Simulation error in maximum likelihood estimation of discrete choice models

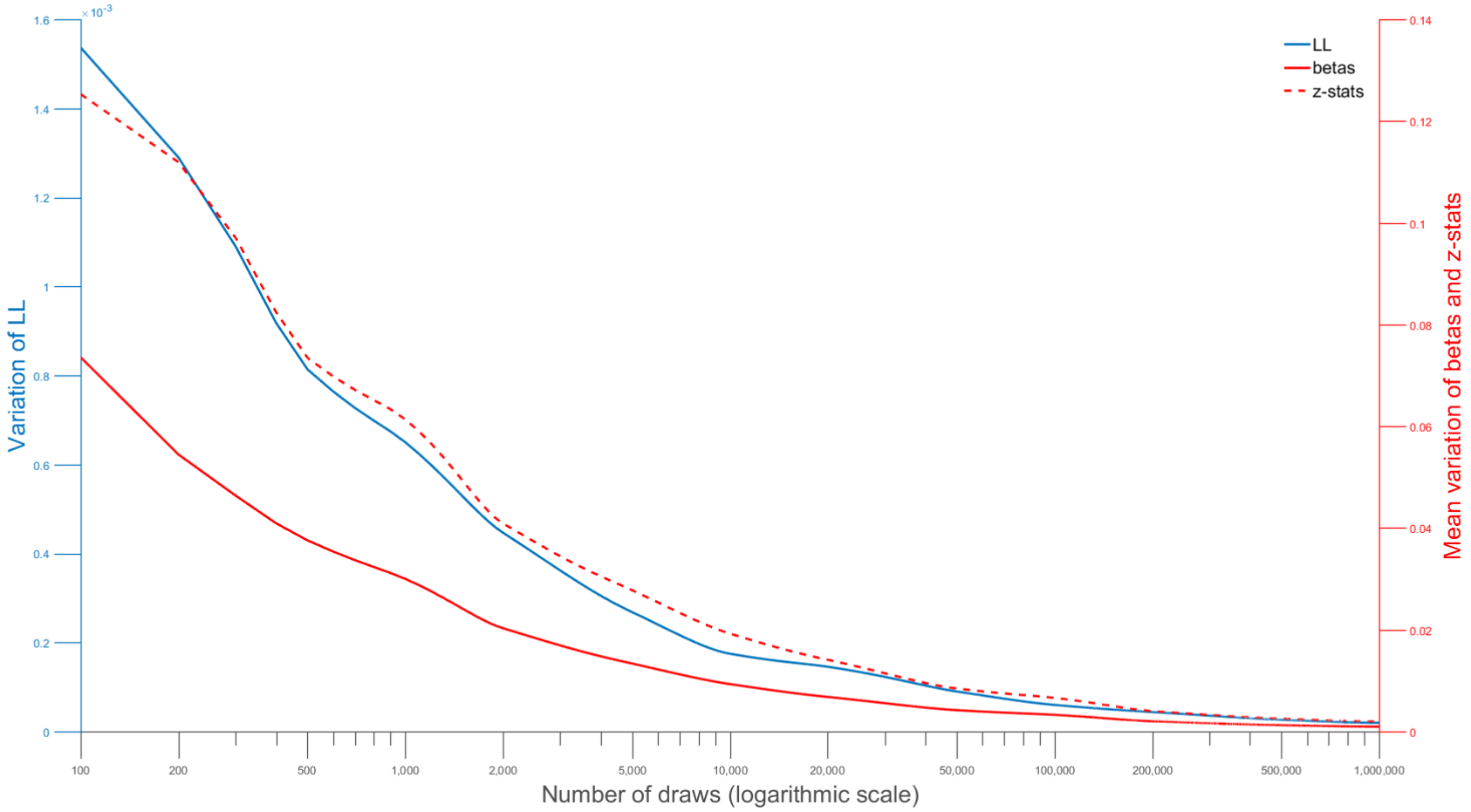
Mikołaj Czajkowski, Wiktor Budziński

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

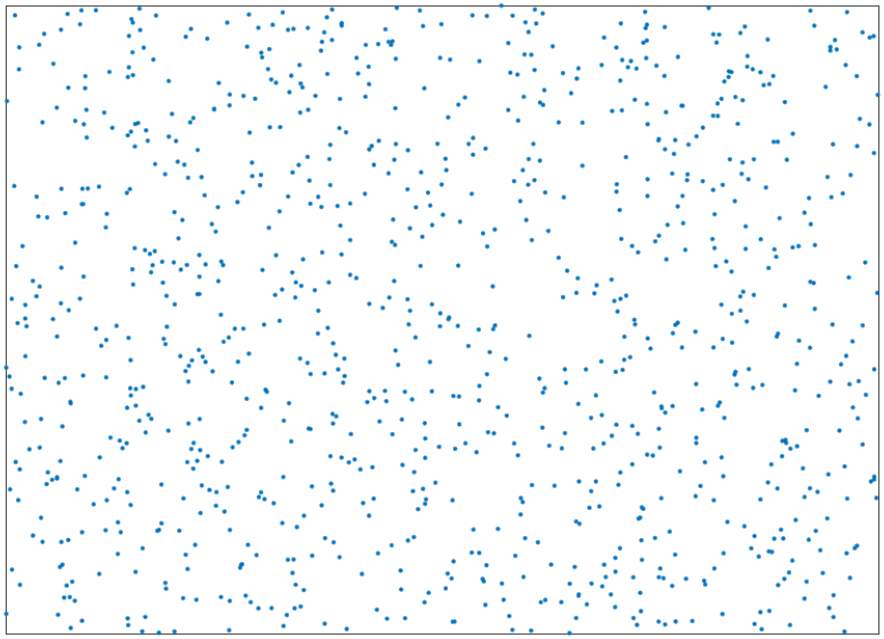


Quasi Monte Carlo methods

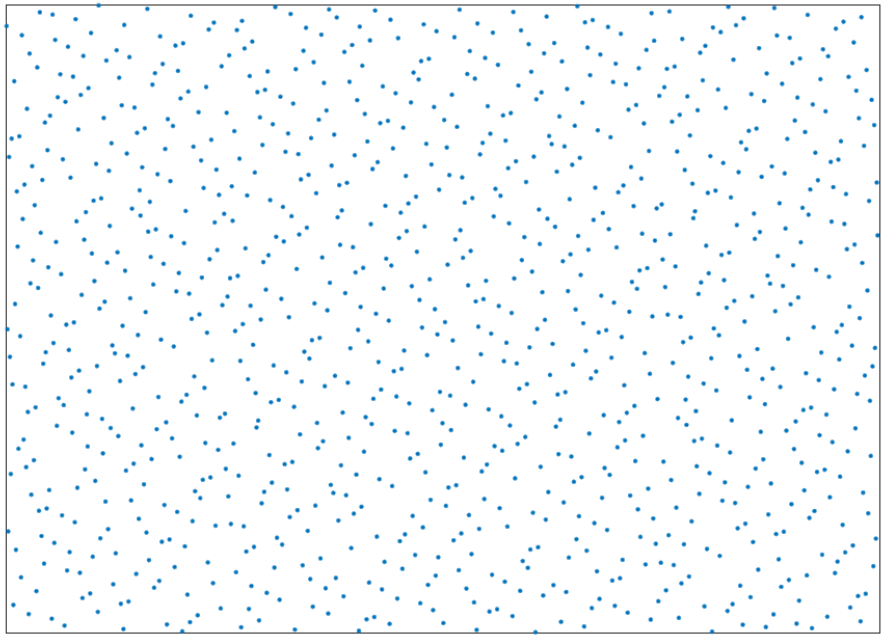
- Quasi Monte Carlo methods reduce simulation-driven variation
  - Halton sequence (Train 2000, Bhat 2001),
  - Sobol sequence (Garrido 2003)
  - Randomized (t,m,s)-nets (Sándor and Train 2004)
  - Modified Latin Hypercube (Hess, Train and Polak 2006)
  - Lattice rules (Munger *et al.* 2012)
  - Generalized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

Scatter plot of 1000 draws for 2 pseudo-random sequences

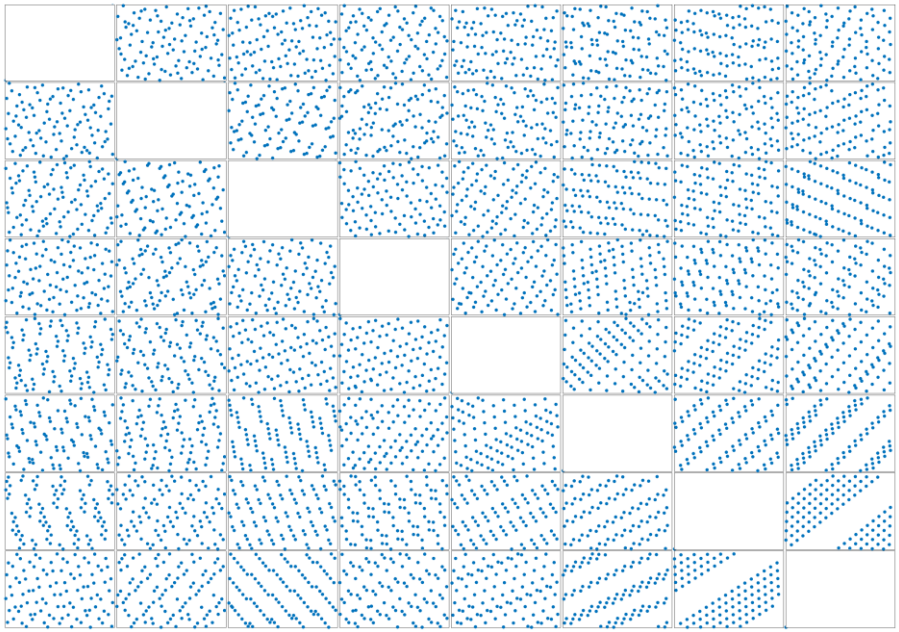


Scatter plot of 1000 draws for 2 Halton sequences

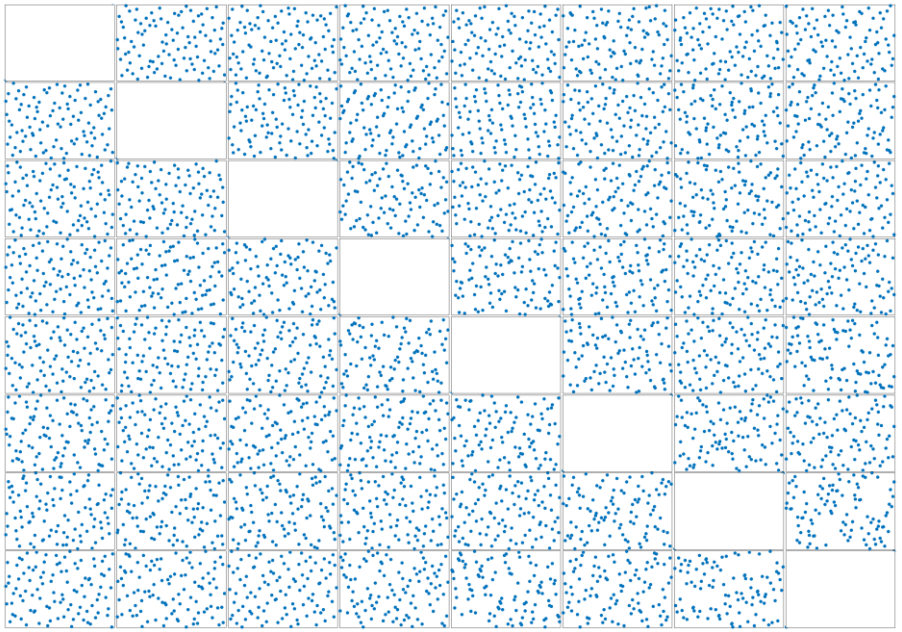


Halton vs. scrambled Halton sequence

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
- 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

Repetitions	Draws		Datasets		
	Types of draws	Number of draws	Number of choice tasks per individual	Number of individuals	Experimental designs
100	*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

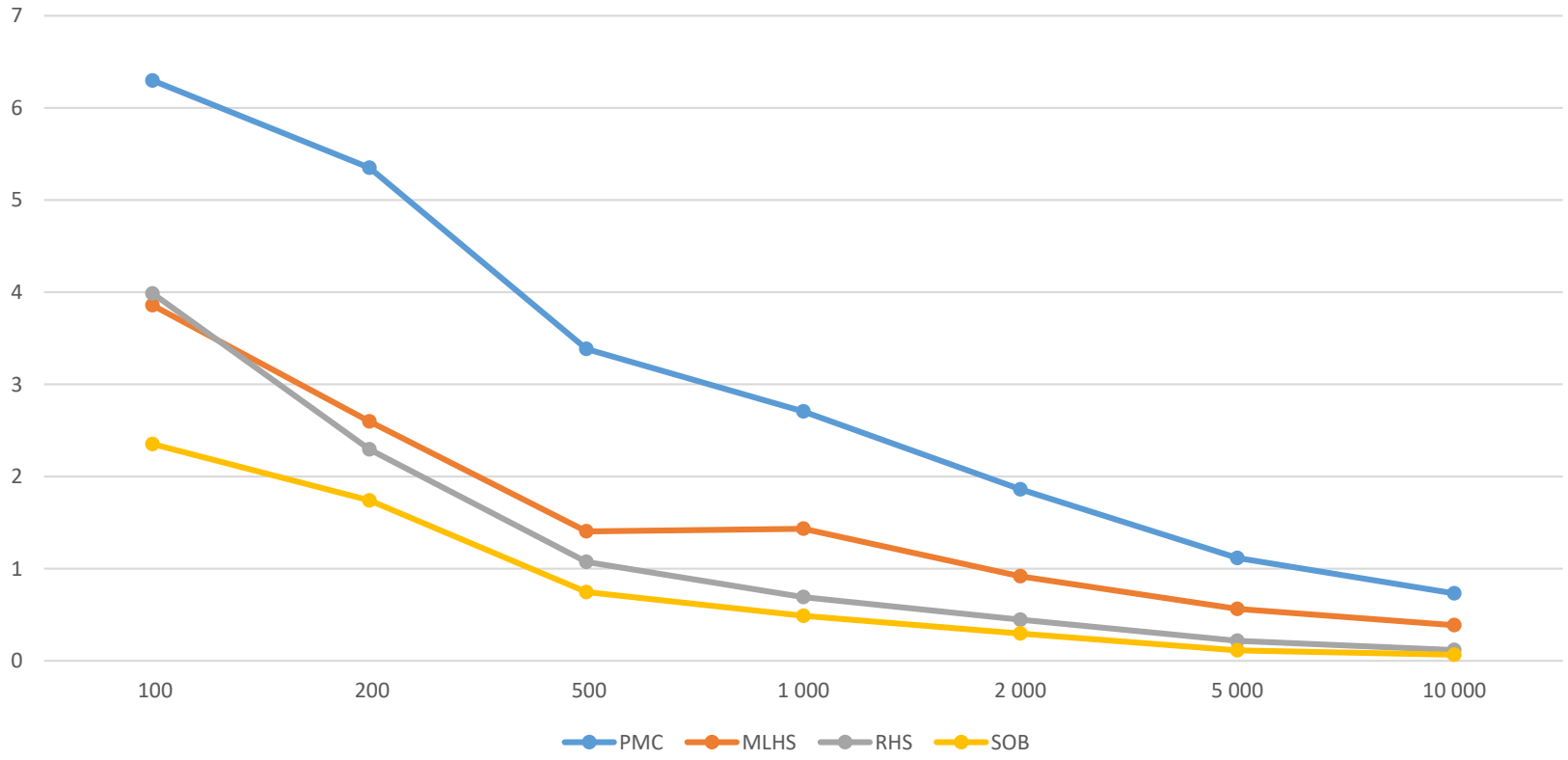
- We want a measure that takes expected values into account but also penalizes variance
  - For typical equality tests – the larger the variance, the more difficult to reject the equality hypothesis
- Testing equivalence instead of equality
  - Reverse the null and the alternative hypotheses
  - Test 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 $\alpha = 0.05$, 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 wrt 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%	3.70%	22.22%	74.08%
200	0.00%	0.00%	0.00%	100.00%
500	0.00%	0.00%	3.70%	96.30%
1,000	0.00%	0.00%	3.70%	96.30%
2,000	0.00%	0.00%	0.00%	100.00%
5,000	0.00%	0.00%	0.00%	100.00%
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	0.37%	7.41%	34.07%	58.15%
200	0.37%	0.00%	25.19%	74.44%
500	0.00%	0.37%	14.81%	84.81%
1,000	0.00%	0.00%	13.70%	86.30%
2,000	0.00%	0.00%	8.89%	91.11%
5,000	0.00%	0.00%	2.22%	97.78%
10,000	0.00%	0.00%	4.07%	95.93%

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	2.22%	6.67%	34.07%	57.04%
200	0.37%	3.33%	28.15%	68.15%
500	0.37%	1.48%	18.15%	80.00%
1,000	2.59%	1.48%	21.11%	74.81%
2,000	0.37%	1.11%	19.26%	79.26%
5,000	3.70%	1.11%	5.56%	89.63%
10,000	0.00%	0.00%	4.44%	95.56%

Results – Sobol draws consistently perform best

- Minimum number of Sobol draws (on average) that outperforms (in terms of $MTL_{0.05}$) 10,000 draws of each type:

	Pseudo-random	*MLHS*	*Halton*
LL	1 167	2 185	8 889
Parameter estimates	2 928	3 648	9 537
z-stats	3 366	4 106	9 374

- How many more draws (relatively, on average) required to perform as good as Sobol draws:

	Pseudo-random	*MLHS*	*Halton*
LL	11.82	5.32	2.18
Parameter estimates	5.54	4.08	2.00
z-stats	5.47	4.34	1.97

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 (at $\alpha = 0.05$)
  - This is exactly what $MTL_{0.05}$ and $MTL_{0.01}$ 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
$p = 0.05$	200	500	500	500	1,000	1,000	1,000	2,000	2,000
$p = 0.01$	500	500	500	1,000	1,000	2,000	1,000	5,000	5,000

Results – how many draws are 'enough'?

- Parameter estimates:
  - No absolute difference level
  - The numbers of draws required for 95% / 99% probability that the difference between parameter estimates < 5%:

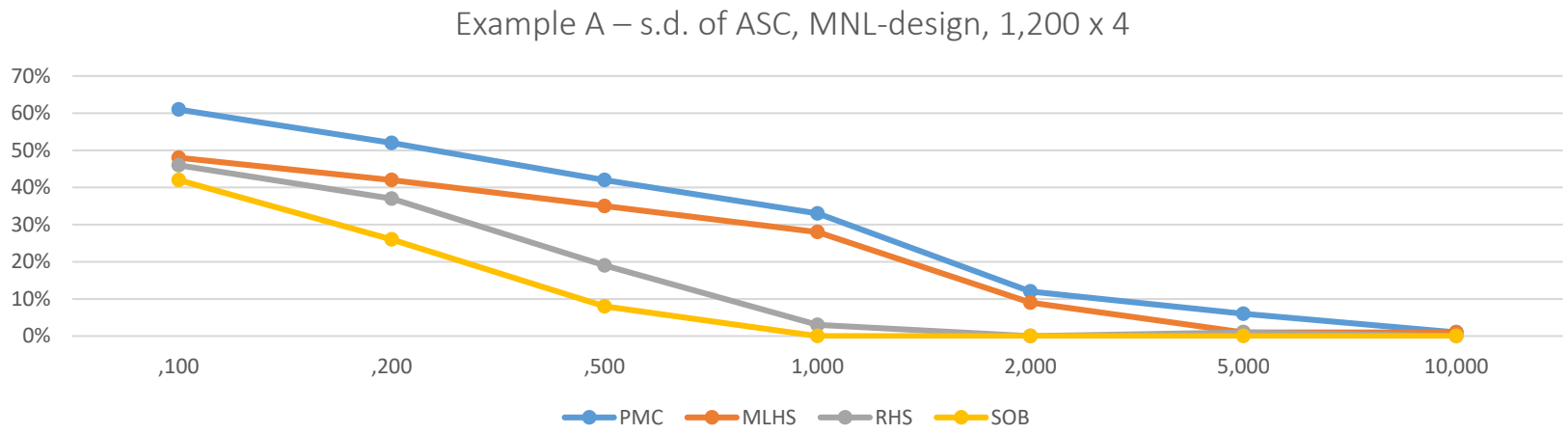
	400 x 4	800 x 4	1,200 x 4	400 x 8	800 x 8	1,200 x 8	400 x 12	800 x 12	1,200 x 12
$p = 0.05$	5,000	2,000	2,000	2,000	1,000	1,000	2,000	2,000	1,000
$p = 0.01$	5,000	2,000	2,000	5,000	2,000	2,000	2,000	2,000	2,000

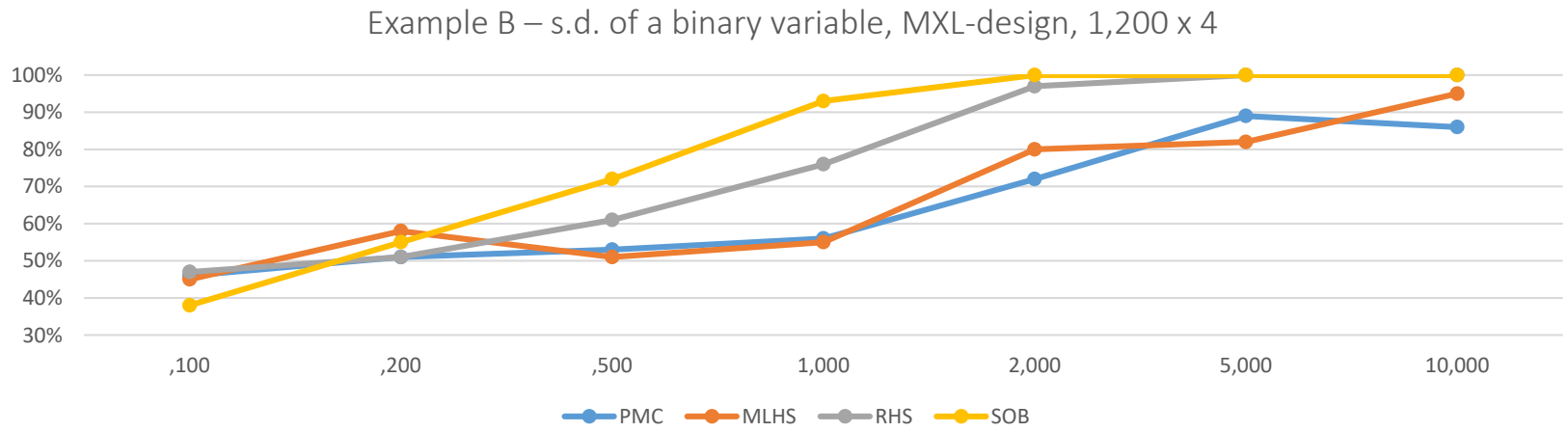
  - The numbers of draws required for 95% / 99% probability that the difference between parameter estimates < 1%:

	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
$p = 0.05$	20,000	10,000	10,000	20,000	10,000	10,000	20,000	20,000	10,000
$p = 0.01$	50,000	20,000	20,000	50,000	20,000	10,000	20,000	20,000	20,000

  - Similar results for comparisons with models estimated using 1,000,000 draws
  - The required number of draws typically higher for standard deviations, lower for means

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





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
  - If 'reliable' is defined as having no more than 5% or 1% chance of erroneous conclusion that a model is significantly better than the same model estimated using a different set of draws (LR-test with 1 d.f.):

Use at least 2,000 (5%) or 5,000 (1%) draws

    - Could be less if the number of individuals is less than 1,200 or the number of choice tasks per individual less than 12
  - If 'reliable' is defined as being 95% sure that one has no more than 5% or 1% simulation-driven variation in parameter estimates:

Use at least 5,000 (5%) or 20,000 (1%) draws

    - Could be less if the number of individuals is more than 400 or the number of choice tasks per individual is more than 4
- Evidence of erroneous inference on significance (both ways) if too few draws are used