ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)

The use of heuristic optimization algorithms to facilitate maximum simulated likelihood estimation of random parameter logit models

Arne Risa Hole

University of Sheffield, UK

and Hong Il Yoo

Durham University, UK

[Received August 2015. Final revision November 2016]

Summary. Applications of random-parameter logit models can be found in various disciplines. These models have non-concave simulated likelihood functions and the choice of starting values is therefore crucial to avoid convergence at an inferior optimum. Little guidance exists, however, on how to obtain good starting values. We apply an estimation strategy which makes joint use of heuristic global search routines and gradient-based algorithms. The central idea is to use heuristic routines to locate a starting point which is likely to be close to the global maximum, and then to use gradient-based algorithms to refine this point further. Using four empirical data sets, as well as simulated data, we find that the strategy proposed locates higher maxima than more conventional estimation strategies.

Keywords: Differential evolution; Generalized multinomial logit; Mixed logit; Particle swarm optimization

1. Introduction

With an increase in desktop computing power, the random parameter logit (RPL) model has become increasingly common in empirical applications. Also known as the mixed logit, the RPL model provides a flexible framework for modelling discrete choice data. It can approximate any random-utility maximization model arbitrarily well subject to specifying a suitable joint distribution of parameters (McFadden and Train, 2000) and incorporate preference heterogeneity between different individuals alongside panel correlation across observations on the same individual (Revelt and Train, 1998). Applications of the RPL model can be found in a range of disciplines including economics, marketing science, transportation studies and health services research.

Although the RPL model is specified by augmenting the parameters of the multinomial logit model with random heterogeneity, it poses some estimation issues which the multinomial logit model does not. Perhaps the best known is that, in most applications, the RPL likelihood is a multi-dimensional integral which has no closed form expression and needs to be numerically approximated by using simulation. This issue has motivated several studies to explore how best to obtain a more accurate approximation from a given number of draws from the

Address for correspondence: Arne Risa Hole, Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield, S1 4DT, UK.
E-mail: a.r.hole@sheffield.ac.uk

 0035–9254/17/66000
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

joint distribution of random parameters (Train (2009), pages 205–236), and their findings have popularized the use of Halton sequences to generate draws. Although progress has also been made on developing estimation methods which are more computationally attractive than the classical method of maximum simulated likelihood (MSL) in certain aspects (Huber and Train, 2001; Harding and Hausman, 2007; Train, 2008), MSL still remains the most commonly used method as it can be readily applied in conjunction with almost any joint distribution of random parameters.

This paper applies an estimation strategy to address another well-known estimation issue, on which limited practical guidance exists. Specifically, in contrast with its multinomial logit counterpart, the RPL likelihood is not globally concave and may feature several local maxima. As in other similar contexts of non-linear estimation, the selection of 'good' starting values for estimated parameters is crucial to avoid an inferior local maximum. In the RPL literature, nevertheless, empirical studies rarely provide an explicit discussion of starting values used, and the question of how to obtain good starting values has not been the subject of inquiry as far as we know.

Our proposed estimation strategy makes joint use of heuristic optimization algorithms and the usual gradient-based algorithms to obtain the MSL estimates of the RPL. Following Dorsey and Mayer (1995), the central idea is to use heuristic algorithms to locate a starting point which is likely to be close to the global maximum, and then to use gradient-based algorithms to refine this point further. For the heuristic search step, we consider two parsimonious but effective algorithms which can be easily implemented by non-specialists in heuristic optimization: the differential evolution (DE) algorithm (Storn and Price, 1997) and the particle swarm optimization (PSO) algorithm (Eberhart and Kennedy, 1995). These population-based algorithms are well suited to the task of locating candidate solutions away from inferior maxima, as they search comprehensively over the parametric space in looking for the directions of improvement (Gilli and Winker, 2009). Like other gradient-free algorithms, however, they tend to be much slower than gradient-based algorithms in refining a candidate solution to a nearby maximum. Our estimation strategy exploits the global search efficiency of the population-based heuristics and the local search efficiency of gradient-based algorithms, in the sense of Dorsey and Mayer (1995).

We provide computational evidence on the performance of the DE- and PSO-assisted estimation strategies in four empirical data sets of varied sizes, as well as in simulated data sets. Although these strategies can be applied to the estimation of any RPL specification, the case-studies primarily focus on the generalized multinomial logit model of Fiebig *et al*. (2010). The results suggest that the DE-assisted strategy is a very effective tool to diagnose whether a solution that is obtained by following the conventional practice is a global maximum. In all four empirical data sets, the DE-assisted strategy locates solutions which improve on the best conventionally obtained solutions in terms of maximized log-likelihood. Under most computational settings improved solutions are found with sufficiently high empirical frequencies to suggest that a small number of DE-assisted estimation runs would be sufficient for detecting whether a preferred conventional solution is at an inferior maximum. Although the PSO-assisted strategy also locates solutions improving on the best conventional solutions in all four empirical data sets, it does so with much lower empirical frequencies. Moreover, in each data set, the best solution that attains the highest likelihood that we have found comes from the DE-assisted strategy. The findings from simulated data sets support these results.

The remainder of this paper is organized as follows. Section 2 reviews the specification and MSL estimation of the generalized multinomial logit model. Section 3 presents the DE and PSO algorithms. Section 4 presents the main case-studies based on two smaller empirical data sets.

Section 5 briefly introduces the further case-studies that are reported in the on-line appendix, which explore the applicability of the findings to two larger empirical data sets, simulated data sets and other computational settings. Section 6 concludes.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. The generalized multinomial logit model

We assume a sample of N individuals who make a choice from J alternatives in each of T choice situations. The utility that person n derives from choosing alternative j in choice situation t is specified as

$$U_{njt} = \mathbf{x}'_{njt}\boldsymbol{\beta}_n + \varepsilon_{njt} \tag{1}$$

where $\mathbf{x}_{njt}$ is an L-vector of alternative attributes, $\boldsymbol{\beta}_n$ is a conformable vector of utility coefficients and ε_{njt} is an idiosyncratic error term which is independent and identically distributed as type 1 extreme value. Specifying a non-degenerate density of $\boldsymbol{\beta}_n$ leads to an RPL model, which allows interpersonal heterogeneity in preferences for variations in different attributes (Revelt and Train, 1998; McFadden and Train, 2000).

In the generalized multinomial logit model of Fiebig *et al*. (2010), $\boldsymbol{\beta}_n$ is specified as

$$\boldsymbol{\beta}_n = \mu_n\boldsymbol{\beta} + \{\gamma + \mu_n(1-\gamma)\}\boldsymbol{\eta}_n \tag{2}$$

where scalar γ and vector $\boldsymbol{\beta}$ are deterministic, and random vector $\boldsymbol{\eta}_n$ is distributed $\mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$. Using $\mathbf{z}_n$ to denote an M-vector of individual n's characteristics, the random scale factor μ_n is further specified as

$$\mu_n = \exp(\bar{\mu} + \mathbf{z}'_n\boldsymbol{\theta} + \tau v_n) \tag{3}$$

where scalar τ and vector $\boldsymbol{\theta}$ are deterministic, and random scalar v_n is distributed $N(0, 1)$. Scalar $\bar{\mu}$ is a normalizing constant which is calibrated to set the mean of μ_n to 1 when $\boldsymbol{\theta} = \mathbf{0}$. This model can be interpreted as a model that accommodates both canonical 'coefficient heterogeneity' through individual-specific deviations $\boldsymbol{\eta}_n$ around population mean coefficients $\boldsymbol{\beta}$, and 'scale heterogeneity' through the individual-specific scale factor μ_n. Its flexibility is enhanced by the γ-parameter which lets scale heterogeneity affect $\boldsymbol{\beta}$ and $\boldsymbol{\eta}_n$ differently.

Conceptually, allowing the scale factor μ_n to vary by n can be motivated by the possibility that some individuals make choices which are 'noisier', or less aligned with variations in the observed attributes, than others. Then, the idiosyncratic unobservables ε_{njt} would have a larger variance for those individuals, making the scale factor smaller. (This directly follows from the usual identification result for discrete choice models that, when ε_{njt} is normalized as an independently and identically distributed variable, the overall scale of utility is inversely related to the true idiosyncratic variance.) As can be seen from equation (2), however, scale heterogeneity is equivalent to a particular type of coefficient heterogeneity, so the two cannot be sharply distinguished from each other (Fiebig *et al*. (2010), page 398). The main empirical attraction of the generalized multinomial logit model is that the random parameter specification in equation (2) can approximate a wide range of preference patterns, some of which would otherwise call for the use of much less tractable specifications (Keane and Wasi, 2013).

Several other discrete choice models can be derived as special cases of the generalized multinomial logit model. The models GMNL-I and GMNL-II (Fiebig *et al*., 2010) are obtained by

setting γ to 1 and 0 respectively. The generalized multinomial logit model reduces to the standard mixed logit model when the scale factor is assumed to be constant ($\mu_n = 1$), whereas the multinomial logit model with scale heterogeneity, SMNL, is obtained by constraining the covariance matrix of $\boldsymbol{\eta}_n$, $\boldsymbol{\Sigma}$, to $\mathbf{0}$. If both of these constraints are imposed simultaneously, the standard multinomial logit model is obtained. The various special cases of the generalized multinomial logit model are as follows:

(a) GMNL-I, $\boldsymbol{\beta}_n = \mu_n \boldsymbol{\beta} + \boldsymbol{\eta}_n$ ($\gamma = 1$);
(b) GMNL-II, $\boldsymbol{\beta}_n = \mu_n(\boldsymbol{\beta} + \boldsymbol{\eta}_n)$ ($\gamma = 0$);
(c) SMNL, $\boldsymbol{\beta}_n = \mu_n \boldsymbol{\beta}$ ($\boldsymbol{\Sigma} = \mathbf{0}$);
(d) standard mixed logit model, MIXL, $\boldsymbol{\beta}_n = \boldsymbol{\beta} + \boldsymbol{\eta}_n$ ($\mu_n = 1$);
(e) standard multinomial logit model, MNL, $\boldsymbol{\beta}_n = \boldsymbol{\beta}$ ($\mu_n = 1$ and $\boldsymbol{\Sigma} = \mathbf{0}$).

The probability that individual n makes a particular sequence of choices is given by

$$S_n = \int \prod_{t=1}^{T} \prod_{j=1}^{J} \left\{ \frac{\exp(\mathbf{x}'_{njt}\boldsymbol{\beta}_n)}{\sum_{j=1}^{J} \exp(\mathbf{x}'_{njt}\boldsymbol{\beta}_n)} \right\}^{y_{njt}} f(\boldsymbol{\beta}_n | \boldsymbol{\beta}, \gamma, \tau, \boldsymbol{\theta}, \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{\beta}_n \tag{4}$$

where $y_{njt} = 1$ if the individual chose alternative j in choice situation t and $y_{njt} = 0$ otherwise and density $f(\boldsymbol{\beta}_n | \boldsymbol{\beta}, \gamma, \tau, \boldsymbol{\theta}, \boldsymbol{\Sigma})$ is implied by equation (2). The parameters $\boldsymbol{\omega} = (\boldsymbol{\beta}, \gamma, \tau, \boldsymbol{\theta}, \boldsymbol{\Sigma})$ can be estimated by maximizing the simulated log-likelihood function

$$\mathrm{SLL}(\boldsymbol{\omega}) = \sum_{n=1}^{N} \ln \left[\frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T} \prod_{j=1}^{J} \left\{ \frac{\exp(\mathbf{x}'_{njt}\boldsymbol{\beta}_n^{[r]})}{\sum_{j=1}^{J} \exp(\mathbf{x}'_{njt}\boldsymbol{\beta}_n^{[r]})} \right\}^{y_{njt}} \right] \tag{5}$$

where $\boldsymbol{\beta}_n^{[r]}$ is the rth draw from the density of $\boldsymbol{\beta}_n$ and R is the total number of draws.

The standard approach to maximizing the simulated log-likelihood function is to use a gradient-based method such as the Newton–Raphson or Broyden–Fletcher–Goldfarb–Shanno algorithms. See Train (2009), pages 185–204, among others for a description of these methods. The researcher starts with an initial guess of the solution—the starting values—which are then improved on by the algorithm until a specified stopping criterion is reached. As is well known, gradient-based methods cannot distinguish between local and global maxima, and will declare convergence if either type of maximum is reached. Thus, unless the function to be optimized is globally concave, it is not guaranteed that the solution is the global maximum. This issue is of practical importance since the simulated log-likelihood function of the generalized multinomial logit model and its special cases (with the exception of the multinomial logit model) is not globally concave, much like that of other RPL models. In particular, different starting values may lead to different solutions, which suggests that applied researchers should try different sets of starting values to investigate how sensitive the results are to the particular values used. The choice of starting values is rarely discussed in applications of generalized multinomial logit and other RPL models, however. We present some of the strategies that researchers may employ in the following section.

3. Population-based optimization heuristics

This section describes alternative estimation strategies which use population-based heuristic optimization algorithms to obtain starting values for the gradient-based methods. We focus on two population-based optimization heuristics, namely the DE algorithm (Storn and Price, 1997)

and the PSO algorithm (Eberhart and Kennedy, 1995), which have been found to outperform many other heuristic algorithms in a wide range of applications (Gilli and Winker, 2009; Das and Suganthan, 2011).

The main operational aspects of these algorithms are as follows. Suppose that there are a total of K parameters in $(\boldsymbol{\beta}, \gamma, \tau, \boldsymbol{\theta}, \boldsymbol{\Sigma})$ and let a candidate solution be the K-vector of guesses about those parameters. Each algorithm is initialized by generating P different random starting points forming the initial 'population' of candidate solutions, where P is a large number. Then, every one of these candidate solutions is updated over G iterations, or 'generations', where G is another large number. Within each generation, the rule for updating each solution takes into consideration the population of solutions at the end of the preceding generation. The rule also features random elements influencing the direction and extent to which each solution is updated. In the end, the terminal population of P candidate solutions is obtained, and the best candidate solution in the sense of giving the highest simulated log-likelihood value is selected as the fully iterated solution.

For further discussion, let $\boldsymbol{\omega}^{g,p} = (\boldsymbol{\beta}^{g,p}, \gamma^{g,p}, \tau^{g,p}, \boldsymbol{\theta}^{g,p}, \boldsymbol{\Sigma}^{g,p})$ denote a K-vector of possible values of model parameters. Superscripts $p = 1, 2, \ldots, P-1, P$ and $g = 0, 1, \ldots, G-1, G$ identify the pth candidate solution at generation g. Let $\boldsymbol{\Omega}^g = (\boldsymbol{\omega}^{g,1}, \boldsymbol{\omega}^{g,2}, \ldots, \boldsymbol{\omega}^{g,P-1}, \boldsymbol{\omega}^{g,P})$ be the collection of P up-to-date candidate solutions as at g. For later use, we define $g' \equiv g - 1$.

Once the initial population $\boldsymbol{\Omega}^0$ has been generated, each algorithm can be implemented by setting up a simple loop as follows:

```
for g = 1 to G {
    for p = 1 to P {
        DE^{g,p}(F, Cr) or PSO^{g,p}(C, D)
    }
}
```

$\mathrm{DE}^{g,p}(F, \mathrm{Cr})$ and $\mathrm{PSO}^{g,p}(C, D)$ are the rules that the respective algorithms apply to compute the updated candidate solution $\boldsymbol{\omega}^{g,p}$. Each rule depends on two 'tuning parameters' (F, Cr) or (C, D), which are user-specified scalar inputs much like the population size P and the number of generations G. We now turn to a more specific description of each rule.

3.1. Updating process under differential evolution

The updating rule $\mathrm{DE}^{g,p}(F, \mathrm{Cr})$ consists of three main stages: mutation, recombination and selection. The first two stages produce a K-vector of trial values $\mathbf{t}^{g,p}$. This is competed against $\boldsymbol{\omega}^{g',p}$ in the last stage, which selects the better of the two vectors as $\boldsymbol{\omega}^{g,p}$.

The mutation stage uses the amplification factor F and constructs a linear combination of three existing candidate solutions other than $\boldsymbol{\omega}^{g',p}$. For this, three vectors are randomly drawn from $\boldsymbol{\Omega}^{g'} \setminus \{\boldsymbol{\omega}^{g',p}\}$ with equal probabilities and without replacement: let these draws be $\boldsymbol{\omega}^{g',z_1}$, $\boldsymbol{\omega}^{g',z_2}$ and $\boldsymbol{\omega}^{g',z_3}$. Their linear combination $\mathbf{d}^{g,p}$ is specified as

$$\mathbf{d}^{g,p} = \boldsymbol{\omega}^{g',z_1} + F(\boldsymbol{\omega}^{g',z_2} - \boldsymbol{\omega}^{g',z_3}). \tag{6}$$

The recombination stage uses the cross-over probability Cr to construct the K-vector $\mathbf{t}^{g,p}$ by combining elements of $\boldsymbol{\omega}^{g',p}$ and $\mathbf{d}^{g,p}$. This step also involves making $K+1$ different random draws: a positive integer $i^{g,p}$ is drawn from $\{1, 2, \ldots, K-1, K\}$, while K scalars $u_k^{g,p}$ for $k = 1, 2, \ldots, K-1, K$ are drawn from the standard uniform distribution. Now, let $\omega_k^{g',p}$, $d_k^{g,p}$ and $t_k^{g,p}$ denote the kth elements of $\boldsymbol{\omega}^{g',p}$, $\mathbf{d}^{g,p}$ and $\mathbf{t}^{g,p}$ respectively. Each element of $\mathbf{t}^{g,p}$ is chosen according to the following criteria:

$$t_k^{g,p} = \begin{cases} d_k^{g,p} & \text{if } u_k^{g,p} \leqslant \text{Cr or } k = i^{g,p}, \\ \omega_k^{g',p} & \text{otherwise.} \end{cases} \tag{7}$$

Because of the role of integer $i^{g,p}$, $\mathbf{t}^{g,p}$ is always different from $\boldsymbol{\omega}^{g',p}$ in at least one element.

The selection stage evaluates the simulated log-likelihood (5) at the updating target $\boldsymbol{\omega}^{g',p}$ and at the trial vector $\mathbf{t}^{g,p}$. The updated solution $\boldsymbol{\omega}^{g,p}$ equals $\mathbf{t}^{g,p}$ if $\text{SLL}(\mathbf{t}^{g,p}) > \text{SLL}(\boldsymbol{\omega}^{g',p})$, and $\boldsymbol{\omega}^{g',p}$ otherwise. The terminal population $\boldsymbol{\Omega}^G$ consists of P candidate solutions which have thus been updated G times. It is the best solution in $\boldsymbol{\Omega}^G$, in the sense of giving the highest simulated log-likelihood, that is passed to a gradient-based algorithm for further improvement.

The role of the amplification factor F can be likened to that of the step size in gradient-based optimization. In the above updating rule, F is the only component that can be systematically increased by the user to induce a large extent of parametric changes between generations. The cross-over probability Cr, in contrast, influences how often the parametric changes are finalized. Storn and Price (1997) found in a range of applications that, although F is not a probability, the DE algorithm tends to perform the best when it is chosen from the $(0, 1)$ interval much like Cr.

3.2. Updating process under particle swarm optimization

The updating rule $\text{PSO}^{g,p}(C, D)$ deviates from $\text{DE}^{g,p}(F, \text{Cr})$ in that now $\boldsymbol{\omega}^{g,p}$ always changes from $\boldsymbol{\omega}^{g',p}$ even when doing so worsens the simulated log-likelihood. Two additional concepts are needed for a further exposition. First, define $\mathbf{s}^{g,p}$ as the best pth candidate solution that has been obtained up to generation g, i.e. $\mathbf{s}^{g,p}$ is the best out of $\boldsymbol{\omega}^{0,p}, \boldsymbol{\omega}^{1,p}, \ldots, \boldsymbol{\omega}^{g-1,p}, \boldsymbol{\omega}^{g,p}$. Likewise, define $\mathbf{q}^g$ as the best candidate solution that has been obtained up to generation g, i.e. the best out of $\mathbf{s}^{g,1}, \mathbf{s}^{g,2}, \ldots, \mathbf{s}^{g,p-1}, \mathbf{s}^{g,p}$.

$\text{PSO}^{g,p}(C, D)$ uses the acceleration constant C and the inertia weight D to 'fly' $\boldsymbol{\omega}^{g',p}$ towards the best-so-far positions at $\mathbf{s}^{g',p}$ and $\mathbf{q}^{g'}$, thereby obtaining the updated solution $\boldsymbol{\omega}^{g,p}$. The extent of the changes involved, or 'velocity of the flight' $\mathbf{v}^{g,p}$, depends also on two scalars $r_1^{g,p}$ and $r_2^{g,p}$, each of which is drawn from the standard uniform distribution:

$$\mathbf{v}^{g,p} = D\mathbf{v}^{g',p} + C\{r_1^{g,p}(\mathbf{s}^{g',p} - \boldsymbol{\omega}^{g',p}) + r_2^{g,p}(\mathbf{q}^{g'} - \boldsymbol{\omega}^{g',p})\}, \tag{8}$$

$$\boldsymbol{\omega}^{g,p} = \boldsymbol{\omega}^{g',p} + \mathbf{v}^{g,p}. \tag{9}$$

The initial velocity $\mathbf{v}^{0,p}$ is set to the K-vector of 0s so that $\mathbf{v}^{1,p}$ equals a randomly weighted sum of the updating target's ($\boldsymbol{\omega}^{g',p}$) deviations from the two types of best-so-far candidate solutions.

Once the updated solution $\boldsymbol{\omega}^{g,p}$ has been thus computed, $\mathbf{s}^{g,p}$ is re-evaluated for use in the next generation: $\mathbf{s}^{g,p}$ equals $\boldsymbol{\omega}^{g,p}$ if $\text{SLL}(\boldsymbol{\omega}^{g,p}) > \text{SLL}(\mathbf{s}^{g',p})$ and $\mathbf{s}^{g',p}$ otherwise. Then, $\mathbf{q}^g$ is also re-evaluated and set to $\mathbf{s}^{g,p}$ when $\text{SLL}(\mathbf{s}^{g,p}) > \text{SLL}(\mathbf{s}^{g,p'})$ for all $p' \neq p$. In the PSO context, the terminal population of P candidate solutions refers to the collection of $\mathbf{s}^{G,p}$ for $p = 1, 2, \ldots, P-1, P$, instead of $\boldsymbol{\Omega}^G$ *per se*. It is the best solution in that collection, which by definition is $\mathbf{q}^G$, that is passed to a gradient-based algorithm for further improvement.

The acceleration constant C can be viewed as a step size parameter, much like the amplification factor F in the DE updating rule. The inertia weight D controls the tendency to continue flying in the existing direction of parametric changes. C is often set to 2 or less, as in the seminal study of Eberhart and Kennedy (1995). Gilli and Schumman (2010) suggested that setting D to a number less than 1 tends to result in better performance than setting it to 1 as in the seminal study.

4. Main case-studies

This section explores the use of the DE- and PSO-assisted strategies to estimate the generalized

multinomial logit model. Each strategy passes a fully iterated DE or PSO solution as a starting point to a gradient-based algorithm to obtain the final solution. The DE- and PSO-assisted strategies are tools to improve the chance of finding the global maximum. Like any other estimation strategy, they are not guaranteed to find the global maximum. From a practitioner's standpoint, two empirical performance issues may thus be of primary interest.

The first issue is how frequently these estimation strategies can find a solution which is at least as good as the best that can be obtained by using a conventional strategy. This directly relates to whether the DE- and PSO-assisted strategies are a useful addition to the practitioner's toolkit. Starting value search strategies are not part of the common reporting practice. Our own experience and conversations with colleagues, however, suggest that most practitioners would follow a similar approach to those of Greene and Hensher (2010), page 418, and Knox *et al.* (2013), page 74: the conventional strategy is to start from the estimated special cases of the generalized multinomial logit model.

The second issue is whether some configurations of DE and PSO algorithms are conducive to finding such a solution repeatedly. This pertains to how easily the DE- and PSO-assisted strategies can be implemented in practice. As discussed earlier, each algorithm involves tuning parameters affecting how candidate solutions become updated over generations. Without knowing what these parameters need be set to, the DE- and PSO-assisted strategies would be only slightly less ambiguous than the generic advice to 'try a range of starting values'.

Two empirical case-studies are presented below to illustrate the performance issues in detail. The data come from the Pap smear test and pizza A choice experiments that were analysed by the developers of the generalized multinomial logit model (Fiebig *et al.*, 2010; Keane and Wasi, 2013) and are available for download from the *Journal of Applied Econometrics* data archive page for Keane and Wasi (2013). Further information on these data sets is given in Fiebig *et al.* (2010), page 404. All the estimation results were obtained by using Stata 12.1 (StataCorp., 2011). Our on-line appendix provides a further summary of the computational settings.

4.1. Conventional estimation strategy

Implementing the conventional estimation strategy is seemingly straightforward. It entails estimating initially a model which is nested within the generalized multinomial logit model, and then using the results to start the generalized multinomial logit model estimation run. This process is to be repeated for various nested models, and the best out of several resulting generalized multinomial logit solutions is picked as the preferred solution.

In practice, the conventional estimation strategy is only slightly more, if at all, straightforward than implementing the DE- and PSO-assisted strategies. Since nested models include fewer parameters, they provide estimated starting values for only some generalized multinomial logit parameters; the practitioner needs to select custom starting values for the rest, and this selection may affect the final generalized multinomial logit solution. The practitioner also needs to decide how the intermediate solutions are to be computed. All nested models except MNL have non-concave simulated likelihoods with potentially many maxima. Moreover, both GMNL-I and GMNL-II nest MIXL and SMNL, both of which in turn nest MNL.

Table 1 summarizes the custom values that we combined with each nested model's estimates to construct a starting point for the generalized multinomial logit model. Starting from MNL draws on the default setting of Stata's `gmnl` command (Gu *et al.*, 2013) and provides a basis for specifying other starting points. MIXL and SMNL were estimated from the same MNL starting point, ignoring irrelevant parameters. GMNL-I (or GMNL-II) was estimated three times, once from each of the MNL, MIXL and SMNL starting points, again ignoring irrelevant parameters;

Table 1. Starting values based on special cases of the generalized multinomial logit model†

Parameter	*Values for the following models:*				
	MNL	*MIXL*	*SMNL*	*GMNL-I*	*GMNL-II*
β	Est.	Est.	Est.	Est.	Est.
σ	0.10	Est.	0.10	Est.	Est.
τ	0.25	0.25	Est.	Est.	Est.
γ	0	0	0	0	0

†Est. indicates that the restricted model produces the relevant parameter estimates that can be directly used as starting values for the generalized multinomial logit model.

the generalized multinomial logit model, in turn, was then estimated once from each of the three potential GMNL-I (or GMNL-II) starting points, though only the best of the three resulting solutions is reported below.

4.2. Differential evolution and particle swarm optimization assisted estimation strategies

The DE and PSO algorithms require, as user inputs, the population size P and the number of generations G. In addition, both algorithms require an initial population of P candidate solutions that they can improve over G generations.

Following common practice, we set $P = 10K$ where K is the number of estimated parameters. We also set $G = 10K$, as preliminary experimentation with simulated data sets suggested that both algorithms tended to slow down substantially around the $10K$th generation. To illustrate this slowdown in an empirical context, Fig. 1 plots how a selection of DE and PSO starting points that were used in the first case-study (Pap smear) would have varied if G had been set to 420 (or $30K$) instead of 140 (or $10K$). It should be emphasized that the DE and PSO solutions at the $10K$th generation are used as starting points for further optimization, not as the final solutions. All the final solutions are obtained by executing the gradient-based algorithm from the DE and PSO starting points.

The initial population of P solutions is generated as follows. For the generalized multinomial logit parameters to be estimated $\boldsymbol{\omega} = \{\beta, \tau, \gamma, \boldsymbol{\sigma}\}$, consider the bounds given by $\mathbf{l} = \{\mathbf{b}_{\text{MNL}}, 0, 0, \mathbf{0}\}$ and $\mathbf{u} = \{3\mathbf{b}_{\text{MNL}}, 2, 1, 1.5\mathbf{b}_{\text{MNL}}\}$, where $\mathbf{b}_{\text{MNL}}$ is the vector of the MNL estimates and $\mathbf{0}$ is the K-vector of 0s. For each initial solution, each element of $\boldsymbol{\omega}$ is independently drawn from a uniform variable lying between the corresponding elements of $\mathbf{l}$ and $\mathbf{u}$.

The updating process of each algorithm requires two tuning parameters as additional user inputs: amplification factor F and cross-over probability Cr in the case of DE, or the acceleration constant C and the inertia weight D in the case of PSO. We follow Gilli and Schumann (2010) in experimenting with 16 pairs, or configurations, of those tuning parameters per algorithm: a DE configuration is in $F = \{0.2, 0.4, 0.6, 0.8\} \times \text{Cr} = \{0.2, 0.4, 0.6, 0.8\}$, whereas a PSO configuration is in $C = \{0.5, 1.0, 1.5, 2.0\} \times D = \{0.5, 0.75, 0.9, 1.0\}$. The resulting configurations are spaced sufficiently broadly to provide indicative evidence for future applications on what tuning parameter values could be narrowly searched over for further fine-tuning of each algorithm.

Since the updating process is partly random, different DE or PSO starting points would result from the same configuration when different random-number seeds are specified for initializa-

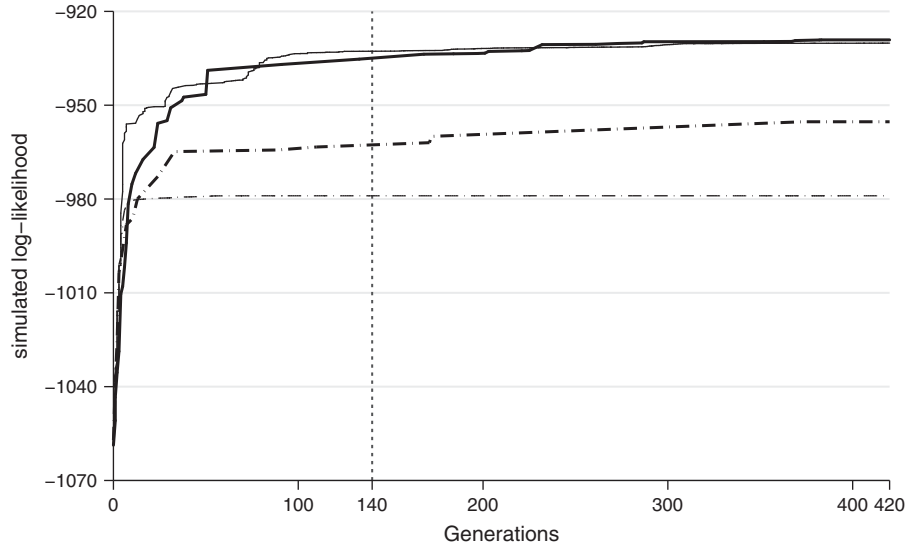

Fig. 1. Pap smear—selected update paths over generations: the best DE (———) and PSO (———) and worst DE (– · – · –) and PSO (– · – · –) refer to the DE and PSO starting points that led to the best and worst final solutions in the Pap smear case-study in Section 4.3; the figure plots how these starting points had been updated until the 140th generation, at the end of which they were passed to the gradient-based algorithm for further optimization to obtain the final estimation results; the figure also plots what these starting points would have become if the algorithm continued without termination until the 420th generation

Table 2. Pap smear: conventional solutions†

Starting point	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
MIXL	−931.065	8.64×10^{-7}	-2.06×10^{-14}	998.8549
GMNL-II	−931.065	2.92×10^{-5}	-4.29×10^{-11}	998.9412
SMNL	−932.133	5.46×10^{-8}	-5.30×10^{-16}	606.9426
GMNL-I	−934.091	1.95×10^{-7}	-2.12×10^{-14}	4732.774
MNL	−960.317	13.04914	-9.22×10^{-6}	1.72×10^{18}‡

†logL, g and H refer to the simulated log-likelihood, its gradient (as a column vector) and Hessian respectively. The infinity norm of g, $\|g\|_\infty$, is the largest element of g in absolute value. $\kappa(H)$ is the 2-norm condition number of H, defined as $\lambda_{\max}/\lambda_{\min}$ where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $-H$.
‡H is ill conditioned (i.e. $\kappa(H) > 6.7 \times 10^7$).

tion. We have obtained 48 DE starting points and 48 PSO starting points, by restarting each configuration three times from the same set of three seeds. In other words, the same set of three different initial populations has been used to obtain the three starting points that are associated with each configuration of each algorithm.

4.3. Results: Pap smear

In the Pap smear data set, each of 79 individuals faced 32 choice scenarios consisting of two options, namely to have a Pap smear test or not. These options are described by six different attributes, including the alternative-specific constant for the have the test option. Estimating the mean $\boldsymbol{\beta}$ and standard deviation $\boldsymbol{\sigma}$ of the canonical random coefficient on each attribute results in 14 generalized multinomial logit model parameters.

Table 3. Pap smear: 10 best DE-assisted solutions†

F	Cr	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
0.8	0.6	*−925.378*	1.44×10^{-7}	-1.78×10^{-15}	2033.113
0.8	0.2	*−925.378*	9.77×10^{-7}	-1.22×10^{-14}	2033.185
0.8	0.2	*−925.378*	6.64×10^{-5}	-3.18×10^{-11}	2033.347
0.6	0.6	*−925.378*	8.07×10^{-5}	-1.35×10^{-10}	2033.3
0.6	0.4	*−925.378*	0.0001	-2.21×10^{-10}	2033.109
0.8	0.2	*−925.378*	0.000438	-3.75×10^{-9}	2033.296
0.8	0.4	*−925.409*	3.92×10^{-7}	-3.02×10^{-15}	3018.498
0.8	0.4	*−925.409*	9.32×10^{-5}	-2.26×10^{-10}	3018.577
0.6	0.2	*−926.308*	5.84×10^{-6}	-1.37×10^{-11}	936.3411
0.6	0.2	*−926.308*	0.00062	-4.03×10^{-8}	936.369

†*F*, Cr, *C* and *D* indicate tuning parameter values leading to relevant starting points. logL is in italics if it is greater than the highest logL (MIXL starting point) in Table 2. See the footnote to Table 2 for other information.

Table 4. Pap smear: 10 best PSO-assisted solutions†

C	D	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
1.5	0.90	*−926.308*	4.74×10^{-6}	-9.87×10^{-13}	936.3309
1.5	1.00	*−926.308*	0.000377	-4.39×10^{-8}	936.2917
2	0.90	*−926.671*	5.08×10^{-8}	-3.56×10^{-17}	1240.651
1.5	0.75	−932.176	5.49×10^{-5}	-6.57×10^{-12}	4973.183
0.5	0.90	−932.176	0.000197	-1.04×10^{-10}	4969.639
1	0.75	−932.376	7.26×10^{-9}	-6.85×10^{-18}	2174.327
2	0.50	−932.376	5.33×10^{-8}	-9.91×10^{-16}	2174.169
0.5	1.00	−932.376	4.00×10^{-7}	-1.64×10^{-14}	2173.287
1	0.90	−934.091	1.90×10^{-8}	-2.92×10^{-16}	512.7021
0.5	0.50	−934.091	1.02×10^{-7}	-6.73×10^{-16}	512.736

†*F*, Cr, *C* and *D* indicate tuning parameter values leading to relevant starting points. logL is in italics if it is greater than the highest logL (MIXL starting point) in Table 2. See the footnote to Table 2 for other information.

Table 2 reports in descending order the simulated log-likelihood values logL of the solutions that were obtained by applying the conventional strategy, along with the usual diagnostics for checking convergence to a local optimum. Stata classifies all solutions as 'converged', implying that the Hessian H is negative definite and the weighted gradient norm $g'H^{-1}g$ is smaller than -1×10^{-5} in magnitude. Further inspection suggests that only the MNL-based solution gives warning signs: the inf-norm of the gradient, $\|g\|_\infty$, deviates far from 0 and the Hessian condition number, $\kappa(H)$, exceeds 1 over the square root of Stata's machine precision. But this is the worst solution which is unlikely to be reported by a practitioner who tries alternative starting points.

The best conventional solution results in logL $= -931.065$. It is also a type of local maximum which practitioners may find particularly convincing as a candidate for the global maximum, because it can be reached from two different starting points, namely MIXL and GMNL-II. The negligible difference between their convergence diagnostics arises because the MIXL-based estimates differ marginally from the GMNL-II-based estimates, in or after the fifth decimal place.

The DE- and PSO-assisted estimation strategies find several solutions which improve on the best conventional solution. The best solution is a DE-assisted solution resulting in logL = −925.378. Table OA1 in the on-line appendix reports the logL-results from all three starts of 16 configurations of each algorithm. The main features of those results may be summarized as follows. 16 of 48 DE-assisted solutions (35%) result in logL > −931.065, ranging from −928.034 to −925.378. Considering that some of the 48 solutions include those resulting from configurations that are not well suited to the present application, a *prima facie* case exists that the DE-assisted strategy is a practically useful complement to the conventional strategy. In contrast, only three out of 48 PSO-assisted estimation runs (6%) result in an improved solution, ranging from −926.671 to −926.308.

Another practically attractive feature of the DE-assisted solutions is clearer indicative evidence on which configurations are likely to work well. Tables 3 and 4 report the top 10 logL-values found with the aid of each algorithm for DE and PSO respectively. A qualitative direction for fine-tuning the DE configuration to the present application would be 'try a big change to the parameter estimates, but accept the resulting change only occasionally'. No similar direction emerges in the case of PSO, as the top 10 solutions are associated with a wider range of configurations. To be specific, the top 10 DE-assisted solutions are overly represented by configurations specifying a large amplification factor F (0.6 and 0.8) and a small cross-over probability Cr (0.2 and 0.4). When restricting attention to the four implied configurations, nine out of 12 DE-assisted estimation runs (75%) find an improved solution, and four of those nine runs reach the highest logL of −925.378. In contrast, a small F (0.2 and 0.4) appears not well suited, regardless of the accompanying Cr: only two of such 28 DE-assisted runs find an improved solution, none of them reaching the highest logL.

The highest logL has been reached from six different DE starting points and displays appropriate convergence diagnostics. Of course, as in the case of the best conventional solution, such repeatability does not imply that the underlying solution is the global maximum. Verifying that a particular solution is the global maximum is considered to be beyond the scope of our study because, as far as we are aware, no definitive guideline exists on how such verification is to be performed. We have, however, verified that the best conventional solution is not the global maximum. Our present and subsequent analysis focuses on the consequences of basing an empirical analysis on the best conventional solution when a DE- or PSO-assisted solution is capable of achieving a higher logL.

Table 5 reports parameter estimates at the second-worst and best conventional solutions, and at the best DE-assisted solution. The second-worst conventional solution (solution A) results from the GMNL-I starting point, and is the worst out of the conventional solutions with acceptable convergence diagnostics. In terms of logL, the best conventional solution (solution B) gains over solution A by about 3 points, and there are marked differences between the coefficient estimates: the mean of the 'alternative-specific constant test', in particular, is about 2.5 times larger in solution A than in solution B (−3.85 *versus* −1.51) and many other estimates disagree even on the first significance figures.

There are less pronounced differences between the best DE-assisted solution (solution C) and the best conventional solution (solution B), although C improves on B by 6 logL-points, or twice as much as B improves on A. The main difference between the solutions is that whereas solution B supports simplifying the model to a more parsimonious GMNL-II model with a non-random-test cost coefficient, solution C does not support such a simplification as both the estimate of γ and the standard deviation of the cost coefficient are significant and non-trivial. The remaining differences are not such that it becomes immediately obvious from simple inspection whether policy relevant statistics derived from these solutions, such as the median willingness to pay

Table 5. Pap smear: generalized multinomial logit parameter estimates†

	Solution A: 2nd-worst conventional		*Solution B: best conventional*		*Solution C: best DE assisted*	
	Estimate	*Standard error*	*Estimate*	*Standard error*	*Estimate*	*Standard error*
If know doctor	1.367‡	0.290	1.202‡	0.240	1.329‡	0.286
	[2.764‡]	0.515	[1.803‡]	0.246	[2.340‡]	0.377
If doctor is male	−3.595‡	0.657	−2.196‡	0.339	−2.775‡	0.556
	[3.828‡]	0.642	[2.760‡]	0.405	[3.472‡]	0.479
If test is due	5.565‡	1.211	4.763‡	0.650	4.969‡	0.824
	[4.691‡]	0.911	[3.530‡]	0.451	[3.478‡]	0.553
If doctor recommends	3.090‡	0.689	1.835‡	0.293	2.226‡	0.422
	[2.943‡]	0.559	[1.681‡]	0.254	[1.201‡]	0.238
Test cost	−0.339‡	0.101	−0.327‡	0.094	−0.245§	0.096
	[0.602‡]	0.165	[0.022]	0.054	[0.180§]	0.076
ASC for test	−3.852‡	1.056	−1.507‡	0.346	−2.281‡	0.512
	[4.140‡]	0.747	[4.447‡]	0.517	[4.099‡]	0.607
γ	0.102§	0.045	0.081	0.054	0.152‡	0.055
τ	1.304‡	0.230	0.940‡	0.144	0.962‡	0.158
logL	−934.091		−931.064		−925.378	

†For each named attribute, the corresponding elements of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ (in square brackets) are reported. The '2nd-worst conventional' and 'best conventional' respectively refer to GMNL-I and MIXL or GMNL-II starting point solutions in Table 2. 'Best DE assisted' refers to the first six solutions in Table 3.
‡Statistical significance at the 1% level.
§Statistical significance at the 5% level.

(WTP) and the predicted choice probability, would be substantively different. (The WTP for a specific attribute is the utility coefficient on that attribute divided by the absolute value of the utility coefficient on the price or cost attribute. The WTP distribution can be simulated first by making simulated draws for all utility coefficients according to equation (2), and then computing relevant ratios of those simulated coefficients.)

To facilitate further comparisons, Table 6 reports selected percentiles of WTP distributions simulated from solutions A, B and C. As expected from the earlier comparison of A with B, these two solutions imply quite a different median WTP, which is the primary statistic on which practitioners are likely to focus (e.g. Small *et al.* (2005)). The implied WTP distributions of solution B and C, in contrast, are only slightly different at the median. The main difference between those two solutions is that, because of heterogeneity in the cost coefficient which is only picked up by solution C, the interpercentile ranges of WTP are much more pronounced for C than B. As a result, conclusions regarding the dispersion of the WTP distribution that is implied by solution B may require reconsideration.

Table 7 compares the three solutions in terms of the predicted changes in the probability of choosing the Pap smear test in response to attribute level variations. The baseline specification of the attribute levels has been motivated by what Johar *et al.* (2013), page 1853, found plausible in the Australian context. As in the case of the median WTP, solutions B and C agree on the substantive conclusions, predicting changes of similar magnitudes and indicating that, under the baseline scenario, the test is more likely to be chosen than not. In this case, however, solution A also yields almost the same results as the others, apart from that in line with its large and negative alternative-specific constant, it predicts a smaller baseline probability of the test (0.45)

Table 6. Pap smear: simulated WTP distributions†

	WTP ($) for the following percentiles:				
	p(10)	*p(25)*	*p(50)*	*p(75)*	*p(90)*
If know doctor:					
2nd-worst conventional	−121	−35	8	51	145
best conventional	−39	−2	36	76	114
best DE assisted	−156	−32	41	122	284
If doctor is male:					
2nd-worst conventional	−244	−96	−27	45	206
best conventional	−182	−128	−67	−8	48
best DE assisted	−455	−212	−83	21	207
If test is due:					
2nd-worst conventional	−292	−64	41	135	340
best conventional	−5	65	144	222	293
best DE assisted	−184	45	156	321	682
If doctor recommends:					
2nd-worst conventional	−162	−35	21	79	205
best conventional	−14	19	57	94	129
best DE assisted	−73	28	69	135	287

†Each WTP distribution has been simulated by making 100000 draws from the joint density of utility coefficients according to the solutions in Table 5. $p(Q)$ denotes the Qth percentile of the simulated distribution.

Table 7. Pap smear: predicted choice probabilities†

	Solution A	*Solution B*	*Solution C*
Base choice probability	0.45	0.57	0.53
Change when test is not due	−0.24	−0.27	−0.26
Change when does not know doctor	−0.06	−0.06	−0.07
Change when doctor is female	0.15	0.12	0.15
Change when doctor recommends	0.12	0.09	0.11
Change when test cost is 0	0.04	0.05	0.04

†Solutions A, B and C are respectively based on 100000 draws from the joint density of utility coefficients according to the '2nd-worst conventional', 'best conventional' and 'best DE-assisted' solutions in Table 5. The base choice probability is the probability of choosing a test (over no test) when the test is due, the patient knows the doctor, the doctor is male, the doctor makes no recommendation and the cost is $30. Each row reports how this probability changes when each attribute changes from its base level.

than B (0.57) and C (0.53). This robustness may stem from the same source as the difficulties of finding the global maximum, namely that different combinations of parametric values lead to similar probabilities or likelihoods.

4.4. Results: pizza A data

In the pizza A data set, each of 178 individuals faced 16 choice scenarios consisting of two hypothetical pizza delivery services. These services are described by eight attributes. Estimating

Table 8. Pizza A: conventional solutions†

Starting point	*logL*	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
GMNL-II	−1361.84	1.45×10^{-5}	-4.88×10^{-12}	173280.2
MIXL	−1365.17	1.30×10^{-6}	-1.58×10^{-12}	20000.77
MNL	−1368.44	3.98×10^{-5}	-1.51×10^{-9}	182428.5
GMNL-I	−1374.45	0.003018	-1.71×10^{-8}	3409.606
SMNL	−1395.5	4.60×10^{-6}	-8.35×10^{-13}	442.66

†See the footnote to Table 2.

Table 9. Pizza A: 10-best DE-assisted solutions†

F	*Cr*	*logL*	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
0.6	0.8	*−1356.8*	18479.79	−51.2587‡	-1.42×10^{19}
0.8	0.4	*−1357.17*	1665.752	−0.16408‡	-3.35×10^{20}
0.6	0.2	*−1357.17*	1887.993	−0.16469‡	-5.19×10^{20}
0.6	0.4	*−1357.17*	298.4716	−0.16521‡	6360351
0.6	0.2	*−1357.53*	0.002223	-2.33×10^{-6}	4232703
0.6	0.4	*−1357.64*	0.000897	-4.58×10^{-7}	2195944
0.8	0.8	*−1357.64*	0.002647	-1.12×10^{-6}	2567936
0.4	0.8	*−1359.03*	0.000146	-4.24×10^{-10}	41664.38
0.8	0.8	*−1359.11*	4.60×10^{-6}	-5.65×10^{-11}	175508.2
0.4	0.2	*−1359.11*	0.001924	-4.17×10^{-9}	171543.3

†logL is in italics if it is greater than the highest logL (GMNL-II starting point) in Table 8. See the footnotes to Table 2 for other information.
‡Stata declared convergence failure since $|g'H^{-1}g|$ exceeds the tolerance criterion (1×10^{-5}).

the mean and standard deviation of the canonical random coefficient on each attribute results in 18 generalized multinomial logit parameters.

Table 8 reports logL-values attained by the conventional solutions. The MIXL and GMNL-II starting points again turn out to be two best conventional starting points. But, this time, only GMNL-II leads to the highest logL of −1361.84. All conventional solutions, including the worst, display acceptable convergence diagnostics.

Tables 9 and 10 report the top 10 logL-values attained by respectively the DE- and PSO-assisted solutions. The full set of the DE- and PSO-assisted estimation runs are available in Table OA2 of the on-line appendix. The results agree with the Pap smear results on two broad conclusions. First, the best solution (logL = −1356.80) is obtained by the DE-assisted strategy. Second, the DE-assisted strategy outperforms the PSO-assisted strategy in terms of finding a solution improving on the best conventional solution: 42% or 20 out of 48 DE-assisted solutions, and 23% or 11 out of 48 PSO-assisted solutions, improve on the best conventional solution.

The current results, however, are quite different from the previous results in one important dimension. 11 DE-assisted solutions (23%) and four PSO-assisted solutions (8%) were declared 'not converged' by Stata, because the associated Hessian is not negative definite and/or $g'H^{-1}g$ exceeds the tolerance level. Importantly, as the upper panel of Table 9 shows, the clear sign of non-convergence is present in the four best solutions that we have obtained.

Since these are symptoms of an empirically underidentified model, we followed the advice

Table 10. Pizza A: 10 best PSO-assisted solutions†

C	D	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
1.5	0.5	*−1359.3*	0.002958	-1.89×10^{-7}	177189.4
1.5	0.75	*−1360*	0.001075	-2.63×10^{-6}	184407.5
1	0.9	*−1360.09*	0.000029	-2.40×10^{-8}	219779.6
2	0.5	*−1360.29*	6.21×10^{-5}	-1.26×10^{-10}	32379.59
2	1	*−1360.29*	0.000188	-4.50×10^{-10}	32251.78
2	1	*−1360.29*	0.000264	-9.16×10^{-10}	32340.78
0.5	1	*−1360.71*	0.00854	-9.80×10^{-9}	218317.4
1	0.9	*−1360.76*	1.71×10^{12}	−0.02901‡	—
1	0.5	*−1360.79*	0.000654	-1.94×10^{-8}	448585.6
2	0.75	*−1360.9*	0.000794	-1.57×10^{-6}	823405.7

†logL is in italics if it is greater than the highest logL (GMNL-II starting point) in Table 8. See the footnotes to Table 2 for other information.
‡Stata declared convergence failure since $|g'H^{-1}g|$ exceeds the tolerance criterion (1×10^{-5}).

of Chiou and Walker (2007) and re-estimated the model with a higher number of simulation draws (10000), using as starting point the best conventional solution. As Chiou and Walker pointed out, using a larger number of draws unmasks empirical underidentification: whereas the best conventional solution displays acceptable convergence diagnostics at 500 draws, the new estimation run failed to attain convergence. Thus, in the present application, the use of the DE- and PSO-assisted strategies leads to a practically different implication from the conventional strategy: namely that the model needs to be simplified before the parameter estimates can be readily interpreted.

5. Further analysis

The results that were described in the previous section suggest that the DE- and PSO-assisted estimation strategies can be a useful tool for improving the chance of finding the global maximum in empirical applications. Between the two strategies, the DE-assisted strategy appears to be the better choice since it improves on the conventional solution more frequently and is more consistent in terms of which configurations are likely to perform well. The best conventional and DE-assisted solutions have led to somewhat (Pap smear) and quite (pizza A) different substantive conclusions based on the estimated generalized multinomial logit models.

The on-line appendix reports an extensive set of results from further case-studies, which echo the relatively superior performance of the DE-assisted strategy. The additional case-studies include applications of the DE- and PSO-assisted strategies to two larger empirical data sets (holiday A and mobile phone) of Fiebig *et al.* (2010), as well as to simulated data sets. We also reanalyse the Pap smear and pizza A data sets, using alternative hybrid estimation strategies which exploit the DE and PSO algorithms jointly with the Nelder–Mead algorithm. (We thank a reviewer for this suggestion.) Finally, we repeat all of our four empirical case-studies in the new context of estimation of the mixed logit model, MIXL, instead of the generalized multinomial logit model.

6. Conclusion

In this paper, we have proposed an estimation strategy which uses DE and PSO algorithms to

obtain starting values for random parameter logit models. Our findings suggest that the DE-assisted strategy can be a very effective tool to diagnose the adequacy of the modelling results that are obtained by using the conventional strategy. The DE configuration ($F = 0.8$, $\mathrm{Cr} = 0.2$) performs particularly well and may serve as a baseline configuration in similar applications.

Our results clearly suggest that repeatedly finding a particular maximum from several starting points is not reliable evidence that it is the global maximum. Given the difficulties of verifying the global maximum in empirical work, it is prudent to embrace the recommendation that Knittel and Metaxoglou (2014) made in a different context of non-linear optimization: namely to report the main differences across several optima found during the estimation process.

We conclude with a few remarks on the estimation run time of the DE-assisted strategy relative to that of the conventional strategy. Comparing the run time is inherently difficult because the estimation issue of interest is not to locate a unique maximum in the fastest time but to locate the best of several possible maxima. The sensitivity of a gradient-based optimizer's run time to starting points poses another source of complication: in the Pap smear case-study, for example, conventionally estimating the generalized multinomial logit model took as little as 17 min (from the MIXL starting point) to 11 h (from the MNL starting point). With these *caveats* in mind, we note that, in all of our empirical case-studies, two DE-assisted estimation runs using ($F = 0.8$, $\mathrm{Cr} = 0.2$) required a comparable amount of time with that of the conventional strategy of searching over major special cases of the final model: continuing with the Pap smear example, each DE-assisted run using this configuration took 1.5 h whereas the conventional strategy took a combined total of 3.5 h even when we overlook the exceptional 11-h run from the MNL starting point. In every empirical case-study and given the same configuration, at least two out of three restarts of the DE-assisted strategy located a higher maximum than the conventional strategy. These findings suggest that running two or three restarts of the DE-assisted strategy would make an effective and computationally feasible addition to the empirical practitioner's toolkit.

Acknowledgements

We thank the Joint Editor, the Associate Editor and two reviewers for their very helpful comments and suggestions. We also thank Jin Yan for alerting us to the literature on population-based optimization heuristics, seminar participants at Copenhagen Business School and the Universities of Aberdeen, Durham and Leeds, and attendees at the Fourth International Choice Modelling Conference.

References

Chiou, L. and Walker, J. (2007) Masking identification of discrete choice models under simulation methods. *J. Econmetr.*, **141**, 683–703.

Das, S. and Suganthan, P. (2011) Differential evolution: a survey of the state-of-the-art. *IEEE Trans. Evoln. Computn*, **15**, 4–31.

Dorsey, R. and Mayer, W. (1995) Genetic Algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *J. Bus. Econ. Statist.*, **13**, 53–66.

Eberhart, R. and Kennedy, J. (1995) A new optimizer using particle swarm theory. In *Proc. 6th Int. Symp. Micromachine and Human Science*, pp. 39–43.

Fiebig, D., Keane, M., Louviere, J. and Wasi, N. (2010) The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marktng Sci.*, **29**, 393–421.

Gilli, M. and Schumann, E. (2010) Robust regression with optimisation heuristics. In *Natural Computing in Computational Finance*, vol. 3 (eds A. Brabazon and M. O'Neill), pp. 9–30. Berlin: Springer.

Gilli, M. and Winker, P. (2009) Heuristic optimization methods in econometrics. In *Handbook of Computational Econometrics* (eds D. Besley and E. Kontoghiorghes), pp. 81–120. Chichester: Wiley.

Greene, W. and Hensher, D. (2010) Does scale heterogeneity across individuals matter?: an empirical assessment of alternative logit models. *Transportation*, **37**, 413–428.
Gu, Y., Hole, A. and Knox, S. (2013) Fitting the generalized multinomial logit model in Stata. *Stata J.*, **13**, 382–397.
Harding, M. and Hausman, J. (2007) Using a Laplace approximation to estimate the random coefficients logit model by nonlinear least squares. *Int. Econ. Rev.*, **48**, 1311–1328.
Huber, J. and Train, K. (2001) On the similarity of Classical and Bayesian estimates of individual mean partworths. *Marktng Lett.*, **12**, 259–269.
Johar, M., Fiebig, D., Haas, M. and Viney, R. (2013) Using repeated choice experiments to evaluate the impact of policy changes on cervical screening. *Appl. Econ.*, **45**, 1845–1855.
Keane, M. and Wasi, N. (2013) Comparing alternative models of heterogeneity in consumer choice behavior. *J. Appl. Econmetr.*, **28**, 1018–1045.
Knittel, C. and Metaxoglou, K. (2014) Estimation of random-coefficient demand models: two empiricists' perspective. *Rev. Econ. Statist.*, **96**, 34–59.
Knox, S., Viney, R., Gu, Y., Hole, A., Fiebig, D., Street, D., Haas, M., Weisberg, E. and Bateson, D. (2013) The effect of adverse information and positive promotion on women's preferences for prescribed contraceptive products. *Socl Sci. Med.*, **83**, 70–80.
McFadden, D. and Train, K. (2000) Mixed MNL models for discrete response. *J. Appl. Econmetr.*, **15**, 447–470.
Revelt, D. and Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level. *Rev. Econ. Statist.*, **80**, 647–657.
Small, K., Winston, C. and Yan, J. (2005) Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica*, **73**, 1367–1382.
StataCorp. (2011) *Stata Statistical Software: Release 12*. College Station: StataCorp.
Storn, R. and Price, K. (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *J. Globl Optimizn*, **11**, 341–359.
Train, K. (2008) EM algorithms for nonparametric estimation of mixing distributions. *J. Choice Modllng*, **1**, 40–69.
Train, K. (2009) *Discrete Choice Methods with Simulation*, 2nd edn. New York: Cambridge University Press.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Online appendix for: The use of heuristic optimization alogrithms to facilitate maximum simulated likelihood estimation of random parameter logit models'.