Make your own errors

Ernesto Carrella

June 21, 2018

The Setup

- You have a simulation model
- You want to tune parameters to fit data
- Likelihoods are intractable

Current Solution

- Summarise data and simulation state space: S
- Compute some distance function given model parameters θ:

$$(S - S(\theta))^T W (S - S(\theta))$$

- Minimize it!

Current Problems:

$$(S - S(\theta))^T W (S - S(\theta))$$

- How to choose summary statistics ?
- How to weigh the distance ?
- How to minimize ?

Turn your minimization into a regression

1. Repeatedly run the model each time supplying it a random vector $\hat{\theta}$
2. Collect for each simulation its statistics $S(\hat{\theta})$
3. Run K separate regressions for each θ against all summary statistics on the data-set just produced:

$$\begin{cases} \theta_1 = r_1(S_1, S_2, \dots, S_M) \\ \theta_2 = r_2(S_1, S_2, \dots, S_M) \\ \qquad \vdots \\ \theta_n = r_n(S_1, S_2, \dots, S_M) \end{cases}$$

4. Plug in the "real" summary statistics S^* in each regression to get the "real" parameters θ^*

Secret ingredient

- Use a regularized regression and a large set of candidate S. Let the regression choose them.

Works for model selection too

1. Repeatedly run each model
2. Collect for each simulation its generated summary statistics $S(m_i)$
3. Build a classifier predicting the model from summary statistics and train it on the data-set just produced:

$$i \sim g(S_1, S_2, \dots, S_M)$$

4. Plug in "real" summary statistics S^* in the classifier to predict which model generated it

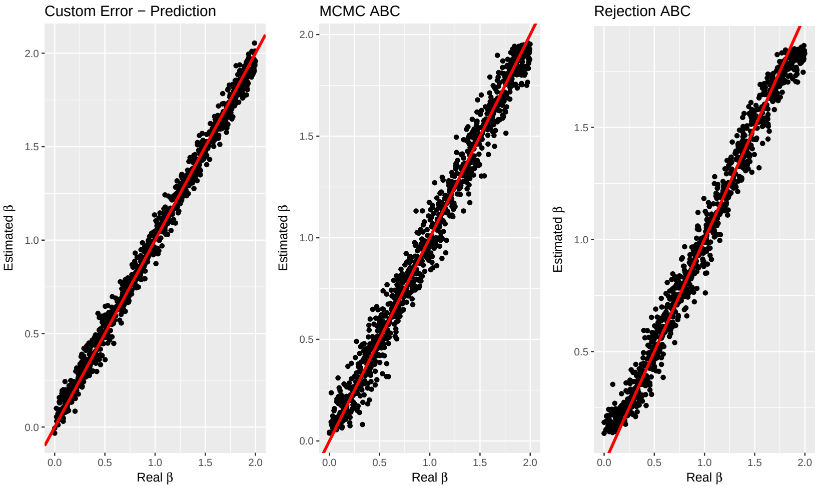
The literature review slide

- Reviews: Hartig et al. (2011); Grazzini and Richiardi (2015)
- Indirect Inference Zhao (2010)
- Selection and Weighing: Liao (2013), Altonji and Segal (1996), Badham et al. (2017)
- Optimization:
  - Usual suspects (Heppenstall, Evans, and Birkin (2007))
  - BACCO: Kennedy and O'Hagan (2001); Salle and Yıldızoğlu (2014); Parry et al. (2013); Ciampaglia (2013)
  - ABC: Beaumont (2010); Grazzini, Richiardi, and Tsionas (2017); Drovandi, Pettitt, and Faddy (2011); Zhang et al. (n.d.)
- "Regression-based methods": Blum and Francois (2010); Blum et al. (2013); Beaumont, Zhang, and Balding (2002)

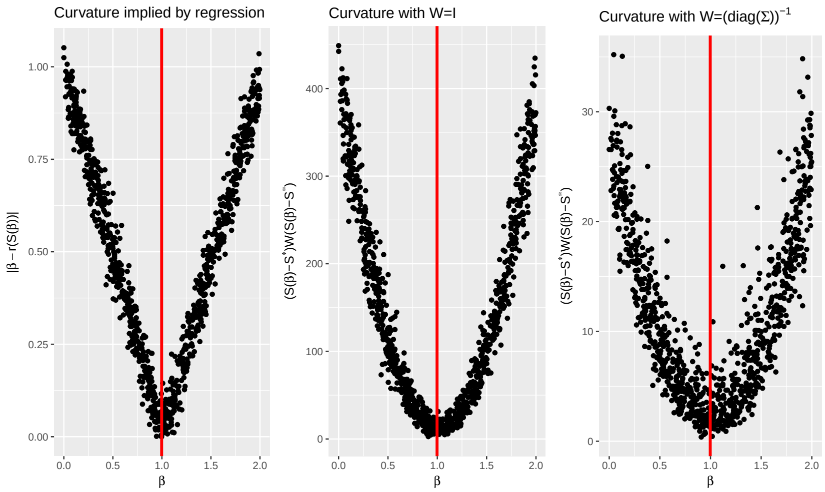
Who needs OLS anyway?

- There are 10 summary statistics $(S_0, \ldots, S_9)$
- Propose the model $S_i = \beta i + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$.
- Assuming that the model is correct, find β given $(S_0, \ldots, S_9)$
    - Train regression $\beta = \text{Intercept} + \sum b_i S_i$

Compare to ABC



Compare to SMD



Regression is informative

- Train regression $\beta = \text{Intercept} + \sum b_i S_i$

Term	Estimate
(Intercept)	0.0279508
b_2	0.0043900
b_3	0.0110419
b_4	0.0125185
b_5	0.0156443
b_6	0.0172918
b_7	0.0214022
b_8	0.0221615
b_9	0.0269217

Broken Lines

- There are 10 summary statistics $(S_0, \ldots, S_9)$
- Propose the model

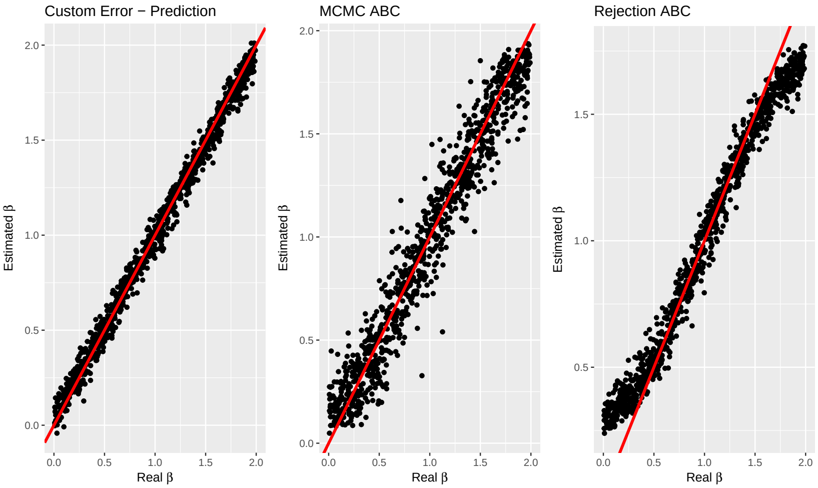
$$S_i = \begin{cases} \epsilon & i < 5 \\ \beta i + \epsilon & i \geq 5 \end{cases}$$

- Assuming that the model is correct, find β given $(S_0, \ldots, S_9)$.
  - Train regression $\beta = \text{Intercept} + \sum b_i S_i$

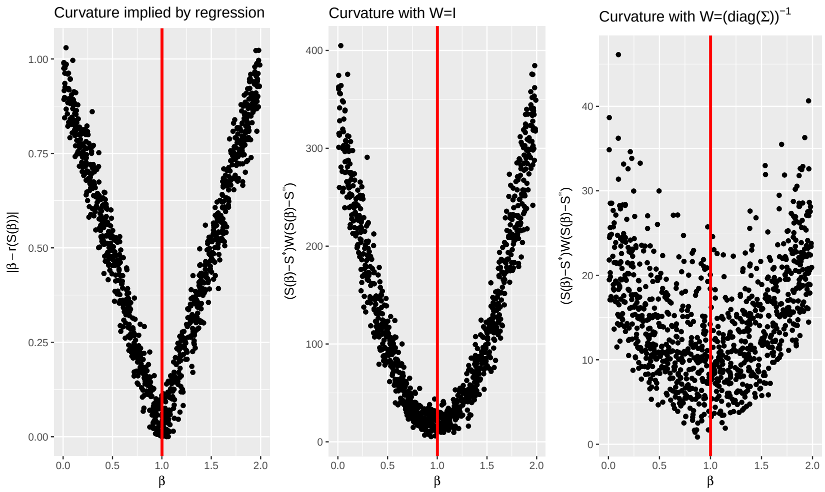
Ignore useless coefficients

Term	Estimate
(Intercept)	0.0313019
b_5	0.0195684
b_6	0.0199352
b_7	0.0237299
b_8	0.0263909
b_9	0.0280121

Compare to ABC



Compare to SMD



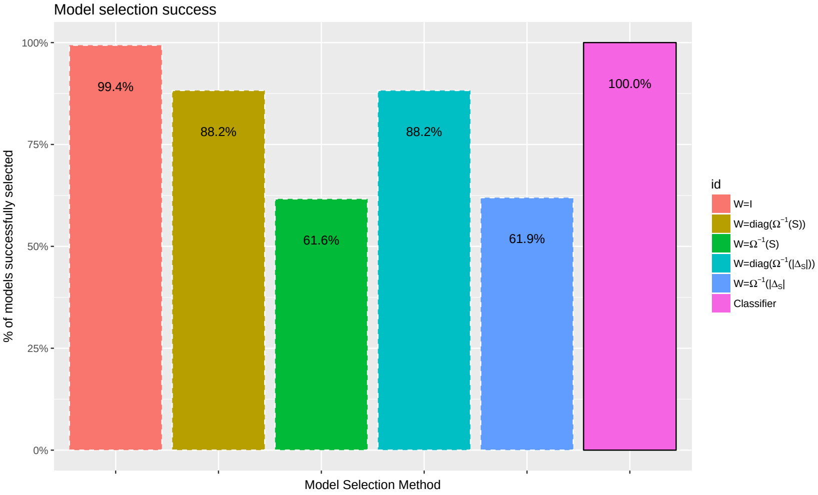
Broken or not broken?

- There are 10 summary statistics $(S_0, \ldots, S_9)$
- Were they generated from the straight line model or broken line model?

Look only at what is important

Term	Estimate
(Intercept)	-9.6861570
b_1	0.1729165
b_2	1.1348468
b_3	1.4334659
b_4	1.9689561

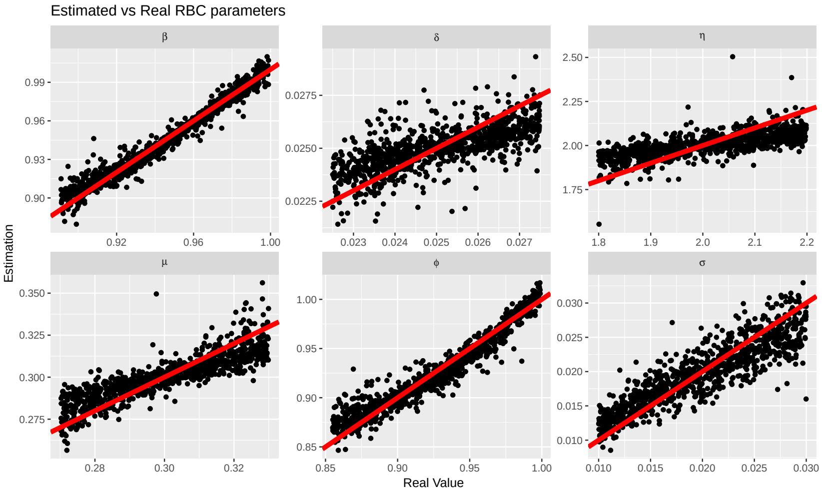
Model Selection Methods



Example 2: RBC Method

- Basic RBC model, 6 parameters: $\beta, \gamma, \eta, \mu, \phi, \sigma$
- Implemented in R (Klima, Podemski, and Retkiewicz-Wijtiwiak 2018)
- Each run, I observe 150 quarters
- Can we find the parameters by looking at:
  1. $t = -5, \ldots, +5$ cross-correlation matrix of Y and r, I, C, L
  2. the lower-triangular covariance matrix of Y, r, I, C, L.
  - And all the squares, and all the pair-wise products (2486 summary statistics)

Out of sample prediction



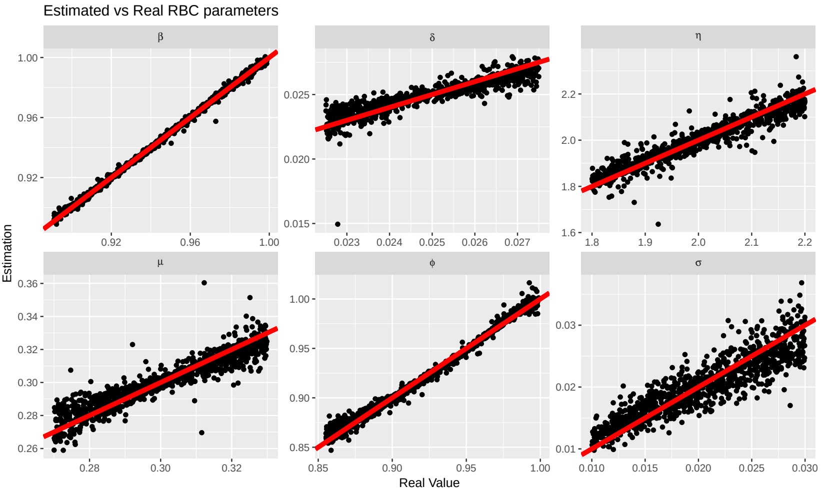
Easy to diagnose

Variable	Variable Bounds	Average Bias	Average RMSE	Predictivity
β	[0.891,0.999]	0.0003105	0.0000351	0.9635677
δ	[0.225,0.275]	-0.0000435	0.0000010	0.4896138
η	[1.8,2.2]	-0.0009048	0.0066570	0.4868081
μ	[0.27,0.33]	0.0001420	0.0000761	0.7462549
σ	[0.01,0.03]	-0.0000385	0.0000071	0.7928857
ϕ	[0.855,0.999]	-0.0001569	0.0001229	0.9249848

RBC - Other summary statistics

- Same RBC model, 6 parameters: $\beta, \gamma, \eta, \mu, \phi, \sigma$
- Implemented in R (Klima, Podemski, and Retkiewicz-Wijtiwiak 2018)
- Can we find the parameters by looking at (40 summary statistics):
  1. Pair-wise VAR-1 fits of Y on r, I, C, L.
  2. the lower-triangular covariance matrix of Y, r, I, C, L.
  3. Linear regression Y on r, C, L
  4. AR(5) parameters of Y
  - And all the squares, and all the pair-wise products (2486 summary statistics)

Better fit



Estimated vs Real RBC parameters
β
δ
η
μ
ϕ
σ
Estimation
Real Value
1.00
0.96
0.92
0.92
0.96
1.00
0.025
0.020
0.015
0.023
0.024
0.025
0.026
0.027
2.2
2.0
1.8
1.6
1.8
1.9
2.0
2.1
2.2
0.36
0.34
0.32
0.30
0.28
0.26
0.28
0.30
0.32
1.00
0.95
0.90
0.85
0.85
0.90
0.95
1.00
0.03
0.02
0.01
0.010
0.015
0.020
0.025
0.030

Easy to diagnose

Variable	Variable Bounds	Average Bias	Average RMSE	Predictivity
β	[0.891,0.999]	0.0000144	0.0000023	0.9975821
δ	[0.225,0.275]	-0.0000311	0.0000004	0.8219723
η	[1.8,2.2]	0.0000300	0.0012209	0.9067242
μ	[0.27,0.33]	-0.0000061	0.0000410	0.8612406
σ	[0.01,0.03]	-0.0000604	0.0000060	0.8121216
ϕ	[0.855,0.999]	0.0002949	0.0000215	0.9878289

Conclusion

- Simple way to parametrise model
- We know regressions
  - No new knowledge required
  - Easy to diagnose
- Ask me for a draft paper

Bibliography

Altonji, Joseph G., and Lewis M. Segal. 1996. "Small-Sample Bias in GMM Estimation of Covariance Structures." *Journal of Business & Economic Statistics* 14 (3). Taylor & Francis, Ltd.American Statistical Association:353. https://doi.org/10.2307/1392447.

Badham, Jennifer, Chipp Jansen, Nigel Shardlow, and Thomas French. 2017. "Calibrating with Multiple Criteria: A Demonstration of Dominance." *Journal of Artificial Societies and Social Simulation* 20 (2). JASSS:11. https://doi.org/10.18564/jasss.3212.

Beaumont, Mark A. 2010. "Approximate Bayesian Computation in Evolution and Ecology." *Annual Review of Ecology, Evolution, and Systematics* 41 (1):379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Beaumont, Mark A, Wenyang Zhang, and David J Balding. 2002. "Approximate Bayesian computation in population genetics." *Genetics* 162 (4):2025–35. https://doi.org/Genetics December 1, 2002 vol. 162 no. 4 2025-2035.

Blum, M G B, M A Nunes, D Prangle, and S A Sisson. 2013. "A comparative review of dimension reduction methods in approximate Bayesian computation." *Statistical Science* 28 (2):189–208. https://doi.org/10.1214/12-STS406.

Blum, Michael G B, and Olivier Francois. 2010. "Non-linear regression models for Approximate Bayesian Computation." *Statistics and Computing* 20 (1):63–73. https://doi.org/10.1007/s11222-009-9116-0.

Ciampaglia, Giovanni Luca. 2013. "A framework for the calibration of social simulation models." *Advances in Complex Systems* 16 (04n05). World Scientific