Application of statistical emulation to an agent-based model: assortative mating and the reversal of gender inequality in education in Belgium

Wim De Mulder[1], André Grow[2], Geert Molenberghs[1], Geert Verbeke[1]

[1] Leuven Biostatistics and Statistical Bioinformatics Centre, Leuven, Belgium
[2] Centre for Sociological Research, Leuven, Belgium

E-mail for correspondence: wim.demulder@cs.kuleuven.be

Abstract: We describe the application of statistical emulation to the outcomes of an agent-based model. The agent-based model simulates the mechanisms that might have linked the reversal of gender inequality in higher education with observed changes in educational assortative mating in Belgium. Using the statistical emulator as a computationally fast approximation to the expensive agent-based model, it is feasible to use a genetic algorithm in finding the parameter values for which the corresponding agent-based model outcome is closest to known empirical output. These optimal parameter values are then interpreted sociologically.

Keywords: Statistical emulation; Gaussian process; Agent-based model; Genetic algorithm.

1 Using agent-based models to simulate assortative mating

1.1 Short introduction to agent-based models

An agent-based model (ABM) is a computational model that simulates the behavior and interactions of autonomous agents. A key feature is that population level phenomena are studied by explicitly modelling the interactions of the individuals in these populations. The systems that emerge from such interactions often are complex and can show regularities that are difficult to anticipate without simulation. For example, with an ABM

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

it has been shown that cultural diversity can persist in a population, even if individuals have a tendency to adapt the cultural traits of those with whom they interact.

1.2 Application of ABM to assortative mating

Educational assortative mating is the sorting of individuals into relations based on educational attainment. Grow et al. (2014) describes an ABM that simulates the mechanisms that might have linked the reversal of gender inequality in higher education with changes in patterns of educational assortative mating. Reversal of such gender inequality refers to the fact that while men have always received more education than women in the past, this imbalance has turned around in large parts of the world. In many countries, women now outperform men in participation and success in higher education. Empirical research shows that this reversal has affected patterns of assortative mating. Earlier, men tended to be similarly or more highly educated than their partners. Today, couples still tend to be similarly educated, but if there is a difference, women tend to be more highly educated. The model studies this association by considering several parameters that describe properties of partner search. For illustration, consider parameters w_s^f and w_s^m that determine the importance that female and male agents attach to the education of prospective partners. The higher their values, the more agents prefer partners with similar educational attainment, and the more willing they become to marry similarly educated agents. We expect that these parameters strongly affect the patterns of educational assortative mating in the agent populations. The simulation outcome considered is the fraction of *hypogamic couples*: couples in which the woman has a higher educational attainment than her partner. We restrict attention to Belgium. The ABM is based on stochastic processes and thus produces a random outcome. We will refer to *the* outcome of the ABM as shorthand for the average outcome over 50 runs. The outcome of the ABM, given w_s^f and w_s^m, is denoted as $f_A(w_s^f, w_s^m)$. The input domain is $[0, 2] \times [0, 2]$.

1.3 Empirical data

To generate realistic agent populations, the ABM is initialized with empirical data from several sources (Grow et al. 2014). The empirical value of the fraction of female hypogamic couples (Section 1.2), denoted as h, is $h = 0.14$ and has been derived from data from the European Social Survey.

1.4 Description of the research question

We consider reverse engineering typical of ABMs: for what parameter values does the ABM produce output that is closest to given empirical output? Applied to our case, this question translates into: for which values of w_s^f

and w_s^m is $|f_A(w_s^f, w_s^m) - h|$ smallest? Obtaining such optimal values provides insight into the processes that might have driven observed population changes and allows validation of the model (e.g., are these parameter values feasible in sociological terms?). See further, Section 3.3.

2 Methodology

2.1 Statistical emulation

Consider an input vector $\mathbf{x}$ and a computer model that maps this input to an output y via some deterministic but possibly unknown function ν, i.e. $y = \nu(\mathbf{x})$. We assume that the considered computer model is highly computationally expensive, so that evaluation of ν in a very large number of different inputs is not feasible. Due to the many nonlinear interactions that are often involved, ABMs are a typical member of the class of highly computationally expensive computer models. Statistical emulation provides an approximation to such a computer model, given training data $(\mathbf{x}_1, \nu(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, \nu(\mathbf{x}_n))$. After training, the resulting so-called emulator is much faster to evaluate than the original computer model. An interesting feature of statistical emulation is that it models the uncertainty in the non-training data points, a feature that is absent in many other approximation techniques such as polynomial approximation. Before training data is presented to the emulator, the output of $\nu(\mathbf{x})$, which is scalar for our ABM case study, is modeled as a Gaussian process with mean $m(\mathbf{x}) = \sum_{i=1}^{q} \beta_i \, h_i(\mathbf{x})$, where β_i represent unknown coefficients, and h_i represent known regression functions. We follow standard practice in emulation and choose them linear. The covariance between the outputs corresponding to inputs $\mathbf{x}$ and $\mathbf{x}'$ is modeled as $\mathrm{Cov}\Big(\nu(\mathbf{x}), \nu(\mathbf{x}')|\sigma^2\Big) = \sigma^2 \, c(\mathbf{x}, \mathbf{x}')$, where σ^2 denotes a variance component and where $c(\mathbf{x}, \mathbf{x}')$ denotes a function that models the correlation between $\nu(\mathbf{x})$ and $\nu(\mathbf{x}')$. We take (in line with standard practice): $c(\mathbf{x}, \mathbf{x}') = \exp\Big[-\sum_i \Big((x_i - x_i')/\delta_i\Big)^2\Big]$, with x_i the ith component of $\mathbf{x}$ and where the δ_i represent parameters, which can be determined by maximizing their posterior distribution $\pi(\delta)$. Notice that the variance component σ^2 in the basic emulation framework is constant.

By applying Bayesian techniques, the prior Gaussian process and the training data can be combined to determine the posterior mean $m^\star(\mathbf{x})$ for any input $\mathbf{x}$, and the posterior standard deviation $\hat{\sigma}\sqrt{c^\star(\mathbf{x}, \mathbf{x})}$, where $\hat{\sigma}$ denotes the estimation of the parameter σ above and where $c^\star(\mathbf{x}, \mathbf{x}')$ denotes the posterior correlation between $\nu(\mathbf{x})$ and $\nu(\mathbf{x}')$. The posterior standard deviation allows to determine a confidence interval around the posterior mean. In this paper we will consider 95% confidence intervals around the posterior mean $m^\star(\mathbf{x})$ which we will denote as $CI(\mathbf{x})$. The application of statistical emulation to ABM is a relatively new research topic, largely neglected in sociological and demographic research. To our knowledge, Bijak

et al. (2013) were the first to illustrate the use of this method in the area of demographic ABMs.

2.2 Genetic algorithms

Genetic algorithms belong to the domain of artificial intelligence and can be used for a.o., function optimization. The optimization is performed via a heuristic search, where a population of candidate solutions is evolved toward better solutions, akin to the biological process of evolution. A main characteristic of genetic algorithms is that they only use function evaluations. This makes them very well suited to solve the optimization problem described in Section 1.4, because an ABM does not provide any other useful information for function optimization than the outcome in given inputs (in particular, information about derivatives is absent). However, since a genetic algorithm considers a large set of candidate solutions, it is typically computationally expensive. This is where the use of the computationally cheap emulator comes in: we will find an approximate solution by replacing the optimization problem by $\arg\min_{w_s^f, w_s^m} |m^\star(w_s^f, w_s^m) - h|$.

To solve this problem, we implemented the genetic algorithm described in Carr (2014, Sec. 2.3). To take into account that $\nu(w_s^f, w_s^m)$ follows a distribution, we use a non-deterministic fitness function FI, namely $FI(w_s^f, w_s^m) = 1/k \sum_{i=1}^{k} |h - z_i|$, where $z_1, \ldots, z_k$ are k values generated from the posterior distribution for $\nu(w_s^f, w_s^m)$, with k a parameter for which the value is chosen by the user. The fitness function measures how good a certain solution is: the lower the fitness value, the better the solution.

3 Results

3.1 Construction of the emulator

The emulator is trained with 100 training data points, for which the inputs are selected according to the Latin hypercube method. We perform a validation of the trained emulator by randomly selecting 5 vectors from the input domain to which the ABM is applied. For these vectors, we also calculate the posterior mean and a 95% confidence interval using the emulator. Finally, the difference, expressed in percentage terms, is calculated between the ABM and the emulator, which we define as $(m^\star(w_s^f, w_s^m) - f_A(w_s^f, w_s^m))/f_A(w_s^f, w_s^m) \times 100\%$. The results are shown in Table 1. The output between the emulator and the ABM is less than 5% in all cases, which we find acceptable. However, four of the five outputs generated by the ABM lie outside the corresponding confidence interval, an indication that the choice of a constant prior variance component might be too restrictive. Some recent research in emulation discusses how to relax this assumption, see e.g. Ba (2012). We plan to consider this and similar research, and to relax the constancy of the variance component, in our future work.

TABLE 1. Validation of the emulator.

w_s^f	w_s^m	$f_A(w_s^f, w_s^m)$	$m^\star(w_s^f, w_s^m)$	% difference	$CI(w_s^f, w_s^m)$
1.398	0.032	0.174	0.169	-2.87%	[0.165,0.172]
1.786	0.68	0.124	0.130	4.84%	[0.126,0.133]
0.716	0.818	0.171	0.170	-0.58%	[0.168,0.172]
0.116	1.358	0.186	0.181	-2.69%	[0.178,0.185]
1.034	1.812	0.127	0.122	-3.94%	[0.119,0.125]

3.2 Application of genetic algorithm

The parameter k, defined in Section 2.2, is chosen as $k = 10$. Other parameters are described in Carr (2014) and are chosen as: mutation rate = 0.2, population size = 500, number of iterations = 100. The initial population is chosen randomly from the input domain, and we run the algorithm 5 times with different initial populations. The results are shown in Table 2. It is clear that there is no global optimum and that all solutions have a very similar and very low fitness value. A highly remarkable observation is that the solutions lie almost exactly on a straight line, the correlation between them being -0.99952.

TABLE 2. Solutions found by the genetic algorithm.

iteration	w_s^f	w_s^m	$FI(w_s^f, w_s^m)$
1	0.873	1.391	0.000216
2	1.176	1.014	0.000181
3	1.026	1.199	0.000206
4	1.394	0.769	0.000225
5	1.306	0.857	0.000200

3.3 Interpretation of the results

The results of the previous section suggest that there is no single optimal value for w_s^f and w_s^m, but that there is an optimal *relationship* between them that is *linear*. Substantively this means that the preferences that the two parameters describe might be partial substitutes in generating the observed level of hypogamy. Thus even if women attach little importance to education (low value of w_s^f) and are thus willing to marry lower educated men, they are unlikely to find lower educated partners who are also willing to marry them if men consider education important (high value of w_s^m).

4 Conclusion

We described the use of statistical emulation and genetic algorithms to find the values of the parameters of an agent-based model that simulates assortative mating. The considered parameters describe the importance that female and male individuals attach to the education of prospective partners. The results indicate that there does not exist a single value for these parameters for which the agent-based model produces output that corresponds to the observed fraction of hypogamic couples. However, there is a single straight line that contains values for the parameters corresponding to the closest match between model output and empirical output.

Acknowledgments: The second author has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. 312290 for the GENDERBALL project.

References

Ba, S. and Joseph, V.R. (2012). Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, **6**, 1838 – 1860.

Bijak, J., Hilton, J., Silverman, E., and Cao, V.D. (2013). Reforging the wedding ring: exploring a semi-artificial model of population for the United Kingdom with Gaussian process emulators. *Demographic Research*, **29**, 729 – 766.

Carr, J. (2014). An introduction to genetic algorithms. See: karczmarczuk.users.greyc.fr/TEACH/IAD/GenDoc/carrGenet.pdf

Grow, A., Van Bavel, J., and De Hauw, Y. (2014). An agent-based computational model of assortative mating and the reversal of gender inequality in education in Europe. In: *European Population Conference 2014*, Budapest, Hungary.