	11.11%	81.48%
10,000	3.70%	3.70%	0.00%	81.48%

Percentage of times each type of draws resulted in the lowest simulation error ($MTL_{0.05}$) for parameter estimates

Number of draws used	*Pseudo-random*	*MLHS*	*Halton*	*Sobol*
100	0.00%	0.37%	42.96%	56.67%
200	0.00%	0.00%	33.33%	66.67%
500	0.00%	0.00%	31.11%	68.89%
1,000	0.00%	0.00%	31.48%	68.52%
2,000	0.00%	0.00%	14.44%	78.15%
5,000	0.00%	0.00%	17.78%	74.81%
10,000	3.70%	3.70%	5.56%	75.93%

Percentage of times each type of draws resulted in the lowest simulation error ($MTL_{0.05}$) for z-stats

Number of draws used	*Pseudo-random*	*MLHS*	*Halton*	*Sobol*
100	0.00%	0.37%	48.15%	51.48%
200	0.74%	1.85%	34.07%	63.33%
500	0.37%	2.22%	32.22%	65.19%
1,000	0.74%	1.85%	26.67%	70.74%
2,000	0.00%	4.44%	22.59%	65.56%
5,000	3.70%	1.11%	19.26%	68.52%
10,000	3.70%	3.70%	5.19%	76.30%

Results – Sobol draws consistently perform best

– Percent of additional draws needed to achieve the same simulation error as Sobol draws:

	Pseudo-random	*MLHS*	*Halton*
LL	889% [776% - 1,020%]	305% [258% - 360%]	66% [47% - 87%]
Parameter estimates	361% [331% - 392%]	209% [189% - 232%]	48% [38% - 58%]
z-stats	347% [321% - 375%]	200% [182% - 219%]	51% [42% - 60%]

* Based on regression analysis

Results – regression results
Dependent variable: log(*MTL*)

	LL	Betas	Z stats
Cons.	3.4432*** (0.0693)	0.5144*** (0.0363)	2.7254*** (0.0316)
Type of draws: Pseudo-random	1.4637*** (0.0365)	0.8803*** (0.0185)	0.8366*** (0.0161)
Type of draws: MLHS	0.8939*** (0.0383)	0.6507*** (0.0195)	0.6140*** (0.0169)
Type of draws: Halton	0.3241*** (0.0384)	0.2261*** (0.0195)	0.2297*** (0.0170)
Design is: MXL	0.1803*** (0.0333)	-0.3372*** (0.0169)	-0.3736*** (0.0147)
Design is: OOD	0.0426 (0.0346)	-0.0124 (0.0176)	-0.1082*** (0.0153)
No. of CT is 8	0.6121*** (0.0323)	-0.4829*** (0.0164)	-0.0355** (0.0143)
No. of CT is 12	0.8894*** (0.0332)	-0.3424*** (0.0168)	0.2058*** (0.0146)
No. of individuals is 800	0.4287*** (0.0326)	-0.3001*** (0.0165)	0.1334*** (0.0144)
No. of individuals is 1200	0.6811*** (0.0329)	-0.4943*** (0.0167)	0.2605*** (0.0145)
Log of No. of draws	-0.6387*** (0.0076)	-0.5764*** (0.0038)	-0.5587*** (0.0033)
Parameter for mean		-1.4881*** (0.0136)	-1.4266*** (0.0118)
SQ		0.3477*** (0.0176)	0.1108*** (0.0153)
Cost		-0.7951*** (0.0176)	0.0302** (0.0153)
R^2	0.9346	0.8535	0.8704
N	783	7830	7830

Results – how many draws are ‘enough’?

- Using more draws is always better to using fewer draws
- How many are ‘enough’ depends on the desired precision level
- Log-likelihood:
  - Imagine you are comparing 2 specifications using LR-test (d.f. = 1)
  - Simulation error low enough to have 95% probability of not erroneously concluding that one model is better than the other
    - In other words, 95% of the times the (simulation driven) difference in LL must be lower than 1.9207 (at α = 0.05)
  - This is exactly what $MTL_{0.05}$ can be used for!

	400 x 4	800 x 4	1,200 x 4	400 x 8	800 x 8	1,200 x 8	400 x 12	800 x 12	1,200 x 12
p = 0.05	120	230	340	300	600	890	470	920	1,370
p = 0.01	300	575	850	750	1,500	2,225	1,175	2,300	3,425

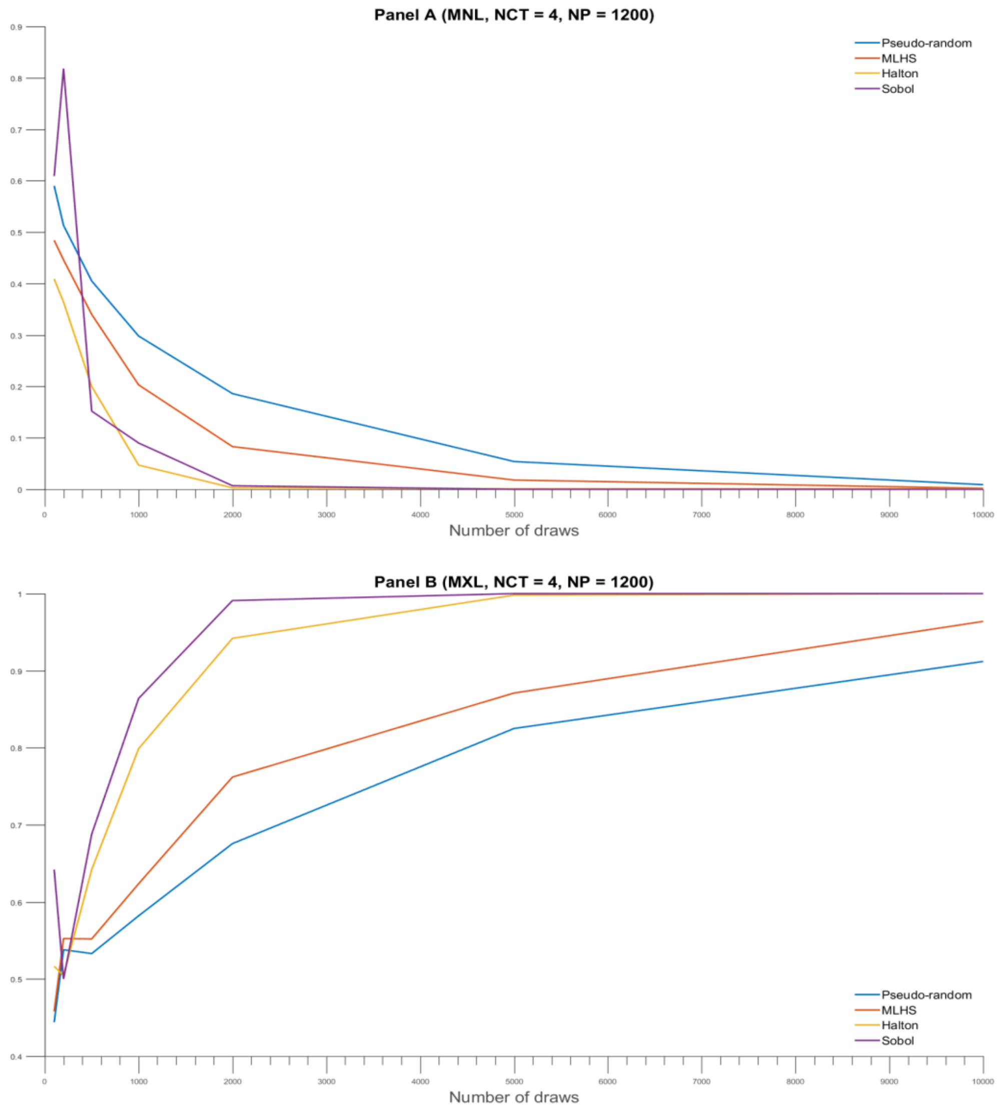
Results – how many draws are ‘enough’?

- Parameter estimates:
  - No absolute difference level
  - The numbers of draws required for 95% probability that the difference between parameter estimates :

	400 x 4	800 x 4	1,200 x 4	400 x 8	800 x 8	1,200 x 8	400 x 12	800 x 12	1,200 x 12
< 5%	2,050	1,220	870	890	530	380	1,130	670	480
< 1%	33,420	19,850	14,180	14,450	8,590	6,130	18,450	10,960	7,820

  - More draws required for standard deviations, ASC, dummies, fewer required for means, cost
  - Similar results for comparisons with models estimated using 1,000,000 draws

Using too few draws and identification problems – percentage of times z-statistics exceeded 1.96



“It must take ages to estimate models with so many draws!”

- Estimation time (1 iteration = LL function evaluation + gradient)
  - Data set: 400 respondents x 4 choice tasks
  - Intel E5-2687W @ 3.00 GHz (12-core) CPU (no GPU used!)
  - Efficient code implementation (Matlab, https://github.com/czaj/dce)

Number of draws	1,000	10,000	100,000	1,000,000
Iteration time	0.2 s	1 s	10 s	100 s

Summary and conclusions

- We investigate the performance of the 4 most commonly used types of draws for simulating log-likelihood in the mixed logit model setting
- We find Sobol draws consistently result in the lowest simulation error

Sobol draws recommended

- Conditional on our simulation setting, we find one needs more draws than typically used for ‘reliable’ estimation results

Use at least 1,000 (5%) or 10,000 (1%) draws

  - mean of the minimums; samples with fewer observations require fewer draws for precise LL and more draws for precise betas, and vice versa

- Evidence of erroneous inference on significance (both ways), if too few draws are used