The University Of Sheffield.

Department Of Economics.

Sheffield Economic Research Paper Series.

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

Arne Risa Hole
Hong Il Yoo

ISSN 1749-8368

SERPS no. 2014021
December 2014

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

Arne Risa Hole[†] and Hong Il Yoo[‡]

[†]Department of Economics, University of Sheffield
a.r.hole@sheffield.ac.uk
[‡]Durham University Business School, Durham University
h.i.yoo@durham.ac.uk

December, 2014

Abstract

The maximum simulated likelihood estimation of random parameter logit models is now commonplace in various areas of economics. Since these models have non-concave simulated likelihood functions with potentially many optima, the selection of "good" starting values is crucial for avoiding a false solution at an inferior optimum. But little guidance exists on how to obtain "good" starting values. We advance an estimation strategy which makes joint use of heuristic global search routines and conventional 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 to a local maximum which stands a good chance of being the global maximum. In the context of a random parameter logit model featuring both scale and coefficient heterogeneity (GMNL), we apply this strategy as well as the conventional strategy of starting from estimated special cases of the final model. The results from several empirical datasets suggest that the heuristically assisted strategy is often capable of finding a solution which is better than the best that we have found using the conventional strategy. The results also suggest, however, that the configuration of the heuristic routines that leads to the best solution is likely to vary somewhat from application to application.

Keywords: mixed logit, generalized multinomial logit, differential evolution, particle swarm optimization

JEL codes: C25, C61

1 Introduction

With an increase in desktop computing power, the estimation of the random parameter logit model (RPL) has become increasingly common in empirical applications. Also known as mixed logit, RPL provides a flexible framework for modeling discrete choice data. RPL can approximate any random utility maximization model arbitrarily well subject to specifying a suitable joint distribution of parameters (McFadden and Train, 2000), and readily incorporate preference heterogeneity between different individuals alongside panel correlation across observations on the same individual (Revelt and Train, 1998). These features make RPL especially attractive when research questions entail the structural analysis of individual preferences from a microeconomic perspective. Related applications can be found in various areas including environmental economics (Layton and Brown, 2000), labor economics (van Soest et al., 2002), transportation economics (Small et al., 2005), international economics (Basile et al., 2008), and health economics (Sivey et al., 2012).

While RPL is specified by augmenting the parameters of the multinomial logit model (MNL) with random heterogeneity, RPL poses a number of estimation issues which MNL does not. Perhaps the best known one is that in most applications, the RPL likelihood is a multidimensional 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 joint distribution of random parameters (Train, 2009, pp.205-236), and their findings have popularized the use of Halton sequences to generate draws. While 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 proposes an estimation strategy to address another well-known estimation issue, on which limited practical guidance exists. Specifically, in contrast to its MNL 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 avoiding potentially false

inferences based on the estimates associated with 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. On the basis of a few studies reporting their starting value search strategies (Greene and Hensher, 2010, p.418; Knox et al., 2013, p.74), the likely conventional practice is to take the starting values from the estimated special cases of a preferred RPL specification.

Our proposed estimation strategy makes joint use of heuristic optimization algorithms and usual gradient-based algorithms to obtain the MSL estimates of RPL. The central idea is to use the 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 to a local maximum which thus stands a good chance of being the global maximum. 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). Sometimes called global search routines (Fox, 2007, p.1013), 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. As 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 investigate the performance of the DE- and PSO-assisted estimation strategies in four different empirical data sets of varied sizes. While these strategies can be applied to the estimation of any RPL specification, the four case studies primarily focus on the generalized multinomial logit model (GMNL) of Fiebig et al. (2010). The traditional RPL specification that augments MNL with normally distributed coefficients is the best known member of the RPL family, so much so that the generic term "mixed logit" is often used to describe this particular specification. GMNL parsimoniously extends it by adding extra parameters to capture interpersonal variations in the overall scale of utility, and tends to perform favorably against other extensions and variants of the traditional specification (Keane and Wasi, 2013). GMNL has been

rapidly gaining influence in the empirical literature, as partly attested by its availability as "canned" commands in software packages like NLOGIT and Stata despite its relative novelty. Our findings do not appear to be exclusively associated with GMNL, however, as they remain qualitatively the same when the four case studies are repeated using the traditional RPL specification.

The results suggest that the DE-assisted strategy is a very effective tool to diagnose whether a solution obtained by following the conventional practice is a global maximum. In all four data sets, the DE-assisted strategy locates solutions which improve on the best conventionally obtained solutions in terms of maximized log-likelihood. Since the updating rules employed by the heuristic algorithms are partly random, the DE- and PSO-assisted strategies may find different solutions over different estimation runs. Under most computational settings we have explored, the DE-assisted strategy finds those improved solutions with high enough 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. While the PSO-assisted strategy also locates solutions improving on the best conventional solutions in all four data sets, it does so with much lower empirical frequencies. Moreover, in each data set, the best solution that attains the highest likelihood we have found comes from the DE-assisted strategy.

In terms of maximized log-likelihood, the best DE-assisted solution is always farther from the best conventional solution than the latter is from the worst conventional solution that displays acceptable convergence diagnostics. Yet, in terms of substantive conclusions, the best DE-assisted and best conventional solutions often show more agreement than the best and worst conventional solutions. The extent of agreement between the solutions is application-specific, however, and the estimation strategy we propose can be used to investigate the robustness of the conclusions drawn from a conventional solution.

The remainder of this paper is organized as follows. Section 2 reviews the specification and MSL estimation of GMNL. Section 3 presents the DE and PSO algorithms. Section 4 presents the main case studies based on two smaller data sets. Section 5 reports further case studies exploring the applicability of the preceding section's findings to two larger data sets and other computational settings. Section 6 concludes.

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 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 a random parameter logit model (RPL), which allows for interpersonal heterogeneity in preferences for variations in different attributes (Revelt and Train, 1998; McFadden and Train, 2000).

In the generalized multinomial logit model (GMNL) 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 $MVN(\mathbf{0}, \boldsymbol{\Sigma})$. Using $\mathbf{z}_n$ to denote an M-vector of individual n's characteristics, the random scale factor μ_n is further specified as

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

where scalar τ and vector $\boldsymbol{\theta}$ are deterministic, and random scalar v_n is distributed $N(0,1)$. Scalar $\overline{\mu}$ 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 one 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 the two components of coefficient heterogeneity 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.[1] 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, p.398). The main empirical attraction of GMNL is that the random parameter specification in (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 GMNL. The GMNL-I and GMNL-II models (Fiebig et al., 2010) are obtained by setting γ to 1 and 0, respectively. The GMNL model reduces to the mixed logit model when the scale factor is assumed to be constant ($\mu_n = 1$), while the the MNL model with scale heterogeneity (SMNL) is obtained by constraining the covariance matrix of $\boldsymbol{\eta}_n$ to $\mathbf{0}$. If both of these constraints are imposed simultaneously, the standard multinomial logit model is obtained. The various special cases of GMNL are summarized below:

- GMNL-I: $\boldsymbol{\beta}_n = \mu_n \boldsymbol{\beta} + \boldsymbol{\eta}_n$ $(\gamma = 1)$
- GMNL-II: $\boldsymbol{\beta}_n = \mu_n(\boldsymbol{\beta} + \boldsymbol{\eta}_n)$ $(\gamma = 0)$
- SMNL: $\boldsymbol{\beta}_n = \mu_n \boldsymbol{\beta}$ $(var(\boldsymbol{\eta}_n) = \mathbf{0})$
- Mixed logit (MIXL): $\boldsymbol{\beta}_n = \boldsymbol{\beta} + \boldsymbol{\eta}_n$ $(\mu_n = 1)$
- Standard multinomial logit (MNL): $\boldsymbol{\beta}_n = \boldsymbol{\beta}$ $(\mu_n = 1$ and $var(\boldsymbol{\eta}_n) = \mathbf{0})$

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

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

where $y_{njt} = 1$ if the individual chose alternative j in choice situation t and 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

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

[1] This directly follows from the usual identification result for discrete choice models that when ε_{njt} is normalized as an iid variable, the overall scale of utility is inversely related to the true idiosyncratic variance.

where $\boldsymbol{\beta}_n^{[r]}$ is the r-th 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 (BFGS) algorithms. See Train (2009, pp.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 upon by the algorithm until a specified stopping criterion is reached. A well-known limitation of gradient-based methods is that the algorithm 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 GMNL model and its special cases (with the exception of the MNL model) is not globally concave, much as 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 GMNL and other RPL models, however. We present some of the strategies that researchers may employ in the following section.

3 Population-based optimization heuristics

A heavily parametrized non-linear model like GMNL is often estimated in two steps. First, a more parsimonious special case of the final model is initially estimated, which in this case ranges from MNL to GMNL-I or GMNL-II. Then, the results are used to specify a starting point for estimation of the final model. While such a procedure provides a data-driven basis for making an initial guess, it does not lead to data-driven guesses about all parameters of the final model, some of which are necessarily constrained and not estimated by the special case. In addition, how closely a constrained maximum resembles an unconstrained maximum is an open question.

This section describes alternative estimation strategies which use population-based heuristic optimization algorithms to obtain initial guesses about all parameters of the unconstrained final model. The central idea here is to use heuristic algorithms to locate a point which is likely to be close to the global maximum, and then to use gradient-based algorithms to improve this point further to a local maximum

which thus stands a good chance of being the global maximum. Heuristic algorithms are often used to optimize non-differentiable functions with multiple optima. The maximum simulated likelihood (MSL) estimation of GMNL involves a differentiable, albeit non-concave, objective function. As Dorsey and Mayer (1995) suggest, such an optimization problem allows practitioners to use both gradient-based and heuristic algorithms in conjunction to exploit the advantage of each. Gradient-based algorithms can locate the global maximum easily if starting values are close to it, but also miss it easily otherwise. Heuristic algorithms may reach a region containing the global maximum more easily because they search through the parametric space more comprehensively for possible directions of improvement. As other gradient-free algorithms, however, they tend to be much slower in refining a candidate solution to a nearby maximum, and are also more prone to solutions which fail the first- and second-order optimality conditions. By exploiting the global search efficiency of heuristic algorithms and the local search efficiency of gradient-based algorithms, our estimation strategies aim to address the practical challenges of finding good starting values and of ensuring that the final solution is at least a local maximum.

We focus on two population-based optimization heuristics, namely the differential evolution (DE) algorithm of Storn and Price (1997) and the particle swarm optimization (PSO) algorithm of Eberhart and Kennedy (1995). Both algorithms can be easily implemented by non-specialists in heuristic optimization, as they require only two tuning inputs to update the model's parameters over iterations; in addition, they have been found to outperform many other heuristic algorithms in a wide range of applications (Gilli and Winker, 2009; Das and Suganthan, 2011). The DE algorithm is much better known in economics, with prior applications in maximum score estimation (Fox, 2007; Fox and Bajari, 2013) and as a building block of a modified Bayesian estimation method (Winchel and Kratzig, 2013), as well as in other classes of numerical optimization tasks (Keller et al., 2004; Krink et al., 2008). Some findings, however, suggest that the PSO algorithm may be better suited to estimation of high-dimensional econometric models (Gilli and Winker, 2009; Gilli and Schumann, 2010).

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 gets updated. In the end, the terminal population of P candidate solutions are 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 $p = 1, 2, \cdots, \mathtt{P}-1, \mathtt{P}$ and g =0,1,$\cdots, \mathtt{G}-1, \mathtt{G}$ identify the pth candidate solution at generation g. Let $\boldsymbol{\Omega}^g = (\boldsymbol{\omega}^{g,1}, \boldsymbol{\omega}^{g,2}, \cdots, \boldsymbol{\omega}^{g,\mathtt{P}-1}, \boldsymbol{\omega}^{g,\mathtt{P}})$ 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)
    }
}
```

$\mathrm{DE}^{g,p}(\mathtt{F}, \mathtt{Cr})$ and $\mathrm{PSO}^{g,p}(\mathtt{C}, \mathtt{D})$ 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" $(\mathtt{F}, \mathtt{Cr})$ or $(\mathtt{C}, \mathtt{D})$, which are user-specified scalar inputs much as 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 (DE)

The updating rule $\mathrm{DE}^{g,p}(\mathtt{F}, \mathtt{Cr})$ 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 $\boldsymbol{\omega}^{g',p}$. To this end, three

vectors are randomly drawn from $\boldsymbol{\Omega}^{g'} \backslash \{\boldsymbol{\omega}^{g',p}\}$ 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

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

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 $\{1, 2, \cdots, K-1, K\}$, while K scalars $u_k^{g,p}$ for $k = 1, 2, \cdots, K-1, K$ 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:

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

Due to 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 $SLL(\mathbf{t}^{g,p}) > SLL(\boldsymbol{\omega}^{g',p})$, 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 $\boldsymbol{\Omega}^G$ 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`, on the other hand, influences how often the parametric changes are finalized. Storn and Price (1997) find in a range of applications that while `F` is not a probability, the DE algorithm tends to perform the best when it is chosen from the $(0, 1)$ interval much as `Cr`.

3.2 Updating process under particle-swarm optimization (PSO)

The updating rule $\text{PSO}^{g,p}(\mathtt{C}, \mathtt{D})$ deviates from $\text{DE}^{g,p}(\mathtt{F}, \mathtt{Cr})$ in that now $\boldsymbol{\omega}^{g,p}$ always changes from $\boldsymbol{\omega}^{g',p}$ even when doing so results in a worse simulated log-likelihood. Two additional concepts 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: that is, $\mathbf{s}^{g,p}$ is the

best one out of $\boldsymbol{\omega}^{0,p}, \boldsymbol{\omega}^{1,p}, \cdots, \boldsymbol{\omega}^{g-1,p}, \boldsymbol{\omega}^{g,p}$. Likewise, define $\mathbf{q}^g$ as the best candidate solution that has been obtained up to generation g: that is, the best one of out $\mathbf{s}^{g,1}, \mathbf{s}^{g,2}, \cdots, \mathbf{s}^{g,p-1}, \mathbf{s}^{g,p}$.

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 involved changes, 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.

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

$$\boldsymbol{\omega}^{g,p} = \boldsymbol{\omega}^{g',p} + \mathbf{v}^{g,p} \quad (9)$$

The initial velocity $\mathbf{v}^{0,p}$ is set to the K-vector of zeros 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 $SLL(\boldsymbol{\omega}^{g,p}) > SLL(\mathbf{s}^{g',p})$ and $\mathbf{s}^{g',p}$ otherwise. Then, $\mathbf{q}^g$ is also re-evaluated and set to $\mathbf{s}^{g,p}$ when $SLL(\mathbf{s}^{g,p}) > SLL(\mathbf{s}^{g,p'})$ 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 $p = 1, 2, \cdots, \mathtt{P}-1, \mathtt{P}$, 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 as 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) suggest 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.

3.3 Further remarks on the use of DE and PSO

In summary, there are three basic user inputs used by both DE and PSO algorithms: the population size P, the number of generations G, and the initial population $\boldsymbol{\Omega}^0$. In addition, there are two tuning inputs used only by a particular algorithm: amplification factor F and cross-over probability Cr for DE, and acceleration constant C and

inertia weight D for PSO.

A full run through the updating loop of either algorithm evaluates the objective function P×G times, with each functional evaluation entailing simulated integration. Specifying larger values for P and G leads to a more comprehensive coverage of the parametric space, but also requires more computer time. This trade-off, and our intended use of a fully iterated DE or PSO solution as starting point for further gradient-based optimization, make it appropriate to exploit somewhat a smaller number of functional evaluations than what would be desirable had the fully iterated solution been intended as the final solution. Much in the same vein as Bhat (1997) sets the maximum number of expectation-maximization (EM) iterations in his hybrid estimation strategy involving the EM and gradient-based algorithms, we will specify moderately large values of P and G such that after P×G computations, the objective function value is likely to vary little with further application of the DE or PSO algorithm: more information is provided below.

It is customary to initialize $\boldsymbol{\Omega}^0$ by taking independent draws from uniform distributions. The selection of bounds for these distributions is not a particularly crucial determinant of either algorithm's performance, as each algorithm allows updated candidate solutions to exceed those bounds. We will choose bounds so that each of the resulting initial candidate solutions may be considered reasonable as a starting point for the GMNL estimation. The configuration of tuning parameters (F, Cr) and (C, D), on the other hand, systematically influences the entire updating path and is known to be a crucial determinant, with most well-suited configurations varying from application to application. We will experiment with a broad range of possible configurations.

4 Main case studies

This section explores the use of the DE- and PSO-assisted strategies to estimate GMNL. 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 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 conversation with colleagues, however, suggest that most practitioners would follow a similar approach as Greene and Hensher (2010, p.418) and Knox et al. (2013, p.74): the conventional strategy is to start from the estimated special cases of GMNL.

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 get 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 Pap Smear test and Pizza A choice experiments analyzed by the developers of GMNL (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 available in Fiebig et al. (2010, p.404). Of 10 empirical illustrations in Fiebig et al. (2010), these two have been selected because, in our view, the required optimization problems are the most representative of what practitioners often face: the number of attributes (6 in Pap smear test, 8 in Pizza A) is within the range commonly seen in modern choice experiments (de Bekker-Grob et al., 2012, p.147) and the GMNL specification to be estimated features uncorrelated normal coefficients.[2]

Both case studies take as given the preferred GMNL specifications of Fiebig et al. (2010) and Keane and Wasi (2013), and aim at estimating parameters $\boldsymbol{\beta}, \tau$, γ and $\boldsymbol{\sigma}$, where the latter denotes the square-root of the elements on the diagonal of $\boldsymbol{\Sigma}$ (the off-diagonal elements are assumed to be zero).[3] The support of γ is the entire real line as in Keane and Wasi (2013), instead of $(0, 1)$ as in Fiebig et al. (2010).

[2]Fiebig et al. (2010) find that in these data sets, the uncorrelated MIXL and GMNL specifications outperform their correlated counterparts in terms of BIC. Keane and Wasi (2013) conduct more extensive model fit comparisons, and find that the uncorrelated GMNL specification also perform favorably against other non-normal mixed logit specifications.

[3]In both case studies $\mathbf{z}_n = \mathbf{0}$ which means that μ_n simplifies to $\exp(\overline{\mu} + \tau v_n)$.

All estimation strategies have been implemented in Stata 12.1, and differ only by which starting points are supplied to the final gradient-based estimation of GMNL. Following Fiebig et al. (2010), the likelihood functions are simulated by taking 500 draws from each random parameter's postulated distribution.[4] The same 500 draws of each parameter are used for all estimation strategies to obviate the interference of simulation noise.

Gradient-based optimization tasks use the *clogit*, *mixlogit* and *gmnl* Stata commands as appropriate, following the default settings of each command unless explained otherwise; these settings include the use of Stata's implementation of the Newton-Raphson algorithm.[5] For the DE and PSO algorithms, we coded our own programs in Stata, using the same simulated likelihood evaluator as *gmnl*. Before progressing to the case studies, we will turn to a further discussion of the implementation details of each estimation strategy.

4.1 Conventional estimation strategy

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

In practice, it 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 of GMNL parameters; the practitioner needs to select custom starting values for the rest, and this selection may affect the final GMNL solution. The practitioner also needs to decide how the intermediate solutions are to be computed. All nested models but 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.

[4] Keane and Wasi (2013) do not report the number of simulated draws used, but comparisons of their MIXL and SMNL results with Fiebig et al. (2010) suggest that it is also 500.

[5] *clogit* is Stata's built-in command for estimating MNL. *mixlogit* is Hole's (2007) command for estimating MIXL. *gmnl* is the same co-author's (Gu et al., 2013) command for estimating GMNL as well as its building blocks, SMNL, GMNL-I, and GMNL-II. Gu et al. (2013) note that in Stata, the Newton-Raphson algorithm tends to outperform quasi-Newton algorithms in terms of the ability to find a GMNL solution satisfying the usual set of convergence diagnostics.

Table 1 summarizes the custom values we combined with each nested model's estimates to construct a starting point for GMNL. The MNL starting point draws on the default setting of the *gmnl* command 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 (GMNL-II) was estimated three times, once from each of the MNL, MIXL and SMNL starting points, again ignoring irrelevant parameters; GMNL, in turn, was then estimated once from each of the three potential GMNL-I (GMNL-II) starting points, though only the best of the three resulting GMNL solutions is reported below.[6]

In our view, this implementation of the conventional strategy is representative of what a typical practitioner would do. A few studies commenting on the estimation process (Greene and Hensher, 2010, p.418; Knox et al., 2013, p.74) only note that starting values have been obtained from nested models. Also, apart from MNL, each nested model requires a non-trivial amount of computer time per estimation run, making it rather cumbersome for the practitioner to experiment with a wide range of custom values, especially when no relevant guidance exists.

In both case studies, our conventional strategy finds solutions which are different from what Keane and Wasi (2013) report. Some of our solutions result in higher, and others worse, log-simulated likelihoods than the corresponding figures in that study. In addition to variations in the process of constructing starting points, such discrepancy may be attributed to different computing environments (Stata and Matlab), for example in terms of pseudo-random number generation. We do not pursue the exact source of the discrepancy because our case studies are not intended as replication exercises. Moreover, even within the Stata computing environment, we find a range of different solutions from different starting points.

4.2 DE- and PSO-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 the common practice, we set `P`=10K where K is the number of estimated parameters. We also set `G`=10K. The choice of `G` varies from application to

[6] In many cases, the GMNL-I (GMNL-II) starting point that led to the best GMNL solution was not the one based on the best of three GMNL-I (GMNL-II) solutions.

application, depending on the nature and purpose of the intended optimization task.[7] In preliminary experimentation with simulated data sets, we noticed that both algorithms tended to slow down substantially around the 10K^{th} generation, motivating our decision to switch to the gradient-based optimization at that point. To illustrate this slowdown in an empirical context, Figure 1 plots how a selection of DE and PSO starting points used in the first case study (Pap Smear) would have varied had `G` been set to 420 (or 30K) instead of 140 (or 10K).

The initial population of `P` solutions is generated as follows. For the GMNL parameters to be estimated $\boldsymbol{\omega} = \{\boldsymbol{\beta}, \tau, \gamma, \boldsymbol{\sigma}\}$, consider the lower and upper bounds given by $\mathbf{l} = \{\mathbf{b}_{MNL}, 0, 0, \mathbf{0}\}$ and $\mathbf{u} = \{3 \times \mathbf{b}_{MNL}, 2, 1, 1.5 \times \mathbf{b}_{MNL}\}$, where $\mathbf{b}_{MNL}$ is the vector of the MNL estimates and $\mathbf{0}$ is the K-vector of zeros. 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 case of DE, or the acceleration constant `C` and the inertia weight `D` in 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 $\texttt{F} = \{0.2, 0.4, 0.6, 0.8\} \times \texttt{Cr} = \{0.2, 0.4, 0.6, 0.8\}$, while a PSO configuration is in $\texttt{C} = \{0.5, 1.0, 1.5, 2.0\} \times \texttt{D} = \{0.5, 0.75, 0.9, 1.0\}$. The resulting configurations are spaced broadly enough 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 initialization. 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 associated with each configuration of each algorithm.

[7] For example, when optimizing a function with a known global optimum, it may be left unspecified to let the optimization run to continue until the optimum is reached (Storn and Price, 1997), whereas when comparing the performance of DE or PSO with that of another algorithm, `G` may be chosen so that with `P`=10K, the same number of functional evaluations results as what the comparator algorithm has performed to find its preferred solution (Gilli and Winker, 2009).

4.3 Results: Pap Smear

In this data set, each of 79 individuals faced 32 choice scenarios consisting of two options, namely get a Pap Smear test or not. These options are described by 6 different attributes, including the alternative-specific constant (ASC) for the get-test option. Estimating the mean ($\boldsymbol{\beta}$) and standard deviation ($\boldsymbol{\sigma}$) of the canonical random coefficient on each attribute results in 14 GMNL parameters.

Table 2 reports in descending order the simulated log-likelihood values (logL hereafter) of the solutions 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 -1E-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 way from zero and the Hessian condition number ($\kappa(H)$) exceeds one 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 solution results in logL of -931.065, which is somewhat higher than -934 in Keane and Wasi (2013). 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.[8] 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 one, resulting in logL of -925.378. Table A1 in Appendix reports the logL results from all 3 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 greater than -931.065, ranging from -928.034 to -925.378. Considering that some of the 48 solutions include those resulting from configurations 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 3 out of 48 PSO-

[8]Both MNL and MIXL starting points led to the same GMNL-II solution that is used as the starting point for GMNL here.

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. Table 3 reports the top ten logL values found with the aid of each algorithm. 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 case of PSO, as the top ten solutions are associated with a wider range of configurations.

To be specific, the top ten 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, 9 out of 12 DE-assisted estimation runs (75%) find an improved solution, and 4 of those 9 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 2 of such 28 DE-assisted runs find an improved solution, none of them reaching the highest logL.

The highest logL has been reached from 6 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 4 reports the second-worst and best conventional solutions, along with the best DE-assisted solution. The second-worst conventional solution (Solution A) results from the GMNL-I starting point, and is the worst one out of conventional solution with acceptable convergence diagnostics. In terms of logL, the best conventional solution (Solution B) gains over Solution A by some 3 points, and there are marked differences between the coefficient estimates: the mean of "ASC test", in particular, is about 2.5 times larger in Solution A than in Solution B (-3.85 vs. -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), despite that 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 while solution B supports simplifying the model to a more parsimonious GMNL-II model with a fixed 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 (WTP) and the predicted choice probability, would be substantively different.[9]

To facilitate further comparisons, Table 5 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 different median WTP, the primary statistic on which practitioners are likely to focus (e.g. Small et al., 2005). The implied WTP distributions of B and C, on the other hand, are only slightly different at the median. The main difference between those two solutions is that due to heterogeneity in the cost coefficient which is only picked up by C, the interpercentile ranges of WTP are much more pronounced for C than B.[10] As a result, conclusions regarding the dispersion of the WTP distribution implied by B may require reconsideration.

Table 6 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, p.1853) find 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

[9]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.

[10]More specifically, the test cost coefficient is very tightly distributed around its mean in B, whereas it is more dispersed in C implying a higher frequency of drawing coefficients close to zero. Since draws from the test coefficient distributions enter the denominators of simulated WTP, this difference can make the WTP distribution of C more dispersed even though other coefficients are similarly dispersed in B and C.

yields almost the same results as the others, apart from that in line with its large and negative ASC, it predicts a smaller baseline probability of the test (0.45) 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 this data set, each of 178 individuals faced 16 choice scenarios consisting of two hypothetical pizza delivery services. These services are described by 8 different attributes. Estimating the mean and standard deviation of the canonical random coefficient on each attribute results in 18 GMNL parameters.

Table 7 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, which lies above -1372 reported in Keane and Wasi (2013).[11] All conventional solutions, including the worst one, display acceptable convergence diagnostics.

The full set of the DE- and PSO-assisted estimation runs are reported in Appendix Table A2, and the results agree with the Pap Smear results on two broad conclusions. First, the best solution is obtained by the DE-assisted strategy and attains logL of -1356.80. Second, the DE-assisted strategy outperforms the PSO-assisted strategy in terms of finding a solution improving on the best conventional solution, even though this time the PSO-assisted strategy does better than in the Pap Smear case study: 42% or 20 out of 48 DE-assisted solutions, and 23% or 11 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 4 PSO-assisted solutions (8%) have been declared "not converged" by Stata, because the associated Hessian is not negative definite and/or $g'H^{-1}g$ exceeds the tolerance level. No solution in the Pap Smear case study displays this issue.

More importantly, the clear sign of non-convergence is present in the four best solutions we have obtained. All these solutions are in the "DE-assisted" panel of Table 8, which reports the 10 best DE-assisted and PSO-solutions. Both $\| g \|_\infty$ and

[11] Only GMNL-II estimated from the SMNL starting point led to this GMNL solution.

$g'H^{-1}g$ of the four solutions evidently deviate from zero, and in the case of the three best solutions $\kappa(H)$ is negative meaning that the Hessian is not negative definite.

Since these are symptoms of an empirically underidentified model, we followed the advice of Chiou and Walker (2007) for further inquiry. Specifically, we re-estimated the model by using as starting point the best conventional solution, and making 10,000 draws to simulate the log-likelihood function. As Chiou and Walker point out, using a larger number of draws unmasks empirical underidentification: while the best conventional solution displays acceptable convergence diagnostics at 500 draws, the new estimation run failed to attain convergence.[12] We note that, in the case of the Pap Smear data, similarly starting an estimation run from the best conventional solution led to convergence within 7 iterations.[13] 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.[14]

Putting the empirical underidentification issue aside, the present case study also yields more ambiguous guidance on configurations of the DE and PSO algorithms. As in the Pap Smear application each PSO configuration tends to perform differently across three restarts, and now the DE configurations also perform somewhat more erratically. The 10 best DE-assisted solutions in Table 8 vary widely in terms of `Cr`, though it still appears to be the case that taking `F` from $\{0.6, 0.8\}$, especially 0.6, is a good choice. The full set of results in Table A2 shows that there are a few more runs with configurations involving `F`=$\{0.2, 0.4\}$ that find an improved solution, on top of the two which already appear in Table 8 (recall that such configurations performed poorly in the Pap Smear application). We note, however, that restricting attention to `F`= $\{0, 6, 0.8\}$ × `Cr`= $\{0.2, 0.4\}$ still seems to be a valid baseline choice: such configurations find an improved solution in 67% or 8 of 12 runs, and encompass (`F`,`Cr`)= $(0.6, 0.4)$ which finds an improved solution in all three restarts.

[12] More specifically, logL rose from -1362.46 to -1353.08 after 31 iterations, at which the Hessian was not negative definite, and no further change occurred during the next 69 iterations.

[13] LogL rose from -938.273 to 936.399.

[14] We found no such evidence of empirical underidentification in a mixed logit model estimated on the same data, where the model followed the same specification as in Section 5.4.

5 Further case studies

The results 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 GMNL models.

In this section, we explore the applicability of the earlier findings to other empirical contexts and computational configurations. The discussion is based on additional sets of estimation results, only a subset of which is reported below for brevity of presentation. Interested readers are referred to our Online Appendix for other discussed results.[15]

5.1 Holiday A data and Mobile Phone data

It is reasonable to ask whether the configurations of DE algorithm which were most likely to improve on the conventional solution in the previous section ($\texttt{F}= \{0, 6, 0.8\} \times$ $\texttt{Cr}= \{0.2, 0.4\}$) will also perform well in other empirical applications. To examine this question, we have applied the same configurations to estimate GMNL using the Holiday A and Mobile Phone data sets from Fiebig et al. (2010) and Keane and Wasi (2013). These data are on individuals' choices from hypothetical holiday packages and from hypothetical mobile phones, respectively. The results are encouraging: the DE-assisted strategy improves on the best conventional solution in 11 out of 12 restarts (92%) in Holiday A, and in all of 12 restarts in Mobile Phone (100%). Furthermore, it is interesting to note that in both data sets the ($\texttt{F} = 0.8$, $\texttt{Cr} = 0.2$) configuration repeatedly locates the best solution we have obtained, just like it did in the Pap Smear application.

The overwhelmingly better performance of the DE-assisted strategy relative to the conventional strategy may be explained by underlying computational difficulties.

[15] The Online Appendix can be accessed at:
https://www.dropbox.com/s/nopimkjotwmsvfu/Dec2014_Hole_and_Yoo_Online_Appendix.pdf?dl=0.

The Holiday A and Mobile Phone data sets have 331 and 493 individuals, respectively, far more than the 79 and 178 individuals in the Pap Smear and Pizza A data sets. Thus, the present cases require many more person-specific likelihoods be simulated. In addition, the Mobile Phone data set requires the estimation of more than 10 extra parameters in comparison with the other data sets. With such factors adding to computational difficulties, the choice of starting values may become even more important.[16]

In both data sets, nevertheless, the best conventional solution and the best DE-assisted solution still show a large amount of agreement on substantive conclusions. Table 9 report the best DE-assisted solution along with the worst conventional and best conventional solutions for Holiday A, and Table 10 report the corresponding results for Mobile Phone.[17] All three solutions display appropriate convergence diagnostics for local maxima in both data sets.

Holiday A yields qualitatively similar results to Pap Smear in the previous section. The best DE-assisted solution achieves a 22.96-point higher logL than the best conventional solution (logL = -2490.92 vs -2513.88), and this difference is much larger than the 9.3 points that the latter gains over the worst conventional solution (logL = -2523.27). Yet, in terms of the parameter estimates, the difference between the best DE-assisted and best conventional solutions is not as evident as that of the best and worst conventional solutions, apart from that the best DE-assisted solution finds much less coefficient heterogeneity for 'Airline' and more for 'Peak season'. The comparisons of simulated WTP distributions lead to the same conclusion.

In Mobile Phone, it is also the case that the best DE-assisted solution gains many more logL points over the best conventional solution than the latter gains over the worst conventional solution. The logL values of the three solutions are -3937.97, -3951.66 and -3954.89 respectively. The comparisons of the parameter estimates are less straightforward in this application as it involves many more parameters, most

[16]This explanation invites the question of why the DE-assisted strategy was found to perform worse in the Pizza A application that involves more individuals and parameters. One possibility is that it is an anomaly due to empirical underidentification. We note that when GMNL is re-estimated by using 10,000 simulated draws and starting from the best conventional solution at 500 draws, the estimation run achieves convergence within a few iterations in both the Holiday A and Mobile Phone data sets, much as in the Pap Smear data set.

[17]Keane and Wasi (2013) report the logL values of -2512 for Holiday A and -3966 for Mobile Phone. In comparison, our best conventional solutions yield -2513.88 and -3951.66 in the respective data sets.

of which are statistically insignificant at all conventional levels. It is, nevertheless, evident that the best DE-assisted solution stands out from both the best and worst conventional solutions. Several standard deviation estimates are significant only in the best DE-assisted solution, and often larger in magnitude than the corresponding estimates in one or both of the conventional solutions. Thus, if significant standard deviations are used to gauge market segments to which particular mobile phone features may appeal, the best DE-assisted solution can lead to quite different marketing decisions than the best conventional solution.

We conclude this subsection with remarks on the PSO-assisted strategy. The previous section suggests that the performance of various PSO configurations tends to be erratic across restarts. In the absence of clearer evidence on suitable baseline configurations, we have applied those drawn from $\mathtt{C} = \{1.5, 2.0\} \times \mathtt{D} = \{0.75, 0.9\}$ to the Holiday A and Mobile Phone data sets by restarting each of the resulting four configurations three times. The results again suggest that the DE-assisted strategy outperforms the PSO-assisted strategy: the latter improves on the best conventional solution less frequently (4 out 12 restarts in Holiday A and 7 out of 12 restarts in Mobile Phone), and the best PSO-assisted solution achieves worse logL than the best DE-assisted solution (-2507.70 in Holiday A and -3949.79 in Mobile Phone).

5.2 All data sets: comparison with the random perturbation strategy

Our use of the DE and PSO algorithms is essentially a sophisticated method for obtaining a suitable random starting point. As the non-identical results across the three starts from the same configurations illustrate, each algorithm works by refining the initial population of several random starting points repeatedly to produce one improved random starting point. In addition, at least in the Pap Smear and Holiday A data sets, the best DE-assisted and best conventional solutions resemble each other closely in terms of parameter estimates. This proximity, together with the inherent randomness of the DE starting points, leads to the question of whether using a randomly perturbed version of the best conventional solution as starting point could be considered as an effective substitute for the DE-assisted estimation strategy.

To address this question, we have applied the following random perturbation strategy to all four data sets as follows. For each parameter estimate in the best conven-

tional solution, a draw is made from a uniform distribution over ±4 standard errors of the estimate. A perturbed starting value for the relevant parameter is obtained by adding up the estimate and the uniform draw. Then, a new random starting point is specified as a vector of the perturbed starting values for all parameters. 20 such starting points have been generated for each data set, and used to estimate GMNL via the Newton-Raphson algorithm.

Table 11 reports the best five logL values resulting from the 20 perturbed starting points. The results suggest that while the perturbation strategy may sometimes be useful in checking for the robustness of the best conventional solution, it cannot readily locate or improve on the best DE-assisted solution. In the Pap Smear data, all five best solutions coincide with the best conventional solution, while in the Pizza A data, all solutions are worse than the best conventional solution. In the Holiday A data, all five best solutions improve on the best conventional solution, but even the very best perturbed solution gains only 1.41 points in terms of logL, much smaller than the best DE-assisted 's gain of 22.96 points. In the Mobile Phone data, the perturbation strategy again turns out to be useful in detecting the inadequacy of the best conventional solution, which is beat by all five best perturbed solutions, but not capable of locating or outperforming the best DE-assisted solution.

5.3 Pap Smear and Pizza A: 20 starts

Our findings so far have suggested that good baseline configurations of the DE algorithm can be drawn from $\texttt{F} = \{0.6, 0.8\} \times \texttt{Cr} = \{0.2, 0.4\}$. As explained earlier the starting point for the algorithm is randomly determined, and we now explore the robustness of the configurations from an alternative angle by restarting each of the four configurations using twenty different random number seeds instead of three as in the previous analysis. For this purpose, we use the Pap Smear and Pizza A data sets whose smaller sizes make them more amenable to a large number of estimation runs.

In each data set, the results over 80 restarts confirm that the performance of these configurations is consistently good. In the Pap Smear data, 49 out of 80 restarts (61.25%) improve on the best conventional solution. The frequency is smaller than the 75% (over 12 comparable restarts) found earlier, but still covers the majority of cases. In the Pizza A data 62 out of 80 restarts (77.5%) improve on the best conventional solution, that is with a higher frequency than the 67% found earlier.

Table 12 reports the ten best DE-assisted solutions found from the 80 restarts in each data set. The results for the Pap Smear data suggest that (as in the case of the best conventional solution) repeatedly finding a particular maximum is not a reliable sign that it is the global maximum. Now, there are two new maxima at the logL values of -924.359 and -924.788, both of which are higher than the logL of -925.378 in the best DE-assisted solution found in the previous section, which was reached four times out of the 12 restarts from the configurations under consideration. An interesting aspect of the parameter estimates at -924.359 is that like the best conventional solution, the standard deviation of the cost coefficient is small and insignificant, in contrast with the "best DE-assisted" solution of the previous section where it is significant. Given the difficulties of verifying the global maximum in empirical work, it appears prudent to report all main differences across several maxima found in estimation runs, as Knittel and Metaxoglou (2014) recommend in the context of the Berry-Levinsohn-Pakes method of demand estimation.

5.4 All data sets: mixed logit case studies

While our focus so far has been on GMNL, the presence of several local maxima is a feature of all random parameter logit (RPL) models. Our DE- and PSO-assisted estimation strategies can be readily adapted to the estimation of other RPL models, and in this section we explore whether the above findings are generalizable to the RPL model with normally distributed coefficients and no scale heterogeneity (MIXL). This model is the best known and arguably most widely estimated RPL specification, to the point where the generic term "mixed logit" is often used to describe this particular model (e.g. Fiebig et al., 2010). For the four data sets in use, the preferred MIXL specification of Fiebig et al. (2010) constrains the off-diagonal elements of $\boldsymbol{\Sigma}$ to zero, like their preferred GMNL specification. We take their preferred MIXL specification as given and estimate the mean ($\boldsymbol{\beta}$) and standard deviations ($\boldsymbol{\sigma}$) of the normally distributed coefficients.

For each data set, several MIXL solutions have been obtained using the same tuning parameter values for the DE and PSO algorithms as in the previous sections. Only one conventional solution has been obtained in this case since using the MNL coefficients as starting values is likely to be the most common strategy for estimating MIXL. As far as we are aware, no previous study has made an explicit mention of

starting values used in estimating the MIXL specification of interest here, presumably because the underlying optimization task may be perceived as numerically simple in that the postulated utility function is linear in parameters and convergence to local maxima can be achieved from a wide range of starting values.[18]

The results across the four data sets suggest that our earlier findings on the performance of DE- and PSO-assisted strategies are not exclusively associated with GMNL (see the Online Appendix for detailed results). Despite the relative numerical simplicity of the MIXL optimization task, the DE- and PSO-assisted strategies perform better than both the conventional strategy and the random perturbation strategy. The DE-assisted strategy still outperforms the PSO-assisted strategy in that the former locates solutions improving on the conventional solution with a greater frequency, and it also finds the best solution out of the ones we have obtained. Moreover, the DE-assisted results from the Pap Smear and Pizza A data sets show that our preferred baseline configurations based on Section 4 are well-suited to the MIXL specification too.

One notable difference when comparing the MIXL and GMNL results is that the MIXL solutions at various local maxima show much greater agreement in terms of policy-relevant statistics than the GMNL solutions do. The conventional solution, the best solution from 20 randomly perturbed starting points and the best DE-assisted solution can be found in the Online Appendix. Presumably because the random scale factor, which can influence all other parameters, is absent in MIXL, all three sets of estimates look very similar and produce almost the same percentiles of the WTP distribution.

6 Conclusion

It is well known that the log-likelihood function of the random parameter logit model may feature multiple maxima, and that the final estimates may be sensitive to the choice of starting values. Only limited documentation and practical guidance exist, however, on the issue of which starting values to use in the estimation process. In this paper, we have proposed an estimation strategy which uses the differential evolution

[18]This contrasts with, for example, the cases of GMNL and an RPL model allowing for log-normally distributed coefficients. The former postulates a non-linear utility function while the latter leads to a likelihood function whose local maxima cannot be easily located (Huber and Train, 2001).

(DE) and particle swarm optimization (PSO) algorithms to obtain starting values. These heuristic algorithms search over the parameter space much more comprehensively than gradient-based algorithms, and can be expected to locate a point close to the global maximum more easily. We have applied this strategy in four different empirical data sets to estimate the generalized multinomial logit model (GMNL), a random parameter logit model featuring both scale and coefficient heterogeneity. The objectives of our empirical applications have been to examine how the DE- and PSO-assisted strategies perform relative to each other as well as to common strategies that most practitioners are likely to be currently following, and also to investigate whether there is a particular configuration of each algorithm which repeatedly results in satisfactory performance. For the common strategies, we have considered (i) the conventional strategy of starting from the estimated special cases of the final model and (ii) the random perturbation strategy of taking random starting points around the best solution found via (i).

Our findings suggest that the DE-assisted strategy can be a very effective tool to diagnose the adequacy of the modeling results obtained using the conventional strategy. In all four data sets, the DE-assisted strategy has located solutions which attain higher log-likelihood values than what the PSO-assisted strategy and the conventional strategy have found. Those improved solutions have been obtained with high enough empirical frequencies to suggest that a small number of DE-assisted estimation runs would be sufficient for detecting the potential inadequacy of a currently preferred conventional solution. In contrast, the random perturbation strategy has failed to improve on the best conventional solution in two of the four data sets, in addition to resulting in worse solutions than the best DE-assisted solution in all data sets.

The best DE-assisted solution has always achieved a larger gain in the log-likelihood over the best conventional solution than the latter has achieved over the worst conventional solution. These larger gains make it interesting to note that in two of the four case studies (Pap Smear and Holiday A), the best DE-assisted and best conventional solutions overlap more in terms of substantive conclusions, such as the median willingness-to-pay, than the best and worst conventional solutions do. It therefore seems possible for the policy implications of a carefully selected conventional solution to remain valid even when the solution is at an inferior local maximum.

An attractive feature of our heuristically assisted estimation strategy is its versatility. Once programmed, it can be readily applied to maximize the log-likelihood of

other random parameter logit models. As briefly discussed, we have obtained qualitatively similar findings on the heuristically assisted estimation of the traditional random parameter logit model featuring normally distributed coefficients and a fixed scale parameter. We leave further application of our estimation strategy to other random parameter logit models to future research.

As a final remark, our results clearly suggest that repeatedly finding a particular maximum from several starting points is not reliable evidence that it is the global maximum. For example, in one empirical data set (Pap smear test), the conventional strategy found a particular maximum repeatedly when the optimization process started from several estimated special cases of GMNL, as well as from randomly perturbed points around that maximum. Yet, the initial runs of the DE-assisted strategy described in Section 4 repeatedly located a higher maximum, and the extra runs of the same strategy described in Section 5 resulted in an even higher maximum. Given the difficulties of verifying the global maximum in empirical work, it appears prudent to embrace the recommendation that Knittel and Metaxoglou (2014) make in a different context of non-linear optimization: namely to report the main differences across several optima found during the estimation process.

References

Basile R, Castellani D, Zanfei A. 2008. Location choices of multinational firms in Europe: the role of EU cohesion policy. *Journal of International Economics* **74**: 328-340.

Bhat C. 1997. An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science* **31**: 34-48.

Chiou L, Walker J. 2007. Masking identification of discrete choice models under simulation methods. *Journal of Econometrics* **141**: 683-703.

de Bekker-Grob E, Ryan M, Gerard K. 2012. Discrete choice experiments in health economics: a review of the literature. *Health Economics* **21**: 145-172.

Das S, Suganthan P. 2011. Differential evolution: a survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation* **15**: 4-31.

Dorsey R, Mayer W. 1995. Genetic Algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *Journal of Business and Economic Statistics* **13**: 53-66.

Eberhart R, Kennedy J. 1995. A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micromachine and Human Science*: pp. 39-43.

Fiebig D, Keane M, Louviere J, Wasi N. 2010. The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science* **29**: 393-421.

Fox J. 2007. Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *Rand Journal of Economics* **38**: 1002-1019.

Fox J, Bajari P. 2013. Measuring the efficiency of an FCC spectrum auction. *American Economic Journal: Microeconomics* **5**: 100-146.

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

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

Greene W, Hensher D. 2010. Does scale heterogeneity across individuals matter? An empirical assessment of alternative logit models. *Transportation* **37**: 413-428.

Gu Y, Hole A, Knox S. 2013. Fitting the generalized multinomial logit model in Stata. *Stata Journal* **13**: 382-397.

Johar M, Fiebig D, Haas M, Viney R. 2013. Using repeated choice experiments to evaluate the impact of policy changes on cervical screening *Applied Economics* **45**: 1845-1855.

Harding M, Hausman J. 2007. Using a Laplace approximation to estimate the random coefficients logit model by nonlinear least squares. *International Economic Review* **48**: 1311-1328.

Hole A. 2007. Fitting mixed logit models by using maximum simulated likelihood. *Stata Journal* **7**: 388-401.

Huber J, Train K. 2001. On the similarity of Classical and Bayesian estimates of individual mean partworths. *Marketing Letters* **12**: 259-269.

Keane M, Wasi N. 2013. Comparing alternative models of heterogeneity in consumer choice behavior.*Journal of Applied Econometrics* **28**: 1018-1045.

Keller K, Bolker B, Bradford D. 2004. Uncertain climate thresholds and optimal economic growth. *Journal of Environmental Economics and Management* **48**: 723-741.

Knittel C, Metaxoglou K. 2014. Estimation of random-coefficient demand models: two empiricists' perspective. *Review of Economics and Statistics* **96**: 34-59.

Knox S, Viney R, Gu Y, Hole A, Fiebig D, Street D, Haas M, Weisberg E, Bateson D. 2013. The effect of adverse information and positive promotion on women's preferences for prescribed contraceptive products. *Social Sciences and Medicine* **83**: 70-80.

Krink T, Paterlini S, Resti A. 2008. The optimal structure of PD buckets. *Journal of Banking and Finance* **32**: 2275-2286.

Layton D, Brown G. 2000. Heterogeneous preferences regarding global climate change. *Review of Economics and Statistics* **82**: 616-624.

McFadden D, Train K. 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics* **15**: 447-470.

Revelt D, Train K. 1998. Mixed logit with repeated choices: households' choices of appliance efficiency level. *Review of Economics and Statistics* **80**: 647-657.

Sivey P, Scott A, Witt J, Joyce C, Humphreys J. 2012. Junior doctors' preferences for specialty choice. *Journal of Health Economics* **31**: 813-823.

Small K, Winston C, Yan J. 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* **73**: 1367-1382.

Storn R, Price K. 1997. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* **11**: 341-359.

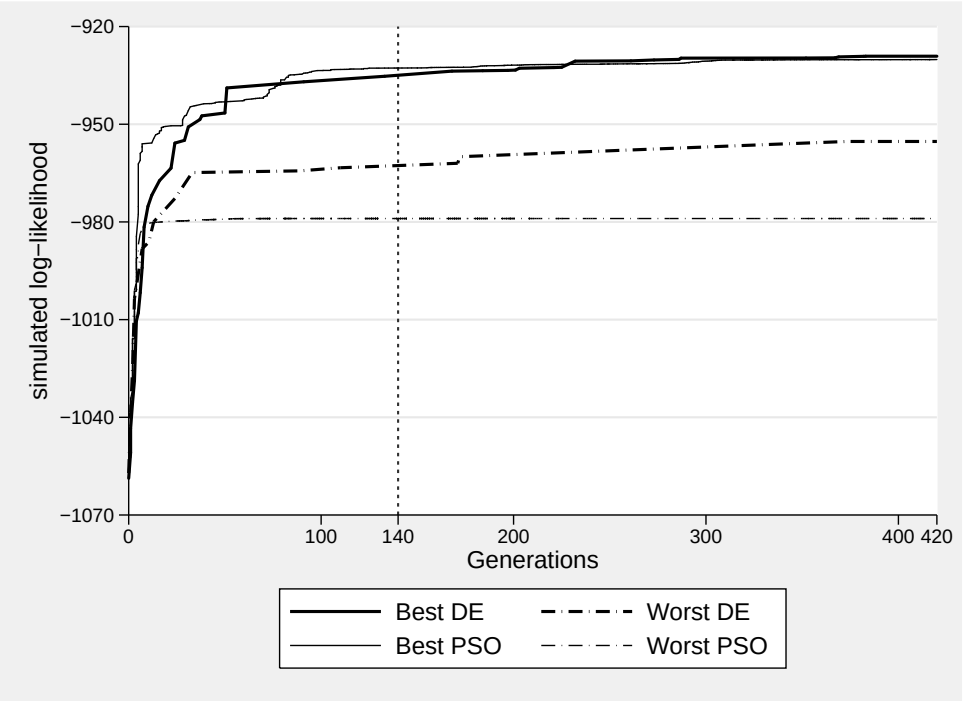
Train K. 2008. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling* **1**: 40-69.

Train K. 2009. *Discrete choice methods with simulation* (2nd ed.). Cambridge University Press: New York.

van Soest A, Das M, Gong X. 2002. A structural labour supply model with flexible preferences. *Journal of Econometrics* **107**: 345-374.

Winschel V, Kratzig M. 2010. Solving, estimating, and selecting nonlinear dynamic models without the curse of dimensionality. *Econometrica* **78**: 803-821.

Figure 1. Pap Smear: selected update paths over generations



Best DE (PSO) and Worst DE (PSO) refer to the DE (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, and also how they would have been updated if the algorithm continued without termination until the 420th generation.

Table 1. Starting values based on special cases of GMNL

	MNL	MIXL	SMNL	GMNL-I	GMNL-II
$\boldsymbol{\beta}$	*Est.*	*Est.*	*Est.*	*Est.*	*Est.*
$\boldsymbol{\sigma}$	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 GMNL.

Table 2. Pap Smear: conventional solutions

Starting point	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
MIXL	-931.065	8.64E-07	-2.06E-14	998.8549
GMNL-II	-931.065	2.92E-05	-4.29E-11	998.9412
SMNL	-932.133	5.46E-08	-5.30E-16	606.9426
GMNL-I	-934.091	1.95E-07	-2.12E-14	4732.774
MNL	-960.317	13.04914	-9.22E-06	1.72E+18[a]

$logL$, 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 λ_{max} and λ_{min} are the largest and smallest eigenvalues of -H. Superscript a indicates that H is ill-conditioned (i.e. $\kappa(H) > 6.7\text{E+}07$).

Table 3. Pap Smear: 10-best DE- and PSO-assisted solutions

A. DE-assisted solutions

F	Cr	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
0.8	0.6	**-925.378**	1.44E-07	-1.78E-15	2033.113
0.8	0.2	**-925.378**	9.77E-07	-1.22E-14	2033.185
0.8	0.2	**-925.378**	6.64E-05	-3.18E-11	2033.347
0.6	0.6	**-925.378**	8.07E-05	-1.35E-10	2033.3
0.6	0.4	**-925.378**	0.0001	-2.21E-10	2033.109
0.8	0.2	**-925.378**	0.000438	-3.75E-09	2033.296
0.8	0.4	**-925.409**	3.92E-07	-3.02E-15	3018.498
0.8	0.4	**-925.409**	9.32E-05	-2.26E-10	3018.577
0.6	0.2	**-926.308**	5.84E-06	-1.37E-11	936.3411
0.6	0.2	**-926.308**	0.00062	-4.03E-08	936.369

B. PSO-assisted solutions

C	D	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
1.5	0.90	**-926.308**	4.74E-06	-9.87E-13	936.3309
1.5	1.00	**-926.308**	0.000377	-4.39E-08	936.2917
2	0.90	**-926.671**	5.08E-08	-3.56E-17	1240.651
1.5	0.75	-932.176	5.49E-05	-6.57E-12	4973.183
0.5	0.90	-932.176	0.000197	-1.04E-10	4969.639
1	0.75	-932.376	7.26E-09	-6.85E-18	2174.327
2	0.50	-932.376	5.33E-08	-9.91E-16	2174.169
0.5	1.00	-932.376	4.00E-07	-1.64E-14	2173.287
1	0.90	-934.091	1.90E-08	-2.92E-16	512.7021
0.5	0.50	-934.091	1.02E-07	-6.73E-16	512.736

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

Table 4. Pap Smear: GMNL parameter estimates

	A. 2nd worst conv.		B. Best conventional		C. Best DE-assisted	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
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 [.]) are reported. "The 2nd worst conv." and "Best conventional" respectively refer to GMNL-I and MIXL/GMNL-II starting point solutions in Table 2. "Best DE-assisted" refers to the first 6 solutions in Table 3. *, **, *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.

Table 5. Pap Smear: simulated WTP distributions

Willingness-to-pay for	p(10)	p(25)	p(50)	p(75)	p(90)
If know doctor:					
2nd worst conv.	-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 conv.	-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 conv.	-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 conv.	-162	-35	21	79	205
Best conventional	-14	19	57	94	129
Best DE-assisted	-73	28	69	135	287

Figures are in $s. Each willingness-to-pay (WTP) distribution has been simulated by making 100,000 draws from the joint density of utility coefficients according to the solutions in Table 4. p(Q) denotes the Q^{th} percentile of the simulated distribution.

Table 6. Pap Smear: predicted choice probabilities

	A	B	C
Base choice probability	0.45	0.57	0.53
Change when test is not due	-0.24	-0.27	-0.26
Change when don't 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 zero	+0.04	+0.05	+0.04

A, B and C are respectively based on 100,000 draws from the joint density of utility coefficients according to "2nd worst conv.", "Best conventional" and "Best DE-assisted" solutions in Table 4. 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.

Table 7. Pizza A: conventional solutions

Starting point	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
GMNL-II	-1361.84	1.45E-05	-4.88E-12	173280.2
MIXL	-1365.17	1.30E-06	-1.58E-12	20000.77
MNL	-1368.44	3.98E-05	-1.51E-09	182428.5
GMNL-I	-1374.45	0.003018	-1.71E-08	3409.606
SMNL	-1395.5	4.60E-06	-8.35E-13	442.66

See notes to Table 2.

Table 8. Pizza A: 10-best DE- and PSO-assisted solutions

A. DE-assisted solutions

F	Cr	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
0.6	0.8	**-1356.8**	18479.79	-51.2587^n	-1.42E+19
0.8	0.4	**-1357.17**	1665.752	-0.16408^n	-3.35E+20
0.6	0.2	**-1357.17**	1887.993	-0.16469^n	-5.19E+20
0.6	0.4	**-1357.17**	298.4716	-0.16521^n	6360351
0.6	0.2	**-1357.53**	0.002223	-2.33E-06	4232703
0.6	0.4	**-1357.64**	0.000897	-4.58E-07	2195944
0.8	0.8	**-1357.64**	0.002647	-1.12E-06	2567936
0.4	0.8	**-1359.03**	0.000146	-4.24E-10	41664.38
0.8	0.8	**-1359.11**	4.60E-06	-5.65E-11	175508.2
0.4	0.2	**-1359.11**	0.001924	-4.17E-09	171543.3

B. PSO-assisted solutions

C	D	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
1.5	0.5	**-1359.3**	0.002958	-1.89E-07	177189.4
1.5	0.75	**-1360**	0.001075	-2.63E-06	184407.5
1	0.9	**-1360.09**	0.000029	-2.40E-08	219779.6
2	0.5	**-1360.29**	6.21E-05	-1.26E-10	32379.59
2	1	**-1360.29**	0.000188	-4.50E-10	32251.78
2	1	**-1360.29**	0.000264	-9.16E-10	32340.78
0.5	1	**-1360.71**	0.00854	-9.80E-09	218317.4
1	0.9	**-1360.76**	1.71E+12	-0.02901^n	.
1	0.5	**-1360.79**	0.000654	-1.94E-08	448585.6
2	0.75	**-1360.9**	0.000794	-1.57E-06	823405.7

$logL$ is in bold if it is greater than the highest $logL$ (GMNL-II starting point) in Table 7. Superscript n indicates that Stata has declared convergence failure since $|g'H^{-1}g|$ exceeds the tolerance criterion (1E-5). See notes to Table 2 for other informaiton.

Table 9. Holiday A: GMNL parameter estimates

	A. Worst conventional		B. Best conventional		C. Best DE-assisted	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
Price	-1.009***	(0.287)	-0.826***	(0.148)	-0.735***	(0.178)
	[0.565***]	(0.159)	[0.921***]	(0.178)	[1.000***]	(0.295)
Overseas destination	0.775***	(0.265)	0.391***	(0.094)	0.411***	(0.088)
	[4.007***]	(1.089)	[3.105***]	(0.602)	[2.922***]	(0.704)
Airline	-0.125	(0.082)	-0.083	(0.070)	-0.062	(0.079)
	[0.183*]	(0.099)	[0.552***]	(0.151)	[0.101**]	(0.047)
Length of stay	1.805***	(0.477)	1.268***	(0.261)	1.354***	(0.320)
	[1.546***]	(0.424)	[1.197***]	(0.233)	[1.306***]	(0.328)
Meal inclusion	1.796***	(0.474)	1.314***	(0.233)	1.478***	(0.403)
	[1.617***]	(0.474)	[1.130***]	(0.217)	[1.564***]	(0.354)
Local tours availability	0.722***	(0.238)	0.552***	(0.128)	0.606***	(0.190)
	[0.636***]	(0.216)	[0.529***]	(0.130)	[0.658***]	(0.178)
Peak season	0.277**	(0.114)	0.143*	(0.073)	0.146**	(0.065)
	[0.913***]	(0.298)	[0.047]	(0.077)	[0.241**]	(0.103)
4-star accommodation	2.817***	(0.763)	2.037***	(0.364)	2.023***	(0.444)
	[2.341***]	(0.627)	[2.061***]	(0.424)	[1.883***]	(0.444)
γ	-0.056**	(0.025)	-0.142***	(0.045)	-0.144***	(0.049)
τ	1.416***	(0.170)	1.205***	(0.132)	1.264***	(0.166)
$logL$	-2523.271		-2513.880		-2490.917	

The worst and best conventional solutions result from MIXL and MNL starting points respectively, and display acceptable convergence diagnostics. The best DE-assisted solution results from configuration `F` = 0.8 and `Cr` = 0.2. Table OA1 and Table OA2 in Online Appendix provide related computational results. See notes to Table 4 for other information.

Table 10. Mobile Phone: GMNL parameter estimates

	A. Worst conventional		B. Best conventional		C. Best DE-assisted	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
No voice comm.	0.045	(0.057)	0.063	(0.055)	0.064	(0.087)
	[0.134]	(0.105)	[0.069]	(0.118)	[0.037]	(0.090)
Voice dialing	0.100*	(0.054)	0.085	(0.058)	0.125	(0.085)
	[0.013]	(0.087)	[0.131]	(0.163)	[0.233***]	(0.067)
Voice operation	-0.155**	(0.063)	-0.136**	(0.060)	-0.099	(0.087)
	[0.167]	(0.103)	[0.098]	(0.143)	[0.254***]	(0.086)
No push to com.	0.056	(0.059)	0.054	(0.058)	0.043	(0.075)
	[0.161**]	(0.081)	[0.005]	(0.083)	[0.213***]	(0.077)
Push to talk	0.059	(0.056)	0.039	(0.060)	0.060	(0.075)
	[0.069]	(0.087)	[0.245***]	(0.080)	[0.206***]	(0.072)
Push to share pics/video	-0.025	(0.061)	-0.021	(0.055)	0.055	(0.074)
	[0.041]	(0.110)	[0.027]	(0.127)	[0.096*]	(0.056)
Personal e-mail	-0.035	(0.071)	-0.059	(0.057)	0.004	(0.082)
	[0.034]	(0.081)	[0.032]	(0.089)	[0.089]	(0.086)
Corporate e-mail	0.080	(0.057)	0.065	(0.056)	0.051	(0.076)
	[0.147]	(0.102)	[0.106]	(0.122)	[0.058]	(0.048)
Both e-mails	-0.060	(0.059)	-0.058	(0.057)	-0.233**	(0.094)
	[0.147*]	(0.087)	[0.085]	(0.085)	[0.085]	(0.054)
WiFi	-0.016	(0.031)	-0.023	(0.031)	-0.059	(0.045)
	[0.006]	(0.051)	[0.012]	(0.051)	[0.008]	(0.044)
USB calbe/cradle	0.095**	(0.043)	0.086**	(0.034)	0.184***	(0.061)
	[0.088]	(0.163)	[0.047]	(0.072)	[0.131***]	(0.048)
Thermometer	0.049	(0.034)	0.063*	(0.037)	0.052	(0.047)
	[0.134]	(0.083)	[0.185***]	(0.066)	[0.151***]	(0.049)
Flashlight	0.063*	(0.034)	0.045	(0.033)	0.083	(0.061)
	[0.029]	(0.061)	[0.075]	(0.069)	[0.003]	(0.062)
Price/100	-1.214***	(0.264)	-1.110***	(0.143)	-1.880***	(0.302)
	[1.096***]	(0.176)	[1.066***]	(0.144)	[0.945***]	(0.210)
ASC for purchase	-0.574***	(0.151)	-0.661***	(0.175)	-1.182***	(0.253)
	[2.542***]	(0.437)	[2.639***]	(0.281)	[4.122***]	(0.907)
γ	-0.108	(0.186)	-0.234**	(0.114)	-0.502***	(0.130)
τ	0.852***	(0.263)	-0.804***	(0.115)	1.715***	(0.162)
logL	-3954.893		-3951.662		-3937.969	

The worst and best conventional solutions result from SMNL and GMNL-I starting points respectively, and display acceptable convergence diagnostics. The best DE-assisted solution results from configurations (`F`,`Cr`) = (0.6,0.2), (0.6,0.4) and (0.8,0.2). Table OA4 and Table OA5 in Online Appendix provide related computational results. See notes to Table 4 for other information.

Table 11. Simulated log-likelihood at 5-best perturbed solutions

	Pap mear	Pizza A	Holiday A	Mobile Phone
DE-assisted	**-925.378**	**-1356.8**	**-2490.92**	**-3937.97**
Conventional	-931.065	-1361.84	-2513.88	-3951.66
Perturbed	-931.065	-1363.2	**-2512.47**	**-3943.01**
Perturbed	-931.065	-1363.99	**-2512.47**	**-3947.98**
Perturbed	-931.065	-1365.21^{n}	**-2512.49**	**-3949.8**
Perturbed	-931.065	-1365.81	**-2512.49**	**-3949.8**
Perturbed	-931.065	-1366.63	**-2512.49**	**-3949.8**

The first two rows report the best DE-assisted and best conventional solutions respectively. The remaining five rows report the five best solutions resulting from 20 random starting points around the best conventional solution: see Section 5.3 for how those starting points have been generated. Boldface indicates a higher simulated log-likelihood than what the best conventional solution has attained. Superscript n indicates that Stata has declared convergence failure since $|g'H^{-1}g|$ exceeds the tolerance criterion (1E-5).

Table 12. 10-best DE-assisted solutions over 20 starts per configuration

A. Pap Smear

F	Cr	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
0.8	0.2	**-924.359**	1.07E-06	-5.45E-15	3741.156
0.8	0.2	**-924.359**	6.55E-07	-9.14E-15	3741.275
0.8	0.2	**-924.359**	2.99E-06	-7.60E-14	3741.364
0.8	0.4	**-924.359**	6.90E-06	-1.35E-13	3740.987
0.8	0.4	**-924.359**	1.27E-05	-3.03E-13	3741.41
0.8	0.2	**-924.359**	2.23E-05	-8.58E-11	3741.099
0.8	0.4	**-924.359**	0.000052	-2.18E-10	3740.992
0.8	0.4	**-924.359**	0.000286	-6.84E-10	3740.659
0.8	0.4	**-924.788**	7.61E-07	-1.34E-14	4324.144
0.8	0.2	**-924.788**	0.000341	-3.07E-10	4322.18

B. Pizza A

F	Cr	$logL$	$\|g\|_\infty$	$g'H^{-1}g$	$\kappa(H)$
0.8	0.4	**-1352.91**	2.39E-04	-7.91E-09	283510.5
0.8	0.2	**-1353.91**	4.01E+02	-1.26E-01^n	1.57E+07
0.8	0.2	**-1354.58**	4.06E-04	-1.44E-08	124508.8
0.8	0.2	**-1355.67**	10.47107	-5.07E-02^n	433304.1
0.8	0.4	**-1355.9**	0.001573	-2.21E-08	601560.5
0.8	0.2	**-1356.84**	7.441122	-6.84E-04^n	-1.31E+19
0.6	0.4	**-1357.17**	12815.33	-4.34E-01^n	-1.17E+17
0.6	0.4	**-1357.17**	1.50E+03	-1.65E-01^n	-7.76E+20
0.6	0.2	**-1357.17**	353.063	-1.65E-01^n	8613358
0.6	0.2	**-1357.17**	503.7171	-0.16532^n	1.82E+07

Information for panel A is the same as in notes to Table 3. Information for panel B is the same as in notes to Table 8. The results in those tables have been obtained by restarting each configuration 3 times. The results in this table have been obtained by restarting each configuration 20 times.

Appendix

Table A1. Pap Smear: all DE- and PSO-assisted solutions in Section 4

A. DE-assisted solutions

F	Cr	Start 1	Start 2	Start 3
0.2	0.2	-937.763	-940.832	-934.466
0.2	0.4	-934.466	-934.466	-935.134
0.2	0.6	-934.091	-934.6	-934.091
0.2	0.8	**-926.384**	-934.091	-940.832
0.4	0.2	-934.603	-934.091	-934.091
0.4	0.4	-934.091	-934.091	-934.814
0.4	0.6	-934.091	-934.091	-934.091
0.4	0.8	**-926.384**	-934.091	-932.376
0.6	0.2	**-926.308**	-934.091	**-926.308**
0.6	0.4	**-926.384**	-934.814	**-925.378**
0.6	0.6	**-926.384**	**-925.378**	**-926.384**
0.6	0.8	**-926.384**	-931.783	-936.769
0.8	0.2	**-925.378**	**-925.378**	**-925.378**
0.8	0.4	-935.455	**-925.409**	**-925.409**
0.8	0.6	**-928.034**	**-925.378**	-937.897
0.8	0.8	-949.114	-934.091	-937.826

B. PSO-assisted solutions

C	D	Start 1	Start 2	Start 3
0.5	0.5	-936.018	-934.091	-936.018
0.5	0.75	-934.603	-937.763	-936.153
0.5	0.9	-934.091	-932.176	-936.043
0.5	1	-934.091	-936.559	-932.376
1	0.5	-934.091	-946.345	-934.091
1	0.75	-932.376	-936.518	-934.49
1	0.9	-934.091	-934.091	-936.018
1	1	-934.091	-934.091	-936.043
1.5	0.5	-936.018	-971.979	-934.091
1.5	0.75	-932.176	-957.961	-942.731
1.5	0.9	-934.603	-934.129	**-926.308**
1.5	1	-938.321	-934.603	**-926.308**
2	0.5	-932.376	-955.963	-935.977
2	0.75	-936.018	-971.979	-959.567
2	0.9	-934.603	-954.306	**-926.671**
2	1	-938.276	-957.961	-938.898

`F`, `Cr`, `C` and `D` indicate tuning parameter values leading to relevant starting points. The simulated log-likelihood at each solution is reported, and is in boldface if it exceeds the highest $logL$ (MIXL starting point) in Table 2.

Table A2. Pizza A: all DE- and PSO-assisted solutions in Section 4

A. DE-assisted solutions

F	Cr	Start 1	Start 2	Start 3
0.2	0.2	-1363.69	-1375.04	-1368.93
0.2	0.4	-1375.04	-1371.05	-1368.93
0.2	0.6	-1366.91	-1374.07	**-1361.37**
0.2	0.8	-1377.3	-1377.17	-1379.4
0.4	0.2	**-1360.9**^{n}	-1366.91	**-1359.11**
0.4	0.4	-1364.58	-1368.29^{n}	-1363.42^{n}
0.4	0.6	-1365.68	**-1360.9**^{n}	**-1360.9**^{n}
0.4	0.8	-1364.76	**-1359.03**	-1363.5
0.6	0.2	-1363.68	**-1357.53**	**-1357.17**^{n}
0.6	0.4	**-1361.8**	**-1357.17**^{n}	**-1357.64**
0.6	0.6	**-1361.8**	**-1360.29**	-1366.91
0.6	0.8	-1362.23	**-1356.8**^{n}	-1363.44
0.8	0.2	-1362.29	**-1360.79**	**-1360.73**
0.8	0.4	**-1357.17**^{n}	-1362.29	-1365.3
0.8	0.6	-1368.43^{n}	-1366.46	**-1360.37**
0.8	0.8	**-1357.64**	**-1359.11**	-1367.92^{n}

B. PSO-assisted solutions

C	D	Start 1	Start 2	Start 3
0.5	0.5	-1375.04	-1380.17	-1369.54
0.5	0.75	-1366.36	-1378.21	-1368.93
0.5	0.9	-1374.82^{n}	-1366.9	-1371.21
0.5	1	**-1360.71**	-1365.36	-1372.35
1	0.5	-1363.03	-1363.48	**-1360.79**
1	0.75	-1363.48	-1388.5	-1372.08
1	0.9	**-1360.09**	**-1360.76**^{n}	-1365.84
1	1	-1370.4	-1365.61	-1377.73
1.5	0.5	-1368.78	**-1359.3**	-1363.48
1.5	0.75	-1364.83	**-1360**	-1375.46
1.5	0.9	-1382.65	-1387.44	-1367.97
1.5	1	-1369.04	-1371.21	-1367.78
2	0.5	-1381.54	-1376.35^{n}	**-1360.29**
2	0.75	-1373.2	-1383.7	**-1360.9**
2	0.9	-1367.69	-1374.84	-1365.83^{n}
2	1	**-1361.65**	**-1360.29**	**-1360.29**

F, Cr, C and D indicate tuning parameter values leading to relevant starting points. The simulated log-likelihood at each solution is reported, and is in bold-face if it exceeds the highest $logL$ (GMNL-II starting point) in Table 7. Superscript n indicates that Stata has declared convergence failure since $|g'H^{-1}g|$ exceeds the tolerance criterion (1E-5).