Simulation error in maximum likelihood estimation of discrete choice models

Mikołaj Czajkowski, **Wiktor Budziński**

czaj.org

Simulation error

- Mixed (random parameters) logit models estimated using the simulated maximum likelihood method
  - Necessarily associated with simulation error
- A different set of draws = somewhat different estimation results

- What type of draws performs best?
- How many draws are „enough”?

Mixed logit model

- Utility function with preference heterogeneity

$$U_{ijt} = \mathbf{X}_{ijt}\beta_i + \varepsilon_{ijt}$$

- Conditional probability of the choice given by the logit formula:

$$P\left(y_{ijt} \mid \beta_i\right) = \frac{\exp\left(\mathbf{X}_{ijt}\beta_i\right)}{\sum_l \exp\left(\mathbf{X}_{ilt}\beta_i\right)}$$

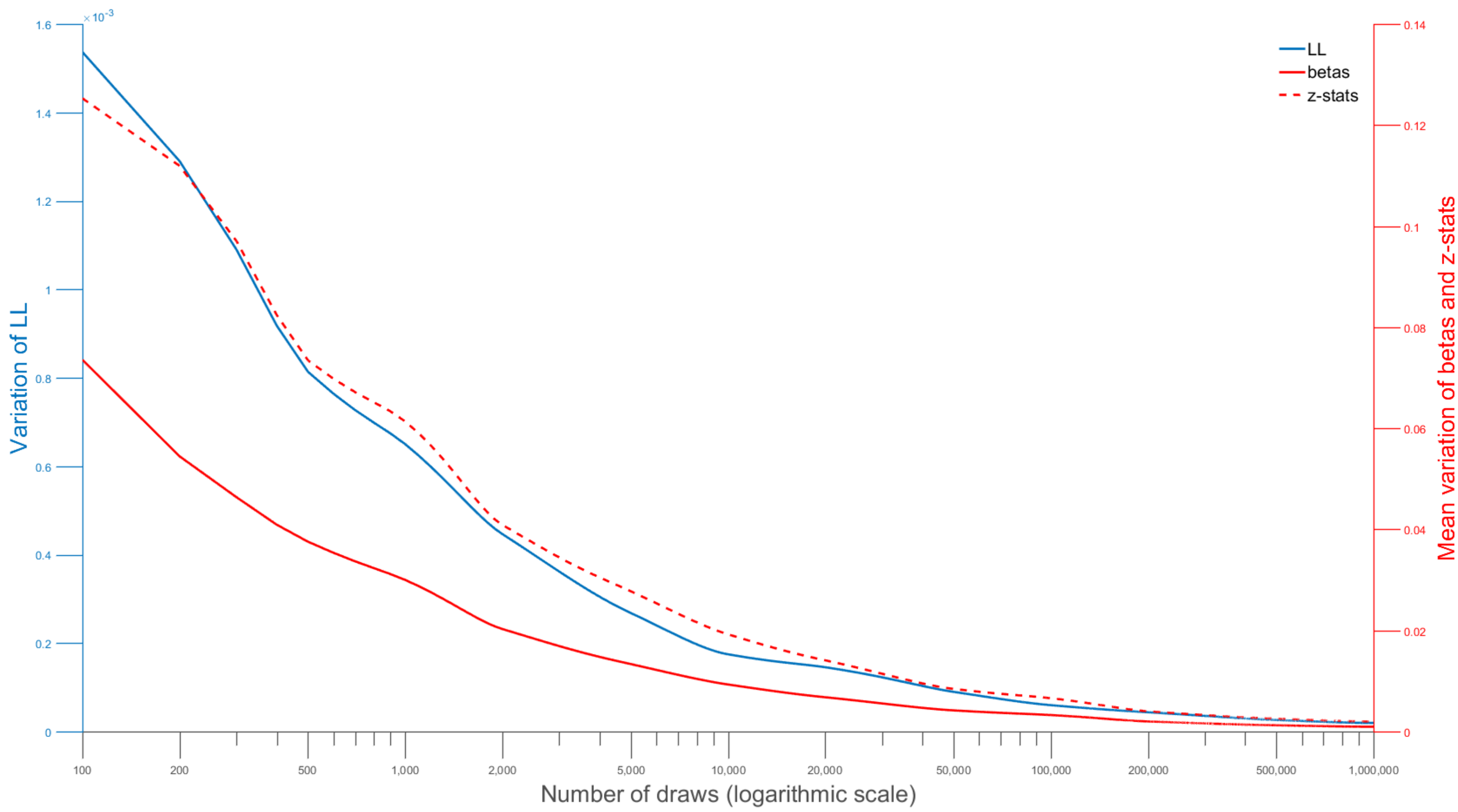
- Unconditional probability given by the integral:

$$P\left(\mathbf{y}_i\right) = \int \prod \left(P\left(y_{ijt} \mid \beta_i\right)^{y_{ijt}} \right) f\left(\beta_i \mid \Omega\right) d\beta_i$$

- Which can be approximated by

$$P\left(\mathbf{y}_i\right) \approx \frac{1}{R} \sum_{r=1}^{R} \prod \left(P\left(y_{ijt} \mid \beta_i^r\right)^{y_{ijt}} \right)$$

Simulation error vs. the number of draws



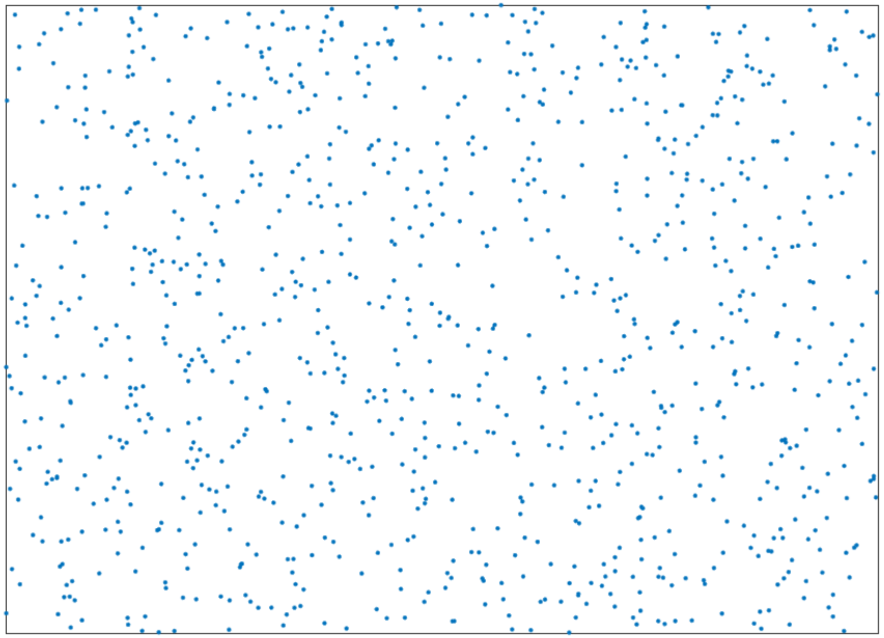
Quasi Monte Carlo methods

- 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)

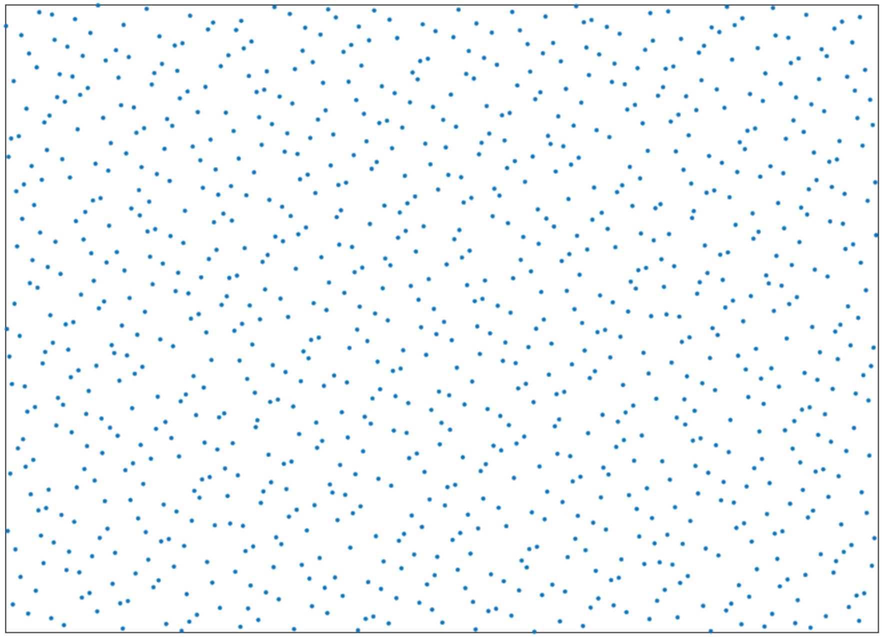
  - Shuffling, scrambling sequences (Bhat 2003, Hess, Polak and Daly 2003, Hess and Polak 2003, Wang and Kockelman 2008)

Pseudo-random vs. Halton sequence

Scatter plot of 1000 draws for 2 pseudo-random sequences

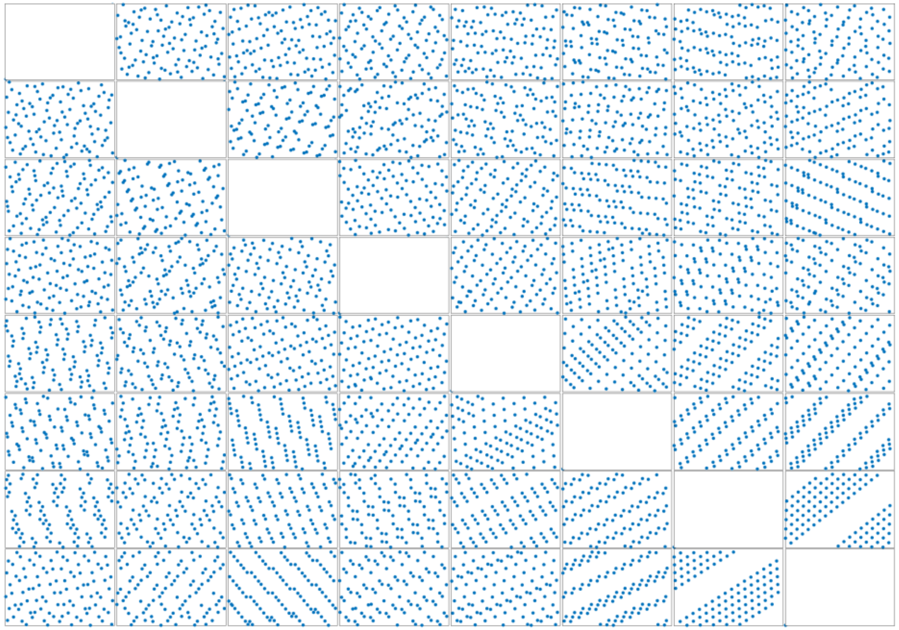


Scatter plot of 1000 draws for 2 Halton sequences

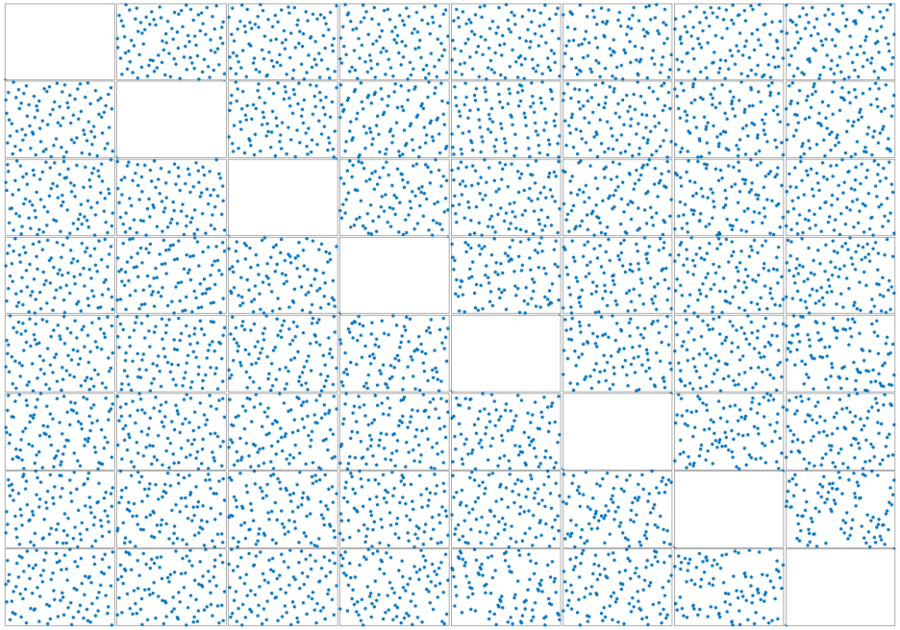


Halton vs. scrambled Halton sequence

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
- 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

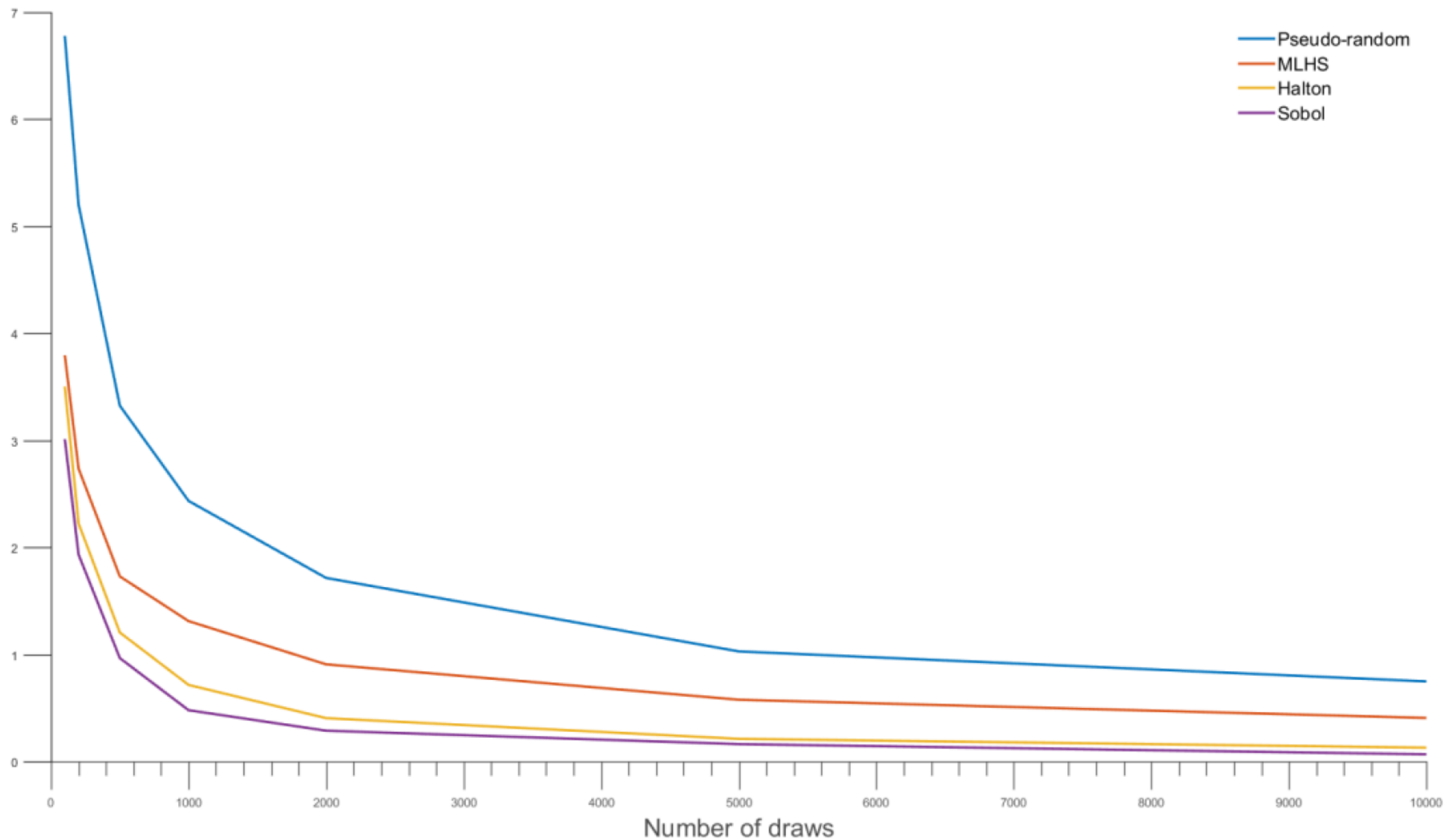
- We need a measure that takes expected values into account but also penalizes variance
  - For typical equality tests – the larger the variance, the more difficult to reject the equality hypothesis
- Testing equivalence instead of equality
  - Reverse the null and the alternative hypotheses
  - Test 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 distributions 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

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%	100.00%
5,000	0.00%	0.00%	14.81%	85.19%
10,000	0.00%	0.00%	0.00%	100.00%

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%	15.93%	84.07%
5,000	0.00%	0.00%	20.74%	79.26%
10,000	0.00%	0.00%	9.26%	90.74%

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

	Log-likelihood	**Parameter estimates**	**z-statistics**
Constant	2.8382*** (0.0817)	-0.9566*** (0.0425)	0.7334*** (0.0362)
log(number of draws)	-0.6338*** (0.0075)	-0.5786*** (0.0038)	-0.5638*** (0.0032)
Pseudo-random draws (*Sobol* used as a reference)	1.4568*** (0.0365)	0.8770*** (0.0186)	0.8360*** (0.0158)
MLHS draws (*Sobol* used as a reference)	0.9021*** (0.0382)	0.6495*** (0.0194)	0.6144*** (0.0166)
Halton draws (*Sobol* used as a reference)	0.3216*** (0.0382)	0.2173*** (0.0194)	0.2209*** (0.0166)
Number of choice tasks	0.1153*** (0.0041)	-0.0450*** (0.0021)	0.0230*** (0.0018)
Number of individuals (in thousands)	0.8695*** (0.0409)	-0.6377*** (0.0208)	0.2956*** (0.0177)
OOD-design (MXL-design used as a reference)	-0.1267*** (0.0329)	0.3125*** (0.0168)	0.2449*** (0.0143)
MNL-design (MXL-design used as a reference)	-0.1495*** (0.0329)	0.3224*** (0.0168)	0.3558*** (0.0143)
Standard deviations (Means used as a reference)		1.4735*** (0.0136)	1.4228*** (0.0116)
X_1 (alternative specific constant)		0.3610*** (0.0176)	0.1193*** (0.0150)
X_5 (discrete variable)		-0.7795*** (0.0176)	0.0373** (0.0150)
R^2	0.9299	0.8465	0.869
n (observations)	816	8160	8160

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

Simulation error – 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 95/99% probability of not erroneously concluding that one model is better than the other
    - In other words, 95/99% of the times the (simulation driven) difference in LL must be lower than 1.9207

	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
α = 0.05	173	277	444	375	602	965	814	1,306	2,095
α = 0.01	238	384	617	517	831	1,337	1,119	1,800	2896

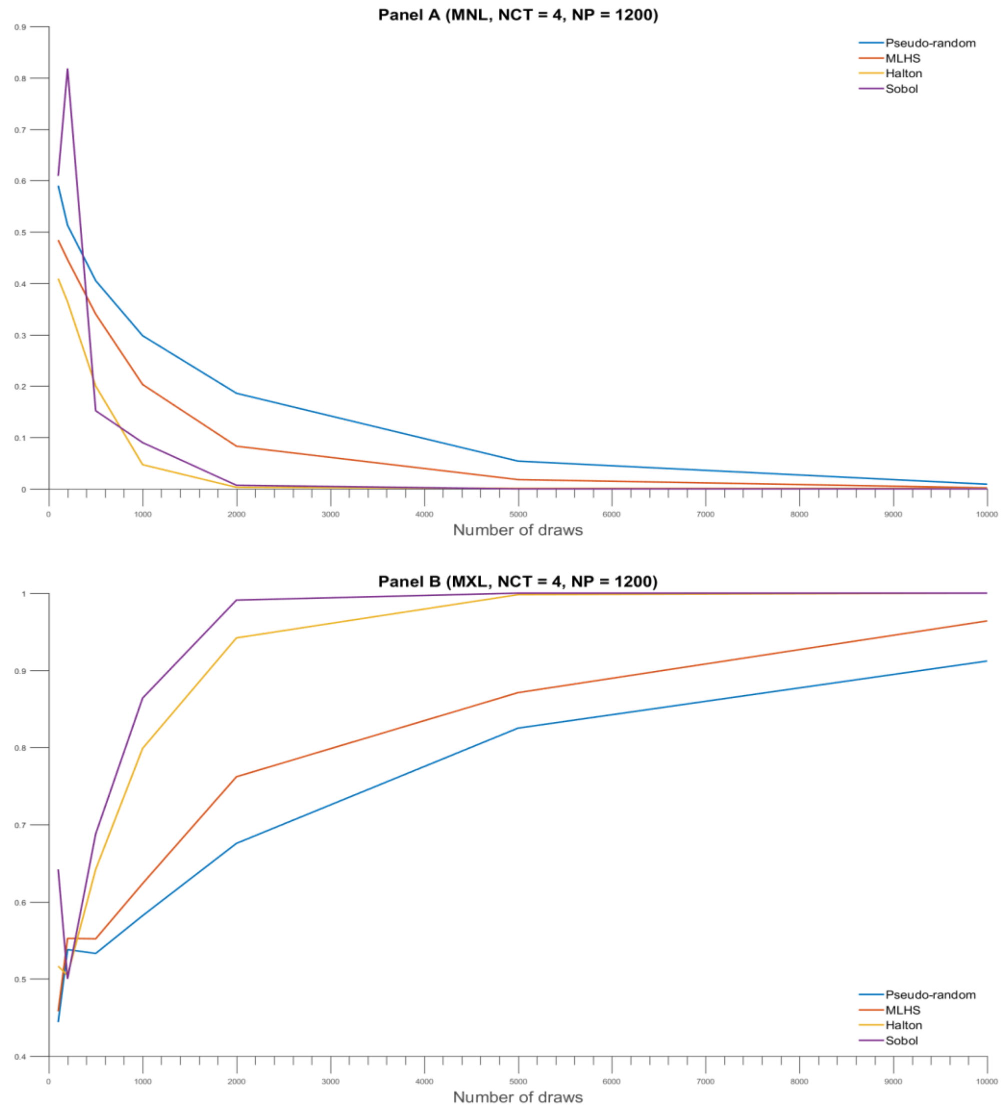
Simulation error – Results: how many draws are 'enough'?

- 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% (α = 0.05)	1,230	895	652	1,155	840	612	1,085	790	575
< 5% (α = 0.01)	1,802	1,312	956	1,686	1,228	894	1,578	1,149	837
< 1% (α = 0.05)	11,321	8,241	5,999	10,637	7,743	5,637	9,994	7,275	5,296
< 1% (α = 0.01)	16,569	12,066	8,787	15,506	11,292	8,224	14,511	10,568	7,696

  - 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

At least 1,000 draws (at 5%)

  - 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