Generalized indirect inference for discrete choice models[☆],[☆☆]

Marianne Bruins [a],[*], James A. Duffy [b], Michael P. Keane [c], Anthony A. Smith Jr. [d]

[a] Nuffield College and Department of Economics, University of Oxford, United Kingdom
[b] Corpus Christi College and Department of Economics, University of Oxford, United Kingdom
[c] School of Economics, UNSW Business School, Australia
[d] Department of Economics, Yale University and National Bureau of Economic Research, United States

ARTICLE INFO

ABSTRACT

This paper develops and implements a practical simulation-based method for estimating dynamic discrete choice models. The method, which can accommodate lagged dependent variables, serially correlated errors, unobserved variables, and many alternatives, builds on the ideas of indirect inference. The main difficulty in implementing indirect inference in discrete choice models is that the objective surface is a step function, rendering gradient-based optimization methods useless. To overcome this obstacle, this paper shows how to smooth the objective surface. The key idea is to use a smoothed function of the latent utilities as the dependent variable in the auxiliary model. As the smoothing parameter goes to zero, this function delivers the discrete choice implied by the latent utilities, thereby guaranteeing consistency. We establish conditions on the smoothing such that our estimator enjoys the same limiting distribution as the indirect inference estimator, while at the same time ensuring that the smoothing facilitates the convergence of gradient-based optimization methods. A set of Monte Carlo experiments shows that the method is fast, robust, and nearly as efficient as maximum likelihood when the auxiliary model is sufficiently rich.

1. Introduction

Many economic models have the features that (i) given knowledge of the model parameters, it is easy to simulate data from the model, but (ii) estimation of the model parameters is extremely difficult. Models with discrete outcomes or mixed discrete/continuous outcomes commonly fall into this category. A good example is the multinomial probit (MNP), in which an agent chooses from among several discrete alternatives the one with the highest utility. Simulation of data from the model is trivial: simply draw utilities for each alternative, and assign to each agent the alternative that gives them the greatest utility. But estimation of the MNP, via either maximum likelihood (ML) or the method of moments (MOM), is quite difficult.

The source of the difficulty in estimating the MNP, as with many other discrete choice models, is that, from the perspective of the econometrician, the probability an agent chooses a particular alternative is a high-dimensional integral over multiple stochastic terms (unobserved by the econometrician) that affect the utilities the agent assigns to each alternative. These probability expressions must be evaluated many times in order to estimate the model by ML or MOM. For many years econometricians worked on developing fast simulation methods to evaluate choice probabilities in discrete choice models (see Lerman and Manski, 1981). It was only with the development of fast and accurate smooth probability simulators that ML or MOM-based estimation in these models became practical (see McFadden, 1989, and Keane, 1994).

A different approach to inference in discrete choice models is the method of "indirect inference". This approach (see Smith, 1990, 1993; Gourieroux et al., 1993; Gallant and Tauchen, 1996), circumvents the need to construct the choice probabilities implied by the economic model, because it is not based on the likelihood, or on moments based on choice frequencies. Rather, the idea of indirect inference (II) is to choose a statistical model that provides a rich description of the patterns in the data. This descriptive model is estimated on both the actual observed data and on simulated data from the economic model. Letting β denote the vector of parameters of the structural economic model, the II estimator is that $\hat{\beta}$ which makes the simulated data "look like" the actual data — in the sense (defined formally below) that the descriptive statistical model estimated on the simulated data "looks like" that same model estimated on the actual data. (The method of moments is thus a special case of II, in which the descriptive statistical model corresponds to a vector of moments.)

Indirect inference holds out the promise that it should be practical to estimate any economic model from which it is practical to simulate data, even if construction of the likelihood or population moments implied by the model is very difficult or impossible. But this promise has not been fully realized because of limitations in the II procedure itself. It is very difficult to apply II to models that include discrete (or mixed discrete/continuous) outcomes for the following reason: small changes in the structural parameters of such models will, in general, cause the data simulated from the model to change discretely. Such a discrete change causes the parameters of a descriptive model fit to the simulated data to jump discretely, and these discontinuities are inherited by the criterion function minimized by the II estimator.

Thus, given discrete (or discrete/continuous) outcomes, the II estimator cannot be implemented using gradient-based optimization methods. One instead faces the difficult computational task of optimizing a multidimensional step function using much slower derivative-free methods. This is very time-consuming and puts severe constraints on the size of the structural models that can be feasibly estimated. Furthermore, even if estimates can be obtained, one does not have derivatives available for calculating standard errors.

In this paper we propose a "generalized indirect inference" (GII) procedure to address this important problem (Section 3). The key idea is to generalize the original II method by applying two different descriptive statistical models to the simulated and actual data. As long as the two descriptive models share the same vector of pseudo-true parameter values (at least asymptotically), the GII estimator based on minimizing the distance between the two models is consistent, and will enjoy the same asymptotic distribution as the II estimator.

While the GII idea has wider applicability, here we focus on how it can be used to resolve the problem of non-smooth objective functions of II estimators in the case of discrete choice models. Specifically, the model we apply to the simulated data does not fit the discrete outcomes in that data. Rather, it fits a "smoothed" version of the simulated data, in which discrete choice indicators are replaced by smooth functions of the underlying continuous latent variables that determine the model's discrete outcomes. In contrast, the model we apply to the actual data is fit to observed discrete choices (obviously, the underlying latent variables that generate actual agents' observed choices are not seen by the econometrician).

As the latent variables that enter the descriptive model applied to the simulated data are smooth functions of the model parameters, the non-smooth objective function problem is obviously resolved. However, it remains to show that the GII estimator based on minimizing the distance between these two models is consistent and asymptotically normal. We show that, under certain conditions on the parameter regulating the smoothing, the GII estimator has the same limiting distribution as the II estimator, permitting inferences to be drawn in the usual manner (Section 4). Our theoretical analysis goes well beyond merely deriving the limiting distribution of the minimizer of the GII criterion function. Rather, in keeping with computational motivation of this paper, we show how the proposed smoothing facilitates the convergence of standard derivative-based optimizers, providing results for selected line-search and trust-region methods.

Finally, we conduct a set of Monte Carlo experiments to assess the performance of the GII estimator, in terms of bias, efficiency, and computation time, for a range of example models (Section 5). For models of only moderate complexity (i.e. on the order of 10 parameters), GII significantly outperforms conventional II (computed using the downhill simplex), in terms of both computation time and efficiency. We look at some cases where simulated maximum likelihood (SML) is also feasible, and show that efficiency losses relative to SML are small. We also show how judicious choice of the descriptive (or auxiliary) model is very important for the efficiency of the estimator. This is true not only here, but for II more generally.

Proofs of all theoretical results stated in the paper are given in the Supplementary Material.

2. The model

We first describe a class of (dynamic) discrete choice models, which motivate the estimation method developed in this paper. However, the ideas underlying the method could be applied to almost any conceivable model involving discrete outcomes, including models with mixed discrete/continuous outcomes (such as Model 5), and even models in which individuals' choices solve forward-looking dynamic programming problems.

Suppose we have a panel of n individuals, each of whom selects a choice from a set of J discrete alternatives in each of T time periods. (We shall always assume that T is 'small' relative to n; all our asymptotic results hold T as fixed as $n \to \infty$.) Let u_{itj} be the (latent) utility that individual i attaches to alternative j in period t. Without loss of generality, set the utility of alternative J in any period equal to 0. In each period, each individual chooses the alternative with the highest utility. Let y_{itj} be equal to 1 if individual i chooses alternative j in period t and be equal to 0 otherwise: that is,

$$y_{itj} = \mathbf{1}\left\{u_{itj} \geq \max_{k \neq j} u_{itk}\right\} = \begin{cases} 1 & \text{if } u_{itj} \geq \max_{k \neq j} u_{itk}; \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Collect $u_{it} := (u_{it1}, \ldots, u_{it,J-1})$ and $y_{it} := (y_{it1}, \ldots, y_{it,J-1})$. The econometrician observes the choices $\{y_{it}\}_{t=1}^{T}$ but not the latent utilities $\{u_{it}\}_{t=1}^{T}$.

The latent utilities themselves follow a stochastic process

$$u_{it} = f(x_{it}, y_{i,t-1}, \ldots, y_{i,t-l}, \epsilon_{it}; \beta), \qquad t = 1, \ldots, T, \tag{2.2}$$

where x_{it} is a vector of (observed) exogenous variables.[1] For each individual i, the (unobserved) disturbances $\epsilon_{it} := (\epsilon_{it1}, \ldots, \epsilon_{it,J-1})$ follow a Markov process

$$\epsilon_{it} = g(\epsilon_{i,t-1}, \eta_{it}; \beta), \qquad t = 1, \ldots, T \tag{2.3}$$

where $\{\eta_{it}\}_{t=1}^{T}$ is a sequence of (unobserved) i.i.d. random vectors having a specified distribution (which does *not* depend on β), and which are also independent of $\{x_{it}\}_{t=1}^{T}$. The initial values $\{y_{it}\}_{t=1-l}^{0}$ and ϵ_{i0} are fixed exogenously. The functions f and g depend on the structural parameters $\beta \in \mathbf{B} \subset \mathbb{R}^{d_{\beta}}$: the econometrician's problem is thus to estimate β, using data on the outcomes $\{y_{it}\}_{t=1}^{T}$ and exogenous variables $\{x_{it}\}_{t=1}^{T}$, for $i \in \{1, \ldots, n\}$.

We give four examples of discrete choice models that fall within this framework below, and which will be used as test cases for GII. Three of these (Models 1, 2, and 4) can be feasibly estimated using simulated maximum likelihood, allowing us to compare its performance with that of GII. In each, the functions f and g take the linear index form

$$f(x_{it}, y_{i,t-1}, \ldots, y_{i,t-l}, \epsilon_{it}; \beta) = (x_{it}, y_{i,t-1}, \ldots, y_{i,t-l}, \epsilon_{it})' \beta_{ut} \tag{2.4}$$
$$g(\epsilon_{i,t-1}, \eta_{it}; \beta) = (\epsilon_{i,t-1}, \eta_{it})' \beta_{\epsilon t}, \tag{2.5}$$

where β_{ut} and $\beta_{\epsilon t}$ are sub-vectors of β. Our theoretical results will also implicitly rely on this linearity, in the sense that the low-level conditions presented in Appendix B are mostly easily verified for a model in which (2.4)–(2.5) hold.

Model 1. $J = 2$, $T > 1$, and $u_{it} = bx_{it} + \epsilon_{it}$, where x_{it} is a scalar, $\epsilon_{it} = r\epsilon_{i,t-1} + \eta_{it}$, $\eta_{it} \sim_{\text{i.i.d.}} N[0, 1]$, and $\epsilon_{i0} = 0$. This is a two-alternative dynamic probit model with serially correlated errors; it has two unknown parameters b and r.

Model 2. $J = 2$, $T > 1$, and $u_{it} = b_1 x_{it} + b_2 y_{i,t-1} + \epsilon_{it}$, where x_{it} is a scalar and ϵ_{it} follows the same process as in Model 1. The initial value y_{i0} is set equal to 0. This is a two-alternative dynamic probit model with serially correlated errors and a lagged dependent variable; it has three unknown parameters b_1, b_2, and r.

Model 3. Identical to Model 2 except that the econometrician does not observe the first $s < T$ of the individual's choices. Thus there is an "initial conditions" problem (Heckman, 1981).

Model 4. $J = 3$, $T = 1$, and the latent utilities obey:

$$u_{i1} = b_{10} + b_{11}x_{i1} + b_{12}x_{i2} + \eta_{i1}$$
$$u_{i2} = b_{20} + b_{21}x_{i1} + b_{22}x_{i3} + c_1\eta_{i1} + c_2\eta_{i2},$$

where $(\eta_{i1}, \eta_{i2}) \sim_{\text{i.i.d.}} N[0, I_2]$. (Since $T = 1$ in this model, the time subscript has been omitted.) This is a static three-alternative probit model; it has eight unknown parameters $\{b_{1k}\}_{k=0}^{2}$, $\{b_{2k}\}_{k=0}^{2}$, c_1, and c_2.

Models involving a mixture of discrete and continuous outcomes, and indeed mixed discrete/continuous outcomes, are also amenable to the estimation procedure proposed in this paper. One example is provided by the following:

Model 5. A static selection model with two equations; the first determines an individual's wage and the second his/her latent utility from working:

$$w_i = b_{10} + b_{11}x_{1i} + c_1\eta_{1i} + c_2\eta_{2i} \tag{2.6a}$$
$$u_i = b_{20} + b_{21}x_{2i} + b_{22}w_i + \eta_{i2}, \tag{2.6b}$$

---

[1] The estimation method proposed in this paper can also accommodate models in which the latent utilities in any given period depend on lagged values of the latent utilities.

where $(\eta_{i1}, \eta_{i2}) \sim_{\text{i.i.d.}} N[0, I_2]$. The unknown parameters are $\{b_{1k}\}_{k=0}^{1}$, $\{b_{2k}\}_{k=0}^{2}$, c_1, and c_2. Let $e_i := \mathbf{1}\{u_i \geq 0\}$ indicate whether individual i works. The econometrician observes the outcome e_i, but not the latent utility u_i, and observes the wage if and only if the individual works ($e_i = 1$). Thus the observed outcomes are

$$y_i = (e_i, e_i w_i)^{\mathsf{T}}.$$

3. Generalized indirect inference

We propose to estimate the model in Section 2 via a generalization of indirect inference. First, in Section 3.1 we exposit the method of indirect inference as originally formulated. In Section 3.2 we explain the difficulty of applying the original approach to discrete choice models. Section 3.3 presents our generalized indirect inference (GII) estimator that resolves this difficulty. The proposed estimator involves a smoothing procedure, which introduces a bias: we accordingly provide a discussion of how this bias may be reduced (Section 3.4), and suggest a procedure for selecting the parameter that regulates the degree of smoothing (Section 3.5). An overview of related literature appears in Section 3.6.

To simplify the exposition, we shall suppose that the structural model to be estimated falls within the class of models delimited by (2.1)–(2.3); but the reader should bear in mind that the proposed estimation method and the results of this paper do not rely upon this restriction. (A set of low-level conditions that are sufficient for our results, and which encompass such models as Model 5, are given in Appendix B.)

We shall use the following notation. Let $y_i := (y_{i1}^{\mathsf{T}}, \ldots, y_{iT}^{\mathsf{T}})^{\mathsf{T}}$ collect the outcomes from every period ($t \in \{1, \ldots, T\}$) for individual i, and $\mathbf{y} := \{y_i\}_{i=1}^{n}$ denote the aggregate of outcomes for all individuals ($i \in \{1, \ldots, n\}$) in the sample. The quantities x_i, η_i, $\mathbf{x}$ and $\boldsymbol{\eta}$ are constructed analogously from the underlying $\{x_{it}\}$ and $\{\eta_{it}\}$. (Recall that all limits are taken as $n \to \infty$, with T (and later, M) held fixed.)

3.1. Indirect inference

Indirect inference exploits the ease and speed with which one can typically simulate data from even complex structural models. The basic idea is to view both the observed data and the simulated data through the "lens" of a descriptive statistical model – henceforth, the *auxiliary model* – characterized by a set of d_θ auxiliary parameters θ. The $d_\beta \leq d_\theta$ structural parameters β are then chosen so as to make the observed data and the simulated data look similar when viewed through this lens.

To formalize these ideas, assume the observed choices $\mathbf{y}$ are generated by the structural discrete choice model (2.1)–(2.3), for a given value β_0 of the structural parameters (i.e. the structural model is "correctly specified"). An auxiliary model can be estimated using the observed data to obtain parameter estimates $\hat{\theta}_n$. Formally, $\hat{\theta}_n$ solves:

$$\hat{\theta}_n := \underset{\theta \in \Theta}{\text{argmax}} \, \mathcal{L}_n(\mathbf{y}, \mathbf{x}; \theta) = \underset{\theta \in \Theta}{\text{argmax}} \, \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i; \theta), \tag{3.1}$$

where $\mathcal{L}_n(\mathbf{y}, \mathbf{x}; \theta)$ is the average log-likelihood function (or more generally, some statistical criterion function) associated with the auxiliary model.

Let $\boldsymbol{\eta}^m := \{\eta_i^m\}_{i=1}^{n}$ denote a set of simulated draws for the values of the unobservable components of the model, with these draws being independent across $m \in \{1, \ldots, M\}$. Then given $\mathbf{x}$ and a parameter vector β, the structural model can be used to generate M corresponding sets of simulated choices, $\mathbf{y}^m(\beta) := \{y_i^m(\beta)\}_{i=1}^{n}$. (Note that the same values of $\mathbf{x}$ and $\boldsymbol{\eta}^m$ are used for all β.) Estimating the auxiliary model on the mth simulated dataset yields

$$\hat{\theta}_n^m(\beta) := \underset{\theta \in \Theta}{\text{argmax}} \, \mathcal{L}_n(\mathbf{y}^m(\beta), \mathbf{x}; \theta). \tag{3.2}$$

The average of these estimates will be denoted by

$$\bar{\theta}_n(\beta) := \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_n^m(\beta). \tag{3.3}$$

Under appropriate regularity conditions, as the observed sample size n grows large (holding M and T fixed), $\bar{\theta}_n(\beta)$ converges uniformly in probability to a non-stochastic limit $\theta(\beta)$, which Gourieroux et al. (1993) term the *binding function*.

Loosely speaking, indirect inference generates an estimate $\hat{\beta}_n$ of the structural parameters by choosing β so as to make $\hat{\theta}_n$ and $\bar{\theta}_n(\beta)$ as close as possible, with consistency following from $\hat{\theta}_n$ and $\bar{\theta}_n(\beta_0)$ both converging to the same pseudo-true value $\theta_0 := \theta(\beta_0)$. To implement the estimator we require a metric for the distance between $\hat{\theta}_n$ and $\bar{\theta}_n(\beta)$. There are three approaches to choosing such a metric, analogous to the three classical approaches to hypothesis testing: the Wald, likelihood ratio (LR), and Lagrange multiplier (LM) approaches.[2]

---

[2] This nomenclature is due to Eric Renault. The Wald and LR approaches were first proposed in Smith (1990, 1993) and later extended by Gourieroux et al. (1993). The LM approach was first proposed in Gallant and Tauchen (1996).

The Wald approach to indirect inference chooses β to minimize the weighted distance between $\overline{\theta}_n(\beta)$ and $\hat{\theta}_n$,

$$Q_n^{\mathrm{W}}(\beta) := \|\overline{\theta}_n(\beta) - \hat{\theta}_n\|_{W_n}^2, \tag{3.4}$$

where $\|b\|_A^2 := b^\mathsf{T} A b$, and W_n is a sequence of positive-definite weight matrices.

The LR approach forms a metric implicitly by using the average log-likelihood $\mathcal{L}_n(\mathbf{y}, \mathbf{x}; \theta)$ associated with the auxiliary model. In particular, it seeks to minimize

$$Q_n^{\mathrm{LR}}(\beta) := -\mathcal{L}_n(\mathbf{y}, \mathbf{x}; \overline{\theta}_n(\beta)) = -\frac{1}{n}\sum_{i=1}^n \ell(y_i, x_i; \overline{\theta}_n(\beta)). \tag{3.5}$$

Finally, the LM approach does not work directly with the estimated auxiliary parameters $\overline{\theta}_n(\beta)$ but instead uses the score vector associated with the auxiliary model.[3] Given the estimated auxiliary model parameters $\hat{\theta}_n$ from the observed data, the score vector is evaluated using each of the M simulated datasets. The LM estimator then minimizes a weighted norm of the average score vector across these datasets,

$$Q_n^{\mathrm{LM}}(\beta) := \left\| \frac{1}{M}\sum_{m=1}^M \dot{\mathcal{L}}_n(\mathbf{y}^m(\beta), \mathbf{x}; \hat{\theta}_n) \right\|_{V_n}^2, \tag{3.6}$$

where $\dot{\mathcal{L}}_n$ denotes the gradient of $\mathcal{L}_n$ with respect to θ, and V_n is a sequence of positive-definite weight matrices.

All three approaches yield consistent and asymptotically normal estimates of β_0, and are first-order asymptotically equivalent in the exactly identified case in which $d_\beta = d_\theta$. In the over-identified case, when the weight matrices W_n and V_n are chosen optimally (in the sense of minimizing asymptotic variance) both the Wald and LM estimators are more efficient than the LR estimator. However, if the likelihood of the auxiliary model is correctly specified, then all three (optimally weighted) estimators are asymptotically equivalent.

3.2. Indirect inference for discrete choice models

Discontinuities arise naturally when applying indirect inference to discrete choice models because any simulated choice $y_{itj}^m(\beta)$ is a step function of β (holding fixed the random draws η^m). Consequently, the sample binding function $\overline{\theta}_n(\beta)$ is discontinuous in β, and these discontinuities are inherited by the II criterion functions (3.4)–(3.6).

Thus in such models, it is difficult to ensure that a gradient-based optimization procedure will converge to an (approximate) maximizer of the II criterion function. One must therefore instead rely on derivative-free methods (such as the Nelder–Mead simplex method); random search algorithms (such as simulated annealing); or abandon optimization altogether, and instead implement a Laplace-type estimator, via Markov Chain Monte Carlo (MCMC; see Chernozhukov and Hong, 2003). But convergence of derivative-free methods is often very slow; while verifying the convergence of MCMC routines is often a challenging task. Thus, the non-smoothness of the criterion functions that define II estimators render them very difficult to use in the case of discrete data.

Despite the difficulties in applying II to discrete choice models, the appeal of the II approach has led some authors to push ahead and apply it nonetheless. Some notable papers that apply II by optimizing non-smooth objective functions are Magnac et al. (1995), An and Liu (2000), Nagypál (2007), Eisenhauer et al. (2015), Li and Zhang (2015) and Skira (2015). Our work aims to make it much easier to apply II in these and related contexts.

3.3. A smoothed estimator (GII)

Here we propose a generalization of indirect inference that is far more practical in the context of discrete outcomes. The fundamental idea is that the estimation procedures applied to the observed and simulated datasets need not be identical, provided that both are consistent for the same binding function. (Genton and Ronchetti, 2003, use a similar insight to develop robust estimation procedures in the context of indirect inference.) We exploit this idea to smooth the function $\overline{\theta}_n(\beta)$, obviating the need to optimize a step function when using indirect inference to estimate a discrete choice model.

Returning to the framework of Section 2, let $u_{itj}^m(\beta)$ denote the latent utility that individual i attaches to alternative $j \in \{1, \ldots, J - 1\}$ in period t of the mth simulated dataset, given structural parameters β (recall that the utility of the Jth alternative is normalized to 0). Rather than use the simulated choice

$$y_{itj}^m(\beta) := \mathbf{1}\left\{u_{itj}^m(\beta) \geq \max_{k \neq j} u_{itk}^m(\beta)\right\} = \prod_{k \neq j} \mathbf{1}\left\{u_{itj}^m(\beta) - u_{itk}^m(\beta) \geq 0\right\}$$

when computing $\overline{\theta}_n(\beta)$, we propose to replace it by the following smooth function of the latent utilities,

$$y_{itj}^m(\beta, \lambda) := \mathcal{K}_\lambda[u_{itj}^m(\beta) - u_{it1}^m(\beta), \ \ldots, \ u_{itj}^m(\beta) - u_{itJ}^m(\beta)], \tag{3.7}$$

---

[3] When the LM approach is implemented using an auxiliary model that is (nearly) correctly specified in the sense that it provides a (nearly) correct statistical description of the observed data, Gallant and Tauchen (1996) refer to this approach as efficient method of moments (EMM).

where $\mathcal{K} : \mathbb{R}^{J-1} \to \mathbb{R}$ is a smooth, mean-zero multivariate cdf, and $\mathcal{K}_\lambda(v) := \mathcal{K}(\lambda^{-1}v)$.[4] If the choice utilities $\{u^m_{itj}(\beta)\}^J_{j=1}$ are distinct – which in these models occurs with probability one – then as the smoothing parameter λ goes to 0, $y^m_{itj}(\beta, \lambda)$ converges to $y^m_{itj}(\beta, 0) = y^m_{itj}(\beta)$.[5]

Let $y^m_i(\beta, \lambda)$ and $\boldsymbol{y}^m(\beta, \lambda)$ be constructed from $\{y^m_{itj}(\beta, \lambda)\}$, in the same manner as y_i and $\boldsymbol{y}$ were from $\{y_{itj}\}$ (see p. 5). Defining

$$\overline{\theta}_n(\beta, \lambda) := \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^m_n(\beta, \lambda) \tag{3.8}$$

where

$$\hat{\theta}^m_n(\beta, \lambda) := \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{L}_n(\boldsymbol{y}^m(\beta, \lambda), \boldsymbol{x}; \theta), \tag{3.9}$$

we may regard $\overline{\theta}_n(\beta, \lambda)$ as providing a smoothed estimate of $\theta(\underline{\beta})$, for which it is consistent so long as $\lambda = \lambda_n \to 0$ as $n \to \infty$. Accordingly, an indirect inference estimator based on $\overline{\theta}_n(\beta, \lambda_n)$, which we shall henceforth term the *generalized indirect inference* (GII) estimator, ought to be consistent for β_0.[6]

Each of the three approaches to indirect inference can be generalized simply by replacing each simulated choice $y^m_{itj}(\beta)$ with its smoothed counterpart $y^m_{itj}(\beta, \lambda_n)$. For the Wald and LR estimators, this entails using the smoothed sample binding function $\overline{\theta}_n(\beta, \lambda_n)$ in place of the unsmoothed estimate $\overline{\theta}_n(\beta)$. (See Section 3.4 for the exact forms of the criterion functions.)

The GII approach was first suggested in an unpublished manuscript by Keane and Smith (2003), but they did not derive the asymptotic properties of the estimator. Despite this, GII has proven to be popular in practice, and has already been applied in a number of papers, such as Gan and Gong (2007), Cassidy (2012), Altonji et al. (2013), Morten (2013), Ypma (2013), Lopez-Mayan (2014) and Lopez Garcia (2015). Given the growing popularity of the method, a careful analysis of its asymptotic properties is obviously needed.

3.4. Bias reduction

GII inherits the consistency of the II estimator, provided that $\lambda_n \to 0$ as $n \to \infty$. However, as smoothing necessarily imparts a bias to the sample binding function $\overline{\theta}_n(\beta, \lambda_n)$, and thence to the GII estimator, we need λ_n to shrink to zero at a sufficiently fast rate if GII is to enjoy the same limiting *distribution* as the unsmoothed II estimator. In general, the smoothing imparts a bias that is of the order $\|\theta(\beta_0, \lambda) - \theta(\beta_0, 0)\| = O(\lambda)$; hence this will be dominated by the estimator variance only if $n^{1/2}\lambda_n \to 0$. On the other hand, if $\lambda_n \to 0$ too rapidly, derivatives of the GII criterion function will become highly irregular, impeding the ability of derivative-based optimization routines to locate the minimum. (For precisely what this may entail for λ_n, see Proposition 4.1.) To allow these conflicting requirements on the smoothing sequence to be more easily reconciled, we introduce the following refinements of the proposed estimator.[7]

3.4.1. Jackknifing

One well-known approach to bias reduction is Richardson extrapolation – commonly referred to as "jackknifing" in the statistics literature – which may be applied directly to the smoothed sample binding function. Provided that the *population* binding function is sufficiently smooth, a Taylor series expansion gives

$$\theta_l(\beta, \lambda) = \theta_l(\beta, 0) + \sum_{r=1}^{s} \alpha_{rl}(\beta)\lambda^r + o(\lambda^s) \qquad \alpha_{rl}(\beta) := \frac{1}{r!} \frac{\partial \theta_l(\beta, \lambda)}{\partial \lambda}\bigg|_{\lambda=0},$$

as $\lambda \to 0$, for $l \in \{1, \ldots, d_\theta\}$. Then, for a chosen $\delta \in (0, 1)$, we have the first-order extrapolation,

$$\theta^1_l(\beta, \lambda) := \frac{\theta_l(\beta, \delta\lambda) - \delta\theta_l(\beta, \lambda)}{1 - \delta} = \theta_l(\beta, 0) + \delta \sum_{r=2}^{s} (\delta^{r-1} - 1)\alpha_{rl}(\beta)\lambda^r + o(\lambda^s),$$

---

[4] Keane and Smith (2003) suggested using the multivariate logistic cdf, $\mathcal{K}(v) := 1/(1 + \sum_{j=1}^{J-1} e^{-v_j})$, and this is used in the simulation exercises presented in Section 5; but this has no particular advantages over other possible choices.

[5] For models in which current utility depends on past *outcomes* (as distinct from past utilities), the performance of GII may be improved through a further refinement to the basic smoothing procedure outlined here: see Appendix A.1 for details.

[6] To implement this procedure in Model 5, which has mixed discrete/continuous outcomes, we would set

$$y_i(\beta, \lambda) = [\mathcal{K}_\lambda(u_i(\beta)), \mathcal{K}_\lambda(u_i(\beta))w_i(\beta)]^\top$$

where $(u_i(\beta), w_i(\beta))$ are as in (2.6).

[7] Independently of our own work on this problem, Kristensen and Salanié (2016) propose two methods for bias reduction that are applicable to a wide range of 'approximate' estimators (which includes simulation-based estimators): one based on a kind of jackknifing procedure, and the other on a Newton–Raphson iteration. While the latter is evidently closely related to what we propose in Section 3.4.2, the former is somewhat different from our own jackknifing procedure (Section 3.4.1). In particular, the structure of the indirect inference problem allows us to apply jackknifing directly to the binding function (so that the bias correction can be 'built in' to the GII criterion function), whereas in our setting, Kristensen and Salanié's (2016) jackknife would have to be implemented by taking an appropriate linear combination of GII estimators computed using different values of λ.

for every $l \in \{1, \ldots, d_\theta\}$. By an iterative process, for $k \leq s - 1$ we can construct a kth order extrapolation of the binding function, which satisfies

$$\theta^k(\beta, \lambda) := \sum_{r=0}^{k} \gamma_{rk} \theta(\beta, \delta^r \lambda) = \theta(\beta, 0) + O(\lambda^{k+1}), \tag{3.10}$$

where the weights $\{\gamma_{rk}\}_{r=0}^{k}$ (which can be negative) satisfy $\sum_{r=0}^{k} \gamma_{rk} = 1$, and may be calculated using Algorithm 1.3.1 in Sidi (2003). It is immediately apparent that the kth order jackknifed sample binding function,

$$\overline{\theta}_n^k(\beta, \lambda_n) := \sum_{r=0}^{k} \gamma_{rk} \overline{\theta}_n(\beta, \delta^r \lambda_n) \tag{3.11}$$

will enjoy an asymptotic bias of order $O_p(\lambda_n^{k+1})$, whence only $n^{1/2} \lambda_n^{k+1} = o_p(1)$ is necessary for the bias to be asymptotically negligible (relative to the estimator variance).[8]

Jackknifed GII estimators of order $k \in \mathbb{N}_0$ may now be defined as the minimizers of:

$$Q_{nk}^e(\beta, \lambda_n) := \begin{cases} \|\overline{\theta}_n^k(\beta, \lambda_n) - \hat{\theta}_n\|_{W_n}^2 & \text{if } e = \text{W} \\ -\mathcal{L}_n(\mathbf{y}; \mathbf{x}, \overline{\theta}_n^k(\beta, \lambda_n)) & \text{if } e = \text{LR} \\ \|\frac{1}{M} \sum_{m=1}^{M} \dot{\mathcal{L}}_n^{mk}(\beta, \lambda_n; \hat{\theta}_n)\|_{V_n}^2 & \text{if } e = \text{LM} \end{cases} \tag{3.12}$$

where $\dot{\mathcal{L}}_n^{mk}(\beta, \lambda; \hat{\theta}_n) := \sum_{r=0}^{k} \gamma_{rk} \dot{\mathcal{L}}_n(\mathbf{y}^m(\beta, \lambda), \mathbf{x}; \hat{\theta}_n)$ denotes the jackknifed score function; the un-jackknifed estimators may be recovered by taking $k = 0$.

3.4.2. Newton–Raphson iterations

By allowing the number of simulations M to increase with the sample size, we can accelerate the rate at which $\overline{\theta}_n^k$ converges to the binding function; the 'effective sample size' used to compute the derivatives of $\overline{\theta}_n^k$ being nM. Larger values of M thus permit us to choose smaller values of λ. Since the evaluation of Q_{nk} is potentially costly when M is very large, a sensible approach is to first minimize Q_{nk} using a very small value of M (e.g. $M = 1$) and large $\lambda = \lambda^{(0)}$, to produce an initial estimate $\hat{\beta}^{(0)}$. One could then increase M to an appropriately large value $M^{(1)}$, allowing λ to be reduced to some $\lambda^{(1)} < \lambda^{(0)}$, and then compute a new estimate by taking (at least) one Newton–Raphson step, applied to the new criterion: that is, by computing

$$\hat{\beta}^{\text{NR}} := \hat{\beta}^{(0)} - [\partial_\beta^2 Q_{nk}(\hat{\beta}^{(0)}, \lambda^{(1)}; M^{(1)})]^{-1} \partial_\beta Q_{nk}(\hat{\beta}^{(0)}, \lambda^{(1)}; M^{(1)}),$$

where $\partial_\beta f$ and $\partial_\beta^2 f$ respectively denote the gradient and Hessian of $f : \text{B} \to \mathbb{R}$.

Although a rigorous analysis of this 'Newton–Raphson' estimator is beyond the scope of this paper (we assume that M is *fixed* throughout Section 4), we do study its performance in the simulation exercises of Section 5. These suggest that its performance is comparable to that of the jackknifed estimator; there is apparently little difference between the two procedures in this respect.

3.5. Smoothing parameter selection

Regarding the choice of λ in practice, we have the following suggestions. We recommend selecting an initial value of $\lambda = \lambda^{(0)}$ that is sufficiently large to ensure the convergence of derivative-based optimizers, when applied to $\beta \mapsto Q_{nk}(\beta, \lambda)$. Such a λ could come from experimentation, or be generated by the automated selection procedure described in Appendix A.2. Because of the possibility of multiple local optima, one would generally optimize Q_{nk} from many possible starting values, before finally settling on at an initial estimate $\hat{\beta}_n(\lambda^{(0)})$. (To save on computational time, a very small value of M should be used in this phase, possibly even $M = 1$.)

In the next phase of the optimization, we would choose a smaller value of $\lambda = \lambda^{(1)} < \lambda^{(0)}$, and reoptimize $Q_{nk}(\cdot, \lambda^{(1)})$, using $\hat{\beta}_n(\lambda^{(0)})$ as a starting value, to obtain a new estimate $\hat{\beta}_n(\lambda^{(1)})$. This process of adjusting λ downwards and reoptimizing the resulting criterion function could be iterated, with say $\lambda^{(i)} := \rho \lambda^{(i-1)}$ for some $\rho \in (0, 1)$ on the ith iteration. We then need to decide at which point these iterations should terminate. One possibility is to terminate once the associated change in the parameter estimates falls below some pre-specified tolerance, that is, once

$$\|\hat{\beta}_n(\lambda^{(i)}) - \hat{\beta}_n(\lambda^{(i-1)})\| < \epsilon \tag{3.13}$$

---

[8] In the case where the auxiliary model estimator can be written as $\hat{\theta}_n^m(\beta, \lambda) = g(T_n^m(\beta, \lambda))$, for some transformation g of a vector T_n^m of sufficient statistics (i.e. statistics that are sufficient for the *auxiliary* model estimator), jackknifing could be applied directly to these statistics. Thus, if we were to set $\hat{\theta}_n^{mk}(\beta, \lambda_n) := g(\sum_{r=0}^{k} \gamma_{rk} T_n^m(\beta, \lambda_n))$, then $\frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_n^{mk}(\beta, \lambda_n)$ would also have an asymptotic bias of order $O_p(\lambda_n^{k+1})$. This approach may have computational advantages if the transformation g is relatively costly to compute (e.g. when it involves matrix inversion).

where ϵ is of the order 10^{-6}, or smaller. Alternatively, since the bias of the GII estimator is ultimately inherited from that of the (smoothed) sample binding function $\overline{\theta}_n(\beta, \lambda)$, we instead could proceed to a point where the estimated (total) bias in $\overline{\theta}_n(\beta, \lambda)$ falls below some pre-specified fraction δ of its overall standard error, where both of these quantities are evaluated at the current estimates. Details of how to compute these quantities via simulation are provided in Appendix A.2.

3.6. Related literature

Our approach to smoothing in a discrete choice model bears a superficial resemblance to that used by Horowitz (1992) to develop a smoothed version of Manski's (1985) maximum score estimator for a binary response model. As here, the smooth version of maximum score is constructed by replacing discontinuous indicators with smooth cdfs in the sample criterion function.

However, there is a fundamental difference in the statistical properties of the minimization problems solved by Manski's estimator, and the (unsmoothed) indirect inference estimator. Specifically, $n^{-1/2}$-consistent estimators are available for the *unsmoothed* problem considered in this paper (see Theorem 4.1, or Pakes and Pollard, 1989); whereas, in the case of Manski's (1985) maximum score estimator, only $n^{-1/3}$-consistency is obtained without smoothing (see Kim and Pollard, 1990), and smoothing yields an estimator with an improved rate of convergence.

A potentially more relevant analogue for the present paper is smoothed quantile regression. This originates with Horowitz's (1998) work on the smoothed least absolute deviation estimator, extended to more general quantile regression and quantile-IV models by Whang (2006), Otsu (2008) and Kaplan and Sun (2012). The latter papers do not smooth the criterion function, but rather the estimating equations (approximate first-order conditions) that equivalently define the estimator. These first-order conditions involve indicator-type discontinuities like those in our problem, smoothed in the same way. Insofar as the problem of solving the estimating equations is analogous to the minimum-distance problem solved by the II estimator, the effects of smoothing are similar: in each case smoothing (if done appropriately) affects neither the rate of convergence nor the limiting distribution of the estimator, relative to its unsmoothed counterpart.

The motivation for smoothing in the quantile regression case involves the potential for higher-order asymptotic improvements.[9] In contrast, in the present setting, which involves structural models of possibly great complexity, the potential for higher-order improvements is limited.[10] The key motivation for smoothing in our case is computational. Accordingly, Section 4.3 is devoted to providing a theoretical foundation for our claim that smoothing facilitates the convergence of standard derivative-based procedures that are widely used to solve (smooth) optimization problems in practice.

For the class of models considered in this paper, two leading alternative estimation methods that might be considered are simulated maximum likelihood (SML) in conjunction with the Geweke, Hajivassiliou and Keane (GHK) smooth probability simulator (see Section 4 in Geweke and Keane, 2001), and the nonparametric simulated maximum likelihood (NPSML) estimator (Diggle and Gratton, 1984; Fermanian and Salanié, 2004; Kristensen and Shin, 2012). However, the GHK simulator can only be computed in models possessing a special structure – which is true for Models 1, 2 and 4, but *not* for Model 3 – while in models that involve a mixture of discrete and continuous outcomes, NPSML may require the calculation of rather high-dimensional kernel density estimates in order to construct the likelihood, the accuracy of which may require simulating the model a prohibitively large number of times.

Finally, an alternative approach to smoothing the II estimator is importance sampling, as in Keane and Sauer (2010) and Sauer and Taber (2013). The basic idea is to simulate data from the structural model only once (at the initial estimate of β). One holds these simulated data fixed as one iterates. Given an updated estimate of β, one re-weights the original simulated data points, so those initial simulations that are more (less) likely under the new β (than under the initial β) get more (less) weight in forming the updated objective function.

In our view the GII and importance sampling approaches both have virtues. The main limitation of the importance sampling approach is that in many models the importance sample weights may themselves be computationally difficult to construct. Keane and Sauer (2010), when working with models similar to those in Section 2, assume that all variables are measured with error, which gives a structure that implies very simple weights. In many contexts such a measurement error assumption may be perfectly sensible. But the GII method can be applied directly to the models of Section 2 without adding any auxiliary assumptions (or parameters).

4. Asymptotic and computational properties

This section presents our results on the asymptotic and computational properties of the GII estimator. These follow from the high-level and regularity conditions stated below (see Assumptions H and R). We shall only provide results for the Wald

---

[9] While potential computational benefits have been noted in passing, we are not aware of any attempt to demonstrate these formally, in the manner of Theorem 4.3.

[10] This is particularly evident when the auxiliary model consists of a system of regression equations, as per Section 5.2.1. For while smoothing does indeed reduce the variability of the simulated (discrete) outcomes $y_{it}^m(\beta, \lambda)$, this may *increase* the variance with which some parameters of the auxiliary model are estimated, if y_{it} appears as a regressor in that model: as will be the case for Models 2 and 3 (see Sections 5.2.3 and 5.2.4). (Note that any such increase, while certainly possible, is of only second-order importance, and disappears as $\lambda_n \to 0$.)

and LR estimators, when these are jackknifed as per (3.11); but it would of course be possible to extend our arguments to cover the LM estimator. All asymptotic results concern limits taken as $n \to \infty$, for fixed values of M and T; and for this reason, we shall generally suppress the t subscript throughout.

Our high-level conditions could be satisfied by a very broad class of models: but verifying that this is so presents a considerable challenge. Appendix B.1 accordingly provides some low-level conditions (Assumption L) that are sufficient for our high-level conditions, and thus for the results of this section. (The statement and discussion of these low-level conditions has been deferred to the Appendix, so to allow our main results to be presented without imposing a further notational burden upon the reader.) Appendix B.2 verifies that our low-level conditions are satisfied by each of Models 1–5, when the auxiliary model is a Gaussian system of seemingly unrelated regressions (see e.g. Section 10.2 in Greene, 2008), which may be estimated either by ordinary least squares, (feasible) generalized least squares, or maximum likelihood. This is a particularly convenient and flexible auxiliary model, and is used throughout the simulation exercises of Section 5.

4.1. High-level conditions

Recall that $\boldsymbol{y}^m(\beta, \lambda)$ denotes the complete set of smoothed outcomes generated by the mth simulation of the model, for $m \in \{1, \ldots, M\}$. If the model is correctly specified, as per R1, we may regard the observed outcomes $\boldsymbol{y}$ as having been generated by a '0th' simulation of the model (without smoothing), denoted $\boldsymbol{y}^0(\beta_0, 0)$. We shall accordingly allow the index m to range over $\{0, 1, \ldots, M\}$, which will permit some of our assumptions and results to be more concisely stated. This device also allows us to write $\hat{\theta}_n = \hat{\theta}_n^0(\beta_0, 0)$, where $\hat{\theta}_n$ denotes the auxiliary model estimate yielded by the observed outcomes.

Let int $\mathscr{X}$ denote the interior of $\mathscr{X} \subseteq \mathbb{R}^k$. The smoothing parameter sequence $\{\lambda_n\}$ below may be sample-dependent, i.e. it is *not* assumed to be a "given" deterministic sequence. Let Λ denote the set of allowable values for λ_n, this may be taken to be $[0, 1]$ without loss of generality.

Assumption R (*Regularity Conditions*)**.**

R1 The structural model is correctly specified: $\boldsymbol{y} = \boldsymbol{y}^0(\beta_0, 0)$ for some $\beta_0 \in$ int B;
R2 $\theta_0 := \theta(\beta_0, 0) \in$ int Θ;
R3 the binding function $\theta(\beta, \lambda)$ is single-valued, and is $(k_0 + 1)$-times differentiable in β for all $(\beta, \lambda) \in (\text{int B}) \times \Lambda$;
R4 $\beta \mapsto \theta(\beta, 0)$ is injective;
R5 the order $k \in \{0, 1, \ldots, k_0\}$ of the jackknifing is chosen such that $n^{1/2}\lambda_n^{k+1} = o_p(1)$; and
R6 $W_n \xrightarrow{p} W$, for some positive definite W.

Remark 4.1. R4 formalizes the requirement that the auxiliary model be "sufficiently rich" to identify the parameters of the structural model; $d_\theta \geq d_\beta$ is essentially necessary for this to be satisfied. R5 ensures that, in conjunction with the choice of λ_n, the order of the jackknifing is such as to ensure that the bias introduced by the smoothing is asymptotically negligible.

Define $\mathcal{L}_n(\theta) := \mathcal{L}_n(\boldsymbol{y}, \boldsymbol{x}; \theta)$, $\mathcal{L}(\theta) := \mathbb{E}\mathcal{L}_n(\theta)$ and

$$\ell_i^m(\beta, \lambda; \theta) := \ell(y_i^m(\beta, \lambda), x_i; \theta).$$

$\dot{\ell}_i^m$ and $\ddot{\mathcal{L}}_n$ respectively denote the gradient of ℓ_i^m and the Hessian of $\mathcal{L}_n$ with respect to θ. Let

$$\phi_n^m := \frac{1}{n^{1/2}} \sum_{i=1}^n \dot{\ell}_i^m(\beta_0, 0; \theta_0)$$

denote the standardized score vector for the mth simulation (at $\lambda = 0$). $\partial_\beta f$ denotes the gradient of $f : \text{B} \to \mathbb{R}^d$ (the transpose of the Jacobian), and $\partial_\beta^2 f$ the Hessian; see Section 6.3 of Magnus and Neudecker, 2007, for a definition of the latter when $d \geq 2$.

Assumption H (*High-level Conditions*)**.**

H1 $\mathcal{L}_n$ is twice continuously differentiable on int Θ;
H2 $[\mathcal{L}_n, \dot{\mathcal{L}}_n, \ddot{\mathcal{L}}_n](\theta) \xrightarrow{p} [\mathcal{L}, \dot{\mathcal{L}}, \ddot{\mathcal{L}}](\theta)$, and

$$\frac{1}{n} \sum_{i=1}^n \dot{\ell}_i^{m_1}(\beta_1, \lambda_1; \theta_1)\dot{\ell}_i^{m_2}(\beta_2, \lambda_2; \theta_2)^\mathsf{T} \xrightarrow{p} \mathbb{E}\dot{\ell}_i^{m_1}(\beta_1, \lambda_1; \theta_1)\dot{\ell}_i^{m_2}(\beta_2, \lambda_2; \theta_2)^\mathsf{T}$$

uniformly on B $\times \Lambda$ and compact subsets of int Θ, for every $m_1, m_2 \in \{0, 1, \ldots, M\}$;
H3 for any random sequences $\beta_n = \beta_0 + o_p(1)$ and $\lambda_n = o_p(1)$,

$$n^{1/2}[\hat{\theta}_n^m(\beta_n, \lambda_n) - \theta(\beta_n, \lambda_n)] = -H^{-1}\phi_n^m + o_p(1) \tag{4.1}$$

where $H := \mathbb{E}\ddot{\mathcal{L}}_n(\theta_0) = \ddot{\mathcal{L}}(\theta_0)$, for every $m \in \{0, 1, \ldots, M\}$;

H4 $\phi_n^m \rightsquigarrow \phi^m$ jointly for $m \in \{0, 1, \ldots, M\}$, where $\{\phi^m\}_{m=0}^M$ is Gaussian with mean zero and

$$\Sigma := \mathbb{E}\phi^{m_1}\phi^{m_1\mathsf{T}} = \mathbb{E}\phi_n^{m_1}\phi_n^{m_1\mathsf{T}} \qquad R := \mathbb{E}\phi^{m_1}\phi^{m_2\mathsf{T}} = \mathbb{E}\phi_n^{m_1}\phi_n^{m_2\mathsf{T}} \tag{4.2}$$

for every $m_1, m_2 \in \{0, 1, \ldots, M\}$; and

H5 $\{\lambda_n\}$ is such that for some $l_0 \in \{0, 1, 2\}$, and every $m \in \{0, 1, \ldots, M\}$,

$$\sup_{\beta \in \mathsf{B}} \|\partial_\beta^l \hat{\theta}_n^m(\beta, \lambda_n) - \partial_\beta^l \theta(\beta, 0)\| = o_p(1)$$

for all $l \in \{0, \ldots, l_0\}$.

Remark 4.2. H2 helps to ensure that the limiting variance of the GII estimator can be consistently estimated. H3 gives an asymptotically linear representation of the auxiliary model estimator, which is standard except for our requirement that this representation holds uniformly over a shrinking neighborhood of $(\beta_0, 0)$. H4 states that (standardized) auxiliary model score is asymptotically Gaussian, as will follow straightforwardly from a central limit theorem.

Remark 4.3. H5, with $l_0 = 0$, underpins our proof of the consistency of GII. The stronger forms of this condition, which require that the l_0th (and lower) derivatives of the sample binding function also converge uniformly to their population counterparts, shall be needed principally to ensure the convergence of derivative-based optimization routines to a (near) minimizer of the GII criterion. That is to say, these stronger forms of H5 are more relevant to the *computation* of the GII estimator than they are to its limiting distribution.

Of all the preceding conditions, it is H5 that will in general impose the most stringent requirements on the smoothing sequence $\{\lambda_n\}$. This is evident from the following result, which relates our high-level conditions to the low-level sufficient conditions given in Appendix B.1 (as Assumption L); its proof appears in Section E of the Supplementary Material. (That these low-level conditions are satisfied by each of Models 1–5 is then verified in Appendix B.2.) The value of $p_0 \in [2, \infty)$ below depends largely on the number of moments that the covariates $\boldsymbol{x}$ are assumed to possess: larger values of p_0 are more restrictive on $\boldsymbol{x}$, but lead to weaker conditions on $\{\lambda_n\}$.

Proposition 4.1. *Suppose Assumption L (in Appendix B.1) and R hold. Then Assumption H holds with $l_0 = 0$ in H5. Further, if $\lambda_n > 0$ for all n, with*

$$n^{1-1/p_0}\lambda_n^{2l-1}/\log(\lambda_n^{-1} \vee n) \xrightarrow{p} \infty \tag{4.3}$$

for some $l \in \{1, 2\}$, then H5 holds with $l_0 = l$.

4.2. Limiting distributions and asymptotic variance estimation

Assumptions H and R shall be maintained throughout the following, even if not explicitly referenced. However, we shall always identify the weakest form of H5 – and implicitly, of (4.3) – that is required for each of our results. The proofs of all the results that follow appear in Section D of the Supplementary Material.

Our first result concerns the limiting distributions of the minimizers of the Wald and LR criterion functions, as displayed in (3.12). For $e \in \{W, LR\}$, let $\hat{\beta}_{nk}^e$ be a near-minimizer of Q_{nk}^e, in the sense that

$$Q_{nk}^e(\hat{\beta}_{nk}^e, \lambda_n) \leq \inf_{\beta \in \mathsf{B}} Q_{nk}^e(\beta, \lambda_n) + o_p(n^{-1}). \tag{4.4}$$

The limiting variance of both estimators will have the familiar sandwich form. To allow the next result to be stated succinctly, define

$$\Omega(U, V) := (G^\mathsf{T}UG)^{-1}G^\mathsf{T}UH^{-1}VH^{-1}UG(G^\mathsf{T}UG)^{-1} \tag{4.5}$$

where $G := [\partial_\beta \theta(\beta_0, 0)]^\mathsf{T}$ denotes the Jacobian of the binding function at $(\beta_0, 0)$, $H := \mathbb{E}\ddot{\mathcal{L}}_n(\theta_0)$, and U and V are symmetric matrices.[11]

Theorem 4.1 (*Limiting Distributions*)**.** *Suppose H5 holds with $l_0 = 0$. Then*

$$n^{1/2}(\hat{\beta}_{nk}^e - \beta_0) \rightsquigarrow N[0, \Omega(U_e, V)], \tag{4.6}$$

where

$$U_e := \begin{cases} W & \text{if } e = W \\ H & \text{if } e = LR \end{cases} \qquad V := \left(1 + \frac{1}{M}\right)(\Sigma - R). \tag{4.7}$$

---

[11] A more general version of the following result, appropriate to the case when $\hat{\theta}_n^m$ maximizes a different criterion from that used to define the LR estimator, was given in Remarks 5.10 and 5.11 in an earlier version of this paper, available as arXiv:1507.06115v1.

Remark 4.4. In view of Proposition 4.1, Theorem 4.1 does *not* restrict the rate at which $\lambda_n \xrightarrow{p} 0$ from *below*; indeed, it continues to hold even if $\lambda_n = 0$ for all n, in which case the estimation problem is closely related to that considered by Pakes and Pollard (1989). Note also that the order of jackknifing does not affect the limiting distribution of the estimator: this has only a second-order effect, which vanishes as $\lambda_n \to 0$.

H5, with $l_0 = 1$, implies that the derivatives of the smoothed criterion function can be used to estimate the Jacobian matrix G that appears in the limiting variances in Theorem 4.1. The remaining components, H and V_e, can be respectively estimated using the data-based auxiliary log-likelihood Hessian, and an appropriate transformation of the joint sample variance of all the auxiliary log-likelihood scores (i.e. using both the data- and simulation-based estimates). Define

$$A^\mathsf{T} := \begin{bmatrix} I_{d_\theta} & -\frac{1}{M} I_{d_\theta} & \cdots & -\frac{1}{M} I_{d_\theta} \end{bmatrix}$$
$$s_{ni}^\mathsf{T} := \begin{bmatrix} \dot{\ell}_i^0(\hat{\theta}_n)^\mathsf{T} & \dot{\ell}_i^1(\hat{\beta}_{nk}^e, \lambda_n; \hat{\theta}_n^1)^\mathsf{T} & \cdots & \dot{\ell}_i^M(\hat{\beta}_{nk}^e, \lambda_n; \hat{\theta}_n^M)^\mathsf{T} \end{bmatrix},$$

where $\hat{\theta}_n^m := \hat{\theta}_n^m(\hat{\beta}_{nk}^e, \lambda_n)$, and $\dot{\ell}_i^0(\theta)$ denotes the gradient of $\ell(y_i, x_i; \theta)$. Then we have

Theorem 4.2 (*Variance Estimation*). *Suppose* H5 *holds with* $l_0 = 0$. *Then*

(i) $\hat{H}_n := \ddot{\mathcal{L}}_n(\hat{\theta}_n) \xrightarrow{p} H$;

(ii) $\hat{V}_n := A^\mathsf{T} \left(\frac{1}{n} \sum_{i=1}^n s_{ni} s_{ni}^\mathsf{T} \right) A \xrightarrow{p} V$; *and*

   if H5 *holds with* $l_0 = 1$, *then*

(iii) $\hat{G}_n := \partial_\beta \overline{\theta}_n(\hat{\beta}_{nk}^e, \lambda_n) \xrightarrow{p} G$, *for* $e \in \{W, LR\}$.

4.3. Convergence of derivative-based optimization procedures

Theorem 4.1 provided the limiting distribution of a near-minimizer of the GII criterion, ignoring how such a minimizer might actually be computed. Ideally, in keeping with the motivation of this paper, it should be possible to achieve this through the application of a derivative-based optimizer (DBO). This section accordingly provides (in Theorem 4.3) conditions on the smoothing $\{\lambda_n\}$ to ensure that if a DBO were applied to the GII criterion:

(i) it would converge to an approximate root of $\partial_\beta Q_{nk}^e(\beta, \lambda_n) = 0$; and

(ii) the sequence of approximate roots thus generated would have the limiting distribution given in Theorem 4.1.

(i) will follow principally from the smoothness of Q_{nk}^e, whereas (ii) will require uniform convergence of the first – and for certain procedures the second – derivatives of Q_{nk}^e. This will in turn impose a lower bound on the rate at which $\lambda_n \xrightarrow{p} 0$, something conspicuously absent from Theorem 4.1.

In large samples, Q_{nk}^e inherits the stationary points of its probability limit Q_k^e. To avoid complications that would arise due to inconsistent roots – which would otherwise interfere with (ii) – we restrict the initialization $\beta^{(0)}$ of the DBO to some $B_0 \subset B$. The precise requirements on B_0 depend on the procedure being analyzed, and are given in Assumption O. We consider two popular line-search procedures – Gauss–Newton (GN), and quasi-Newton (QN) with BFGS updating – and a trust-region (TR) algorithm (see Appendix C).[12] Let $\varrho_{\min}(D)$ and $\sigma_{\min}(D)$ respectively denote the smallest eigenvalue and smallest singular value of a matrix D and recall $G(\beta) := [\partial_\beta \theta(\beta, 0)]^\mathsf{T}$, the Jacobian of the binding function. Then

$$Q_k^W(\beta) := Q_k^W(\beta, 0) = -\|\theta(\beta, 0)\|_W^2 \qquad Q_k^{LR}(\beta) := Q_k^{LR}(\beta, 0) = \mathbb{E}\ell(y_i, x_i; \theta(\beta, 0)).$$

Assumption O (*Optimization Routines*). Let $Q \in \{Q_k^W, Q_k^{LR}\}$. Then $B_0 = B_0(Q)$ may be chosen as any compact subset of int B for which $\beta_0 \in \text{int } B_0$ and $B_0 = \{\beta \in B \mid Q(\beta) \le Q(\beta_1)\}$ for some $\beta_1 \in B$; and either

O-GN $\|G(\beta)^\mathsf{T} W \theta(\beta, 0)\| \ne 0$ for all $\beta \in B_0 \setminus \{\beta_0\}$ and $\inf_{\beta \in B_0} \sigma_{\min}[G(\beta)] > 0$;

O-QN Q is strictly convex on B_0; or

O-TR for every $\beta \in B_0 \setminus \{\beta_0\}$, $\|\partial_\beta Q(\beta)\| = 0$ implies $\varrho_{\min}[\partial_\beta^2 Q(\beta)] < -\epsilon < 0$.

Remark 4.5. Both O-GN and O-QN imply that Q has no stationary points in B_0, other than a minimum at β_0; O-TR permits such points to exist, provided that they are not local minima. In this respect, it places the weakest conditions on Q, and does so because the trust-region method utilizes second-derivative information in a manner that the other two methods do not.

To state our result on the convergence (and limiting distribution) of these optimizers, we must first specify the conditions governing their termination. Let $\{\beta^{(s)}\}$ denote the sequence of iterates generated by a given routine r, from some starting point $\beta^{(0)}$. When $r \in \{GN, QN\}$, we terminate the optimization at the first iterate, denoted s^*, for which an approximate root

---

[12] Note that Gauss–Newton can only be applied to the Wald criterion, since only this criterion has the least-squares form required by that method.

is located; if no such root is ever found, we record the initial value $\beta^{(0)}$ as the outcome of the optimization. This motivates the definition, for $r \in \{GN, QN\}$, of

$$\overline{\beta}^e_{nk}(\beta^{(0)}, r) := \begin{cases} \beta^{(s^*)} & \text{if } \|\partial_\beta Q^e_{nk}(\beta^{(s)})\| \le c_n \text{ for some } s \in \mathbb{N} \\ \beta^{(0)} & \text{otherwise,} \end{cases} \tag{4.8}$$

where $c_n = o_p(n^{-1/2})$. For $r = $ TR, we terminate only at those approximate roots at which the second-order sufficient conditions for a local minimum are also satisfied. In this way, s^* now becomes the smallest s for which $\|\partial_\beta Q^e_{nk}(\beta^{(s)})\| \le c_n$ and $\varrho_{\min}[\partial^2_\beta Q^e_{nk}(\beta, \lambda_n)] \ge 0$; $\overline{\beta}^e_{nk}(\beta^{(0)}, \text{TR})$ may then be defined analogously to (4.8).

Theorem 4.3 (*Derivative-based Optimizers*)**.** *Suppose* $r \in \{GN, QN, TR\}$ *and* $e \in \{W, LR\}$, *and that the corresponding part of* Assumption O *holds for some* B_0. *Then*

$$\sup_{\beta^{(0)} \in B_0} \|\overline{\beta}^e_{nk}(\beta^{(0)}, r) - \hat{\beta}^e_{nk}\| = o_p(n^{-1/2})$$

holds if either

  (i) $(r, e) = (GN, W)$ *and* H5 *holds with* $l_0 = 1$; *or*
  (ii) $r \in \{QN, TR\}$ *and* H5 *holds with* $l_0 = 2$.

In particular, $n^{1/2}[\overline{\beta}^e_{nk}(\beta^{(0)}, r) - \beta_0]$ *has the limiting distribution given in* (4.6).

Remark 4.6. Convergence of the Gauss–Newton procedure occurs under the weakest conditions. This is because the approximate Hessian used by that routine involves only the Jacobian $G_n(\beta) := [\partial_\beta \overline{\theta}^k_n(\beta, \lambda_n)]^\mathsf{T}$ (see (C.3)): thus the uniform convergence of $G_n(\beta)$ is sufficient in this case, whence only H5 with $l_0 = 1$ is required.

The proof of Theorem 4.3 relies on the derivatives of the sample criterion Q_{nk} converging uniformly to their population counterparts, which follows directly from H5.[13] Low-level sufficient conditions for this convergence were provided by Proposition 4.1: notably, when $l_0 \in \{1, 2\}$, (4.3) imposes exactly the sort of lower bound on λ_n that is absent from Theorem 4.1. However, the convergence of these derivatives is not alone sufficient: the smoothness of Q_{nk} is itself important. This smoothness – together with the conditions on B_0 specified in Assumption O – facilitates the application of existing results on the convergence of DBOs to approximate roots in deterministic settings (see Proposition D.6 in the Supplementary Material).

5. Monte Carlo results

In this section we conduct a set of Monte Carlo experiments to assess the performance of the GII estimator, in terms of bias, efficiency, and computation time. We divide the analysis into two parts: In Section 5.1 we explore computational aspects of GII. First, we study the performance of implementations of GII based on alternative search algorithms (quasi-Newton (QN) in conjunction with various bias reduction techniques). Second, we study the behavior of the automated bandwidth selection procedure described in Appendix A.2 (using the MSE appropriate for the level of the binding function). Third, we compare the performance of GII to the conventional II approach that seeks to optimize a non-smooth objective function using the simplex algorithm. Then, in Section 5.2, we analyze the efficiency of GII: specifically, how efficiency depends on the choice of auxiliary model, and (when possible) how it compares to maximum likelihood.

5.1. Computational aspects of GII estimators

The main motivation of our paper is the idea that GII, by smoothing the non-smooth objective functions that arise in models with discrete outcomes, leads to important computational advantages. In this section we present a Monte Carlo analysis that explores this claim.

For this analysis we focus on Models 1 and 4 described in Section 2. We compare the following alternative estimators and optimization methods: First, we report results from the conventional indirect inference approach (II) using the downhill simplex method to deal with the non-smooth objective function. Then, we report results for GII using the QN search algorithm. All optimization procedures are initialized at the true parameter values. We report results for alternative choices of the smoothing parameter λ and number of draws M, both with and without bias correction. All simulation exercises reported in this subsection use 1000 Monte Carlo replications; the reported (average) running times are for Fortran code compiled with G77, running on an Intel i7 4770k computer with Windows 10.

---

[13] Here, we are concerned exclusively with the limiting behavior of the *analytical* derivatives of Q_{nk}, ignoring any errors that might be introduced by numerical differentiation. Since Q_{nk} is smooth by construction (when $\lambda_n > 0$), it seems appropriate to assume that such numerical errors would be orders of magnitude smaller than any random variability in Q_{nk} (and its derivatives).

To implement bias correction we need a reasonable (or more formally, consistent) initial estimate of β. In practice, as discussed in Section 3.5, such initial estimates be obtained using a rather large value of λ, a small value of M, and no jackknifing ($k = 0$). These choices should generate a smooth objective that the QN algorithm can navigate quickly and easily. The initial estimator, which we denote by $\hat{\beta}_0$, is consistent, but likely to suffer from significant bias due to the large value of the smoothing parameter, as well as substantial simulation error because M is small.

We consider three forms of bias correction:

(i) *One extra NR step*: Starting from $\hat{\beta}_0$, choose a (much) smaller value of λ and larger value of M, and then take one additional NR step to obtain a new estimate of β. The reduction in λ will reduce bias, while the increase in M helps to maintain smoothness of the objective function, while also reducing simulation noise (at the cost of increased computation time). See Section 3.4.2 for more details.

(ii) *Jackknife (J/K) after QN*: Starting from that same $\hat{\beta}_0$, we take one further NR step using a (once) jackknifed criterion function (i.e. $k = 1$). This requires a choice of the extrapolation parameter $\delta \in (0, 1)$. See Section 3.4.1 for more details.

(iii) *J/K within QN*: using the same (large) λ as was used to compute $\hat{\beta}_0$, directly apply the QN algorithm to the once jackknifed criterion function ($k = 1$). Unlike the preceding two methods, this method employs a bias-reduced estimate of the binding function at every iteration, not merely at the final step.

In numerous applications (here and elsewhere) we have found that the first approach works well; the results below suggest that the other two procedures perform comparably to the first, in terms of their relative biases and efficiencies.

The QN algorithm employed here is a version of the Davidon–Fletcher–Powell algorithm (as implemented in Chapter 10 of Press et al., 1993); this is closely related to the quasi-Newton routine analyzed in Section 4.3. The initial parameter vector in the hill-climbing algorithm is always the true parameter vector. Most of the computation time in generalized indirect inference lies in computing ordinary least squares (OLS) estimates of auxiliary models. The main cost in computing OLS estimates lies, in turn, in computing the $X^\mathsf{T}X$ part of $(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$. We use blocking and loop unrolling techniques to speed up the computation of $X^\mathsf{T}X$ by a factor of 2 to 3 relative to a "naive" algorithm.[14]

In all cases, we use the LR approach to (generalized) indirect inference to construct our estimates. Unlike the Wald and LM approaches, the LR approach does not require the estimation of a weight matrix. In this respect, the LR approach is easier to implement than the other two approaches. Furthermore, because estimates of optimal weight matrices often do not perform well in finite samples (see e.g. Altonji and Segal, 1996), the LR approach is likely to perform better in small samples.

Given this background we turn to the Monte Carlo results.

5.1.1. Results for Model 1

Model 1 is a two-alternative panel probit model with serially correlated errors and one exogenous regressor. It has two unknown parameters: the regressor coefficient b, and the autoregressive coefficient r. We assume $T = 5$, set $b = 1$ and consider $r \in \{0, 0.40, 0.85\}$. We generate $n = 1000$ artificial datasets from this model. The exogenous variables (the x_{it}'s) are i.i.d. draws from a $N[0, 1]$ distribution, drawn anew for each Monte Carlo replication. We estimate the structural parameters on each dataset using each of the estimation methods discussed above.

In all cases, we use an auxiliary model consisting of T linear probability models of the form

$$y_{it} = z_{it}^\mathsf{T}\alpha_t + \xi_{it}$$

where $\xi_{it} \sim_{\text{i.i.d.}} N[0, \sigma_t^2]$, z_{it} denotes the vector of regressors for individual i in time period t, and α_t and σ_t^2 are reduced form parameters to be estimated. The natural regressors to include in z_{it} are lagged choices and polynomials in current and lagged exogenous variables. Here we simply set $z_{it} = (1, x_{it}, y_{i,t-1})$, $t = 1, \ldots, T$, where the unobserved y_{i0} is set equal to 0. (We consider using richer sets of regressors in Section 5.2, where we discuss efficiency.) Thus, the auxiliary model is characterized by the parameters $\theta = \{\alpha_t, \sigma_t^2\}_{t=1}^T$. We further impose the restrictions $\alpha_t = \alpha_q$ and $\sigma_t^2 = \sigma_q^2$ for $t = 2, \ldots, T$. This is because the time variation in the estimated coefficients of the linear probability models comes mostly from the non-stationarity of the errors in the structural model,[15] and this turns out to be negligible after the first time period. The auxiliary model parameters are estimated by maximum likelihood, which corresponds to OLS given the distributional assumptions on ξ_{it}.[16]

Table 1 reports, for each method, the means and empirical standard deviations of the estimates, and the computation time. The first panel of the table reports results for the case of $r = 0$ (i.e., No serial correlation), while the bottom two panels correspond to $r = 0.4$ and $r = 0.85$, respectively. The first two rows (of each panel) report results for conventional (non-smooth) II estimators, obtained using the downhill simplex method. In our notation these methods correspond to setting $\lambda = 0$. We consider both $M = 10$ and $M = 50$.

---

[14] To avoid redundant calculations, we also precompute and store for later use those elements of $X^\mathsf{T}X$ that depend only on the exogenous variables. We are grateful to James MacKinnon for providing code that implements the blocking and loop unrolling techniques.

[15] Note that we do not draw the initial error from the stationary distribution implied by the law of motion for the errors.

[16] It is worth emphasizing that we include lagged choices (and lagged x's) in the auxiliary model despite the fact that the structural model does not exhibit true state dependence. But in Model 1 it is well-know that lagged choices are predictive of current choices (termed "spurious state dependence" by Heckman). This is a good illustration of how a good auxiliary model should be designed to capture the correlation patterns in the data, as opposed to the true structure.

Table 1
Monte Carlo results for Model 1.

	M/λ	δ	Mean		Std. Dev.		avg. time (s)
			b	r	b	r	
$b = 1, r = 0$							
Downhill simplex	10/0		1.002	0.000	0.041	0.069	0.209
	50/0		1.003	−0.001	0.039	0.066	0.996
Initial QN estimates	10/0.03		0.955	−0.002	0.039	0.062	0.229
One extra NR step	50/0.003		0.998	−0.001	0.039	0.065	0.316
	150/0.003		0.998	−0.001	0.039	0.065	0.499
	300/0.003		0.998	−0.001	0.039	0.064	0.899
J/K after QN	10/0.03	0.66	1.001	−0.002	0.041	0.069	0.250
	10/0.03	0.33	1.002	−0.002	0.041	0.070	0.294
J/K within QN	10/0.03	0.66	1.002	−0.000	0.041	0.068	0.415
	10/0.03	0.33	1.003	−0.000	0.041	0.068	0.513
	10/0.005[a]	0.66	1.002	−0.002	0.041	0.063	
$b = 1, r = 0.4$							
Downhill simplex	10/0		1.006	0.403	0.050	0.067	0.215
	50/0		1.004	0.401	0.047	0.062	1.011
Initial QN estimates	10/0.03		0.948	0.365	0.045	0.059	0.225
One extra NR step	50/0.003		0.998	0.398	0.046	0.062	0.312
	150/0.003		0.998	0.397	0.046	0.062	0.500
	300/0.003		0.998	0.397	0.046	0.061	0.888
J/K after QN	10/0.03	0.66	1.002	0.400	0.048	0.065	0.244
	10/0.03	0.33	1.003	0.400	0.049	0.065	0.286
J/K within QN	10/0.03	0.66	1.003	0.401	0.049	0.066	0.422
	10/0.03	0.33	1.004	0.401	0.049	0.065	0.532
	10/0.005[a]	0.66	1.004	0.401	0.047	0.059	
$b = 1, r = 0.85$							
Downhill simplex	10/0		1.020	0.863	0.089	0.073	0.207
	50/0		1.009	0.856	0.081	0.069	1.016
Initial QN estimates	10/0.03		0.922	0.786	0.068	0.063	0.268
One extra NR step	50/0.003		0.993	0.845	0.077	0.066	0.348
	150/0.003		0.993	0.846	0.077	0.066	0.537
	300/0.003		0.992	0.845	0.075	0.065	0.915
J/K after QN	10/0.03	0.66	0.999	0.851	0.081	0.071	0.287
	10/0.03	0.33	1.000	0.851	0.083	0.072	0.333
J/K within QN	10/0.03	0.66	1.006	0.852	0.082	0.069	0.510
	10/0.03	0.33	1.006	0.853	0.081	0.068	0.650
	10/0.014[a]	0.66	1.004	0.852	0.069	0.060	

[a] Reports the mean λ selected using minimum-MSE procedure described in Model 5, using the MSE for the level of the binding function.

The next nine rows of each panel correspond to GII estimators, implemented using different optimization methods. The first row, labeled "Initial QN estimates", reports results for $(\lambda, M) = (0.03, 10)$. In this model 0.03 is a large value for λ. This is clear from intuition, as this is a probit model where the errors have a standard deviation of 1.0,[17] and it is confirmed by the optimal bandwidth selection procedure of Section 3.5 that we implement below. Obviously, $M = 10$ is also a small number of draws. Thus, these estimates correspond to the initial 'consistent' estimator $\hat{\beta}_0$ discussed above.

The subsequent rows correspond to the three bias correction methods. The first, "One extra NR step", starts from $\hat{\beta}_0$ and takes one additional NR step using after reducing the smoothing parameter by an order of magnitude, to $\lambda = 0.003$, and employs a larger simulation size of $M = 50$, 150 or 300. The second method, "J/K after QN", is the jackknife (with $k = 1$) starting from $\hat{\beta}_0$ and using an extrapolation parameter of $\delta = 0.66$ or 0.33. The third method, "J/K within QN", uses the same $(\lambda, M) = (0.03, 10)$ as we used in the "Initial QN estimates", but implements the jackknife bias correction ($k = 1$)

---

[17] In order to gain intuition for the magnitude of smoothing parameters, it is very useful to imagine that smoothing is induced by optimization error, and to consider what this implies about the probability an agent makes choices that depart from utility maximization. For example, imagine that for agent i at time t, option 1 has a utility of 0.10 and option 2 has a utility of 0. Of course, in the absence of smoothing, option 1 is chosen. But, with $\lambda = 0.03$, option 1 is instead assigned a probability of $\exp(0.10/0.03)/[1 + \exp(0.10/0.03)] = 0.9655$. Thus, the agent has a 3.45% chance of making the "wrong" choice when inferior option is within an 0.10 standard deviation taste shock of the preferred option. This is relatively large degree of smoothing.

If instead we were to set $\lambda = 0.003$, then this same agent would be assigned a probability of 1 for option 1 (up to machine accuracy). The inferior option would have to be much closer to the preferred option (in terms of standard deviations of the taste shock) to be assigned a non-negligible probability. For example, an option that was inferior by 0.01 standard deviations would be assigned a 3.45% probability, and any options that are inferior by more than 0.02 standard deviations would be assigned negligible probability.

Table 2
Distribution of automatically selected λ's (Model 1).

	Mean	Median	Std. Dev.	Max.	Min.
$b = 1, r = 0$	0.005	0.003	0.005	0.029	0.001
$b = 1, r = 0.4$	0.005	0.003	0.006	0.033	0.001
$b = 1, r = 0.85$	0.014	0.007	0.017	0.101	0.001

$\diamond$ For each Monte Carlo replication, λ is chosen so as to minimize the estimated MSE of the (level of the) binding function, evaluated at the values of the structural parameters from which the optimization commences (see Appendix A.2). λ is computed via a grid search, using a grid that spans from 0.0005 to 0.635, whose intermediate points are given by $0.0005 \times (1.1)^i$, for $i \in \{0, 1, \ldots, 75\}$.

throughout every iteration of the QN algorithm (again, using $\delta = 0.66$ or 0.33). Finally, the last row of each panel reports "J/K within QN" estimates obtained when λ is chosen by minimizing the estimated MSE of the sample binding function (see Appendix A.2).

It is evident from Table 1 that the "Initial QN" estimates exhibit significant biases. While the true value of the regressor coefficient b is 1.0, the means of the "Initial QN" estimates are 0.955, 0.948 and 0.922 when $r = 0$, 0.4 or 0.85, respectively; mean estimates of r in these three cases are $-.002$, 0.365 and 0.786, respectively. Notice that biases for both parameters grow larger as serial correlation (r) increases. However, all three proposed bias correction methods do an excellent job of removing this bias, at all levels of serial correlation. The "One extra NR step" method exhibits (very) slightly smaller empirical standard errors, but the difference is not great enough to make any one method clearly preferable.

While we implement the "One extra NR step" method using three different values of M (50, 150 and 300), both the bias and efficiency (i.e., the empirical standard deviation of the estimates) are little affected by the choice of M. This is perhaps not so surprising, since the standard deviation of the GII estimator is proportional to $(1 + M^{-1})^{1/2}$, which evaluates to 1.01 when $M = 50$ (see Theorem 4.1). While computation time for a single iteration is proportional to M, overall computation time increases much less than proportionately with M, because the larger value of M is only used in the one final step.

Both variants of the jackknifed estimators were implemented for two values of the extrapolation parameter, $\delta = 0.66$ and 0.33. It is reassuring that the choice of δ has essentially no bearing on the results. Not surprisingly, the "J/K within QN" procedure is almost twice as slow as the "J/K after QN" procedure. This is because the former requires the calculation of a jackknifed binding function on each iteration (of the search algorithm), while the latter only requires that this be done once, starting from the "Initial QN" estimates.

Overall, computation times for the "One extra NR step" method with $(\lambda, M) = (0.003, 50)$ and the "J/K after QN" method are quite comparable. Thus, there is little to chose between these methods in terms of either bias, efficiency or computation time. But they may both be preferred over "J/K within QN", which is somewhat slower.

Automated smoothing parameter selection. The final row of each panel of Table 1 reports the estimates obtained when λ is chosen so as to minimize the estimated mean-squared error of the sample binding function (evaluated at the starting point of the optimization), as per the procedure described in Appendix A.2. Though the mean value of λ selected by the procedure varies with r, in all cases the procedure delivers values of λ that are generally consistent with the jackknifed estimator producing estimates with little bias, and enjoying an efficiency comparable to (indeed slightly better than) those corresponding to a fixed value of λ (i.e. 0.03).

Table 2 gives further details on the distribution of the λ selected by the procedure. The median selected λ is 0.003 in Model 1, both when $r = 0$ and $r = 0.40$; the standard deviation is a substantial 0.005. When $r = 0.85$, the median λ increases to 0.007, and the standard deviation also increases to 0.017. It is interesting that the value of $\lambda = 0.003$ that we chose for the "One extra QN step" procedure, based on both the intuition discussed in footnote 17 and experimentation with what worked well in practice, is well within the ballpark of the λ selected by this procedure (which were computed *ex post*). Our view is that, as a practical matter, an applied researcher using GII would typically start with a rather large λ to initiate iterations, and then reduce λ until results stabilize, along the lines suggested in Section 3.5. The automated selection procedure may therefore be more useful *ex post* as means of checking that one has not settled on a λ that is too large.

5.1.2. Comparison of GII with the simplex method: Model 1

Another notable result in Table 1 is that the conventional II estimator, implemented using a simplex algorithm designed to optimize non-smooth objective functions, performs as well, in terms of bias, efficiency and computation time, as the bias-adjusted GII methods. This may seem to contradict the main premise of this paper. However, it should be unsurprising that the simplex has no difficulty maximizing a non-smooth objective functions with respect to only two parameters. In more complex models, as we will now show, conventional II is liable to perform poorly.

To examine how II performs in a model with more parameters, in Table 3 we report results for an expanded version of Model 1 in which the regressor x_{it} is now a vector of *fourteen* i.i.d. $N[0, 1]$ random variables. Thus, the model has fifteen total parameters. We consider only the case where the autoregressive parameter $r = 0.85$. We set $b_1 = b_2 = b_3 = b_4 = 0.5$ and $b_k = 0$ for $k = 5, \ldots, 14$; these parameter values were chosen so that the variance of the deterministic part of utility

Table 3
Monte Carlo results for Model 1 with additional variables ($b_1 = b_2 = b_3 = b_4 = 0.5$, $b_5 = b_6 = \cdots = b_{14} = 0$, $r = 0.85$).

		Downhill simplex		One extra NR step		J/K + QN	
						After	Within
M		10	50	50	300	10	10
λ		0	0	0.003	0.003	0.03	0.03
Mean							
b_1		0.504	0.510	0.501	0.499	0.505	0.503
b_2		0.519	0.514	0.503	0.501	0.507	0.506
b_3		0.545	0.528	0.504	0.502	0.508	0.507
b_4		0.540	0.521	0.501	0.499	0.505	0.505
$b_5 - b_{14}$	min	−0.014	−0.015	−0.001	−0.001	−0.002	−0.002
	max	−0.040	0.009	0.001	0.001	0.002	0.002
r		0.887	0.874	0.855	0.852	0.861	0.857
Std. deviation							
b_1		0.061	0.061	0.054	0.051	0.056	0.052
b_2		0.065	0.064	0.055	0.053	0.058	0.054
b_3		0.089	0.075	0.056	0.053	0.058	0.053
b_4		0.090	0.072	0.053	0.052	0.058	0.052
$b_5 - b_{14}$	min	0.047	0.041	0.035	0.035	0.037	0.036
	max	0.090	0.065	0.037	0.037	0.040	0.039
r		0.084	0.082	0.077	0.073	0.081	0.070
Computation							
Time		14.9	92.3	10.40	24.47	10.27	25.84
Avg. iterations		586.1	754.1	13.89	13.89	13.88	21.51
Avg. iter stuck		8.0	7.9				
% finish stuck		10%	17%				

◇ All results using the jackknifing (J/K) use $\delta = 0.66$. 'Avg. iterations' refers to the number of iterations used in either the downhill simplex algorithm or the initial quasi-Newton algorithm, averaged over all Monte Carlo replications. 'Avg. iter stuck' refers to the average number of times that the downhill simplex algorithm fails to find an improvement from reflecting or contracting the worst point in the simplex. '% finish stuck' refers to the percentage of Monte Carlo replications for which the downhill simplex algorithm finished on an iteration where reflecting or contracting failed to produce an improvement in the worst point in the simplex.

($b^{\mathsf{T}}x_{it}$) remains equal to 1.0, exactly as in the original, smaller formulation of Model 1. Thus, the explanatory power of the regressors is the same as in the original model, and the magnitude of the smoothing parameter (λ) is unchanged relative to the variance of the deterministic and stochastic parts of utility (see footnote 17).

The results in Table 3 show that the performance of the conventional II estimator deteriorates as expected in this larger model, while all three bias-adjusted GII approaches work just as well as they did for the smaller model. For example consider the mean estimates of b_1 through b_4. In the case of the II-simplex method with $M = 50$, these are all biased upward by an average of 3.7%. In contrast, the GII approach with "One extra NR step" pins down all four parameters very precisely. Similarly, the mean II estimate of r is biased upward by 2.8%, whereas those produced by the "One extra QN step" are not.

Admittedly, these II biases are modest, but more problematic for II is the issue of efficiency. For instance, the empirical standard deviations of the II estimates (based on $M = 50$) of b_1 through b_4 average 0.068, while those of GII with "One extra NR step" average 0.052. Thus, II suffers from a 30% efficiency loss for these parameters. Efficiency losses for b_5 through b_{14} are comparable.

Given the symmetry of the problem, the empirical standard errors of b_1 through b_4 should be nearly equal to each other. This is roughly true for GII with "One extra NR step", where they are tightly bunched around 0.052. But for II these values range from 0.061 to 0.075. This signals to us that the simplex algorithm is encountering difficulties. The same problem of uneven standard errors is apparent for the II estimates of b_5 through b_{14} (not reported).

Finally, when we compare computation times, we see that the GII method with "One extra NR step", is roughly four times faster than the II-simplex method with $M = 50$. If we reduce M to speed up the II estimator, its performance deteriorates further, as can be seen by comparing columns 1 and 2 of Table 3. Furthermore, the "One extra NR step" is implemented here using $M = 300$, and, as we saw in Table 1, this can be reduced considerably without adverse consequences. Finally, note that the "J/K after QN" method (Table 3, col. 4) is 50% faster than the II-simplex method using only $M = 10$ draws, yet it produces results almost as good (in terms of bias and efficiency) as GII implemented via the "One extra NR step" approach.

5.1.3. Results for Model 4

Model 4 is a (static) three-alternative probit model with eight unknown parameters: three coefficients in each of the two latent utility equations ($\{b_{1i}\}_{i=0}^{2}$ and $\{b_{2i}\}_{i=0}^{2}$) and two parameters governing the covariance matrix of the stochastic terms

Table 4
Monte Carlo results for Model 4.

	M/λ	b_{10}	b_{11}	b_{12}	b_{20}	b_{21}	b_{22}	c_1	c_2	Time[a]
Mean										
True values		0	1	1	0	1	1	1.33	1	
Downhill simplex	10/0	0.020	0.956	0.956	−0.038	0.987	1.006	1.365	1.053	1.788
	50/0	0.014	0.950	0.952	−0.029	0.973	0.989	1.331	1.048	7.820
Initial QN estimates	10/0.03	0.004	0.944	0.953	−0.024	0.957	0.973	1.292	1.041	1.562
One extra NR step	50/0.003	0.000	0.993	0.998	−0.010	1.000	1.005	1.346	0.990	2.229
	150/0.003	0.000	0.993	0.998	−0.010	1.000	1.005	1.346	0.993	2.560
	300/0.003	0.000	0.993	0.998	−0.011	1.000	1.006	1.347	0.993	3.149
J/K after QN										
$\delta = 0.66$	10/0.03	−0.001	1.000	1.005	−0.009	1.005	1.010	1.350	0.990	2.365
$\delta = 0.33$	10/0.03	−0.002	1.001	1.006	−0.008	1.005	1.010	1.349	0.989	2.406
J/K within QN										
$\delta = 0.66$	10/0.03	−0.001	0.998	1.003	−0.012	1.005	1.011	1.352	0.987	3.726
$\delta = 0.33$	10/0.03	−0.001	0.998	1.003	−0.012	1.005	1.011	1.352	0.987	4.678
Std. deviation										
Downhill simplex	10/0	0.085	0.076	0.093	0.104	0.138	0.175	0.287	0.176	
	50/0	0.079	0.070	0.087	0.093	0.123	0.157	0.252	0.159	
Initial QN estimates	10/0.03	0.078	0.067	0.085	0.092	0.110	0.143	0.233	0.154	
One extra NR step	50/0.003	0.081	0.071	0.090	0.091	0.117	0.151	0.253	0.162	
	150/0.003	0.079	0.070	0.088	0.090	0.114	0.147	0.246	0.159	
	300/0.003	0.079	0.070	0.088	0.090	0.114	0.147	0.245	0.160	
J/K after QN										
$\delta = 0.66$	10/0.03	0.084	0.076	0.095	0.095	0.123	0.158	0.260	0.175	
$\delta = 0.33$	10/0.03	0.084	0.076	0.096	0.095	0.124	0.160	0.262	0.176	
J/K within QN										
$\delta = 0.66$	10/0.03	0.082	0.073	0.094	0.096	0.119	0.158	0.256	0.172	
$\delta = 0.33$	10/0.03	0.082	0.073	0.094	0.095	0.118	0.157	0.253	0.171	

[a] Average running time in seconds.

in these equations (c_1 and c_2). We set $b_{10} = b_{20} = 0$, $b_{11} = b_{12} = b_{21} = b_{22} = 1$, $c_2 = 1$, and $c_1 = 1.33$ (implying that the stochastic terms in the latent utilities have a correlation of 0.8). We set $n = 2000$.

The auxiliary model is a pair of linear probability models, one for each of the first two alternatives:

$$y_{i1} = z_i^\mathsf{T} \alpha_1 + \xi_{i1}$$
$$y_{i2} = z_i^\mathsf{T} \alpha_2 + \xi_{i2},$$

where $\xi_i \sim_{\text{i.i.d.}} N[0, \Sigma_\xi]$. The natural regressors to include in z_i are polynomial functions of the exogenous variables $\{x_{ij}\}_{j=1}^3$. Here, we set $z_i = (1, x_{i1}, x_{i2}, x_{i3})$, giving a total of 11 auxiliary model parameters $\theta = (\alpha_1, \alpha_2, \Sigma_\xi)$. These parameters are estimated by OLS. This corresponds to maximum likelihood here, even though Σ_ξ is not diagonal, because the same regressors appear in both equations.

Table 4 presents the results. Again, the GII "Initial QN" estimates exhibit significant biases. This is particularly true for the parameters b_{11}, b_{12} and b_{21}, which are all biased downward by roughly 5%. The error correlation is also biased downward. However, all three proposed bias correction methods do an excellent job of removing these biases. As in Model 1, the "One extra NR step" method exhibits slightly smaller empirical standard errors than "J/K after QN" or "Jackknife within QN", but the difference is not great enough to make any one method clearly preferable.

As in Table 1, we again find that the "One extra NR step" results are not sensitive to whether we choose $M = 50$, 150 or 300. Similarly, the jackknife results are not sensitive to whether we use an extrapolation parameter of $\delta = 0.66$ or 0.33. And again, both the "One extra NR step" and the "J/K after QN" methods are faster to compute than the "J/K within QN" method.

A notable aspect of Table 4 results is that the conventional II-simplex estimator exhibits substantial biases, regardless of whether $M = 10$ or 50. For example, when $M = 50$, the II estimates of b_{11} and b_{12} are biased downward by 5%, and c_1 is biased upward by 11%. This is in sharp contrast to Table 1 results, but it is consistent with the results in Table 3. This suggests that even though the multinomial probit model contains only 8 parameters, this is enough to create some problems for the simplex algorithm.

Finally, it is worth stressing that in our work we have used relatively small models that are amenable to Monte Carlo analysis. Yet we have still found that GII out-performs II in models with 8 (Table 4) or 15 (Table 3) parameters. In the larger (more highly parametrized) models that typically arise in empirical practice, the limitations of the II-simplex approach, and the advantages of GII, will be even more evident.

5.2. Relative efficiency of GII estimators

5.2.1. Choosing the auxiliary model

In this sub-section we investigate the efficiency of GII estimators. The issue of efficiency is closely related to the issue of how to choose an auxiliary model. As discussed in Section 3.1, indirect inference (generalized or not) has the same asymptotic efficiency as maximum likelihood when the auxiliary model is correctly specified − in the sense that the auxiliary model provides a correct statistical description of the observed data (Gallant and Tauchen, 1996). Thus, from the perspective of efficiency, it is important to choose an auxiliary model (or a class of auxiliary models) that is flexible enough to provide a good description of the data.

Another important consideration is computation time. For the Wald and LR approaches to indirect inference, the auxiliary parameters must be estimated repeatedly using different simulated datasets. For this reason, it is critical to use an auxiliary model that can be estimated quickly and efficiently.[18]

To meet the twin criteria of statistical efficiency and computational speed, we recommend using linear probability models (or sets of linear probability models) as the auxiliary model in all of the pure discrete choice models that we use as test cases for GII (i.e., Models 1–4 of Section 2) The class of linear probability models is flexible in the sense that an individual's current choice can be allowed to depend on polynomial functions of lagged choices and of current and lagged exogenous variables. Linear probability models can also be estimated very quickly and easily using ordinary least squares. For Model 5, the Heckman selection model, an appropriate auxiliary model would be a set of OLS regressions with mixed discrete/continuous dependent variables.

The subsequent sub-sections report Monte Carlo experiments on Models 1–4 of Section 2. We focus on how the efficiency of the GII estimators varies with the choice of auxiliary model. For Models 1, 2, and 4, we consider both GII estimators and the simulated maximum likelihood (SML) estimator in conjunction with the GHK smooth probability simulator (cf. Keane, 1994; Lee, 1997). Model 3, which cannot easily be estimated via SML, is estimated using only GII. We omit Model 5 from our analysis, as Altonji et al. (2013) already present results showing that GII performs well for Heckman selection-type models.

5.2.2. Results for Model 1

Model 1 is a two-alternative panel probit model with serially correlated errors and one exogenous regressor. We described it in detail in Section 5.1.1. Recall that the auxiliary model consists of T linear probability models of the form

$$y_{it} = z_{it}^{\mathsf{T}}\alpha_t + \xi_{it}$$

where $\xi_{it} \sim_{\text{i.i.d.}} N[0, \sigma_t^2]$ and where $y_{i0} = 0$. The vector of regressors z_{it} includes both lagged choices and polynomial functions of current and lagged exogenous variables. The set of variables included in z_{it} may grow over time so as to incorporate the additional lagged information that is available in later periods. The auxiliary model is thus characterized by the parameters $\theta = \{\alpha_t, \sigma_t^2\}_{t=1}^T$, which are estimated by OLS.

To examine how increasing the "richness" of the auxiliary model affects efficiency of the structural parameter estimates, we conduct Monte Carlo experiments using four nested auxiliary models. In all four, we impose the restrictions $\alpha_t = \alpha_q$ and $\sigma_t^2 = \sigma_q^2, t = q + 1, \ldots, T$, for some $q < T$. That is, we assume the process is approximately stationary from *period* $q + 1$ onward.

In auxiliary model #1, $q = 1$ and the regressors in the linear probability model are given by: $z_{it} = (1, x_{it}, y_{i,t-1})$, $t = 1, \ldots, T$. This is the same auxiliary model that we used in Section 5.1.1.

In auxiliary model #2, $q = 2$ and the regressors are

$$z_{i1} = (1, x_{i1}) \qquad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}), \ t \in \{2, \ldots, T\},$$

giving a total of 18 parameters. Auxiliary model #3 has $q = 4$, and regressors

$$z_{i1} = (1, x_{i1}, x_{i1}^3) \qquad\qquad z_{i3} = (1, x_{i3}, y_{i2}, x_{i2}, y_{i1}, x_{i1})$$
$$z_{i2} = (1, x_{i2}, y_{i1}, x_{i1}) \qquad\qquad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}), \ t \in \{4, \ldots, T\},$$

and 24 parameters. Finally, auxiliary model #4 has the same regressors as #3, except that

$$z_{i4} = (1, x_{i4}, y_{i3}, x_{i3}, y_{i2}, x_{i2}, y_{i1}, x_{i1})$$
$$z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}, x_{i,t-3}, y_{i,t-4}), \ t \in \{5, \ldots, T\}$$

so $q = 5$ and there are 35 parameters.

Table 5 presents the results of six sets of Monte Carlo experiments, each with 2000 replications. The first two sets of experiments report the results for simulated maximum likelihood, based on GHK, using 25 draws (SML #1) and 50 draws (SML #2). The remaining four sets of experiments report the results for generalized indirect inference, where GII #i refers to generalized indirect inference using auxiliary model #i. In each case, we report the average and the standard deviation of the parameter estimates. We also report the efficiency loss of GII #i relative to SML #2 in the columns labeled $\sigma_{\text{GII}}/\sigma_{\text{SML}}$,

---

[18] This consideration is less important for the LM approach, as it does not work directly with the estimated auxiliary parameters, but instead uses the first-order conditions (the score vector) that define these estimates.

Table 5
Monte Carlo results for Model 1.

	Mean		Std. dev.		$\sigma_{GII}/\sigma_{SML}$		Time (s)
	b	r	b	r	b	r	
$b = 1, r = 0$							
SML #1	1.000	−0.002	0.0387	0.0454	–	–	0.76
SML #2	1.001	−0.000	0.0373	0.0468	–	–	1.53
GII #1	0.998	0.002	0.0390	0.0645	1.05	1.37	0.67
GII #2	0.993	0.001	0.0386	0.0490	1.03	1.05	0.72
GII #3	0.992	0.001	0.0393	0.0490	1.05	1.05	0.91
GII #4	0.988	0.001	0.0390	0.0485	1.05	1.04	0.99
$b = 1, r = 0.4$							
SML #1	0.995	0.385	0.0400	0.0413	–	–	0.78
SML #2	0.999	0.392	0.0390	0.0410	–	–	1.54
GII #1	0.998	0.399	0.0454	0.0616	1.16	1.50	0.70
GII #2	0.993	0.396	0.0410	0.0456	1.05	1.11	0.72
GII #3	0.991	0.395	0.0417	0.0432	1.07	1.05	0.91
GII #4	0.987	0.392	0.0416	0.0432	1.07	1.05	0.97
$b = 1, r = 0.85$							
SML #1	0.984	0.833	0.0452	0.0333	–	–	0.74
SML #2	0.993	0.842	0.0432	0.0316	–	–	1.47
GII #1	0.994	0.846	0.0791	0.0672	1.83	2.13	0.71
GII #2	0.991	0.845	0.0511	0.0412	1.18	1.30	0.74
GII #3	0.992	0.846	0.0492	0.0357	1.14	1.13	0.93
GII #4	0.988	0.841	0.0490	0.0357	1.13	1.13	1.00

where we divide the standard deviations of the GII estimates by the standard deviations of the estimates for SML #2. Finally, we report the average time (in seconds) required to compute estimates (we use the Intel Fortran Compiler Version 7.1 on a 2.2 GHz Intel Xeon processor running Red Hat Linux).[19]

Table 5 contains several key findings.

First, both SML and GII generate estimates with very little bias.

Second, GII is less efficient than SML, but the efficiency losses are small provided that the auxiliary model is sufficiently rich. For example, auxiliary model #1 leads to large efficiency losses, particularly for the case of high serial correlation in the errors ($r = 0.85$). For models with little serial correlation ($r = 0$), however, auxiliary model #2 is sufficiently rich to make GII almost as efficient as SML. When there is more serial correlation in the errors, auxiliary model #2 leads to reasonably large efficiency losses (as high as 30% when $r = 0.85$), but auxiliary model #3, which contains more lagged information in the linear probability models than does auxiliary model #2, reduces the worst efficiency loss to 13%. Auxiliary model #4 provides almost no efficiency gains relative to auxiliary model #3.

Third, GII is faster than SML: computing a set of estimates using GII with auxiliary model #3 takes about 30% less time than computing a set of estimates using SML with 50 draws.

For generalized indirect inference, we also compute (but do not report in Table 5) estimated asymptotic standard errors, using the estimators described in Theorem 4.2. In all cases, the averages of the estimated standard errors across the Monte Carlo replications are very close to (within a few percent of) the actual standard deviations of the estimates, suggesting that the asymptotic results provide a good approximation to the behavior of the estimates in samples of the size that we use.

5.2.3. Results for Model 2

Model 2 is a panel probit model with serially correlated errors, a single exogenous regressor, and a lagged dependent variable. It has three unknown parameters: b_1, the coefficient on the exogenous regressor, b_2, the coefficient on the lagged dependent variable, and r, the serial correlation parameter. We set $b_1 = 1$, $b_2 = 0.2$, and consider $r \in \{0, 0.4, 0.85\}$; $n = 1000$ and $T = 10$.

Table 6 presents the results of six sets of Monte Carlo experiments, each with 1000 replications; the labels SML #i and GII #i are to be interpreted exactly as for Table 1. The results are similar to those for Model 1. Both SML and GII generate estimates with very little bias. SML is more efficient than GII, but the efficiency loss is small when the auxiliary model is sufficiently rich (i.e., 17% at most for model #3, 15% at most for model #4). However, auxiliary model #1 can lead to very large efficiency losses, as can auxiliary model #2 if there is strong serial correlation.

Again, average asymptotic standard errors are close to the standard deviations obtained across the simulations (not reported). Finally, GII using auxiliary model #3 is about 25% faster than SML using 50 draws.

---

[19] Note that the Monte Carlo analysis reported in Section 5.1 was done in 2016, while that reported in this section (Section 5.2) was done several years earlier. Thus, the absolute times are not comparable across sections, nor are timings reported in Section 5.2 indicative of the computation times that would be required for these methods today. Only the *relative* timing across estimation methods reported in Section 5.2 are of interest.

Table 6
Monte Carlo results for Model 2.

	Mean			Std. dev.			$\sigma_{GII}/\sigma_{SML}$			Time (s)
	b_1	r	b_2	b_1	r	b_2	b_1	r	b_2	
$b_1 = 1, r = 0, b_2 = 0.2$										
SML #1	1.000	0.001	0.200	0.0274	0.0357	0.0355	–	–	–	2.47
SML #2	1.002	0.002	0.199	0.0273	0.0362	0.0365	–	–	–	4.89
GII #1	0.999	0.001	0.199	0.0267	0.0571	0.0437	0.98	1.58	1.20	2.72
GII #2	0.996	0.000	0.199	0.0267	0.0379	0.0379	0.98	1.05	1.04	2.80
GII #3	0.995	0.001	0.199	0.0269	0.0377	0.0376	0.99	1.04	1.03	3.66
GII #4	0.993	0.000	0.198	0.0270	0.0377	0.0375	0.99	1.04	1.03	4.06
$b_1 = 1, r = 0.4, b_2 = 0.2$										
SML #1	0.994	0.379	0.214	0.0278	0.0314	0.0397	–	–	–	2.42
SML #2	0.999	0.389	0.206	0.0287	0.0316	0.0397	–	–	–	4.82
GII #1	0.997	0.397	0.198	0.0339	0.0587	0.0544	1.18	1.86	1.37	2.73
GII #2	0.994	0.396	0.198	0.0293	0.0386	0.0462	1.02	1.22	1.16	2.82
GII #3	0.993	0.396	0.197	0.0289	0.0343	0.0431	1.01	1.09	1.09	3.64
GII #4	0.991	0.395	0.196	0.0289	0.0348	0.0434	1.01	1.10	1.09	4.02
$b_1 = 1, r = 0.85, b_2 = 0.2$										
SML #1	0.974	0.831	0.220	0.0321	0.0174	0.0505	–	–	–	2.78
SML #2	0.987	0.840	0.208	0.0327	0.0159	0.0507	–	–	–	5.47
GII #1	1.000	0.854	0.183	0.0952	0.0633	0.1185	2.91	3.98	2.34	3.01
GII #2	0.992	0.852	0.190	0.0417	0.0266	0.0721	1.28	1.67	1.42	2.92
GII #3	0.992	0.851	0.191	0.0383	0.0179	0.0547	1.17	1.13	1.08	3.68
GII #4	0.990	0.850	0.188	0.0379	0.0175	0.0548	1.15	1.10	1.09	4.06

5.2.4. Results for Model 3

Model 3 is identical to Model 2, except there is an "initial conditions" problem: the econometrician does not observe individuals' choices in the first s periods. This is an excellent example of the type of problem that motivates this paper: SML is extremely difficult to implement, due to the problem of integrating over the initial conditions. But II is appealing, as it is still trivial to simulate data from the model. However, we need GII to deal with the discrete outcomes.

To proceed, our Monte Carlo experiments are parametrized exactly as for Model 2, except that we set $T = 15$, with choices in the first $s = 5$ time periods being unobserved (but note that exogenous variables *are* observed in these time periods).

Auxiliary model #1 is as for Models 1 and 2: $q = 1$ and the regressors are $z_{it} = (1, x_{it}, y_{i,t-1})$, $t = s+1, \ldots, T$, where the unobserved y_{is} is set equal to 0. In auxiliary model #2, $q = 2$ and the regressors are:

$$z_{i,s+1} = (1, x_{i,s+1}, x_{is}) \qquad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}), \ t \in \{s+2, \ldots, T\},$$

for a total of 19 parameters. In auxiliary model #3, $q = 4$ and there are 27 parameters:

$$z_{i,s+1} = (1, x_{i,s+1}, x_{i,s+1}^3, x_{is}, x_{i,s-1})$$
$$z_{i,s+2} = (1, x_{i,s+2}, y_{i,s+1}, x_{i,s+1}, x_{is})$$
$$z_{i,s+3} = (1, x_{i,s+3}, y_{i,s+2}, x_{i,s+2}, y_{i,s+1}, x_{i,s+1})$$
$$\quad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}), \ t \in \{s+4, \ldots, T\}.$$

Finally, in auxiliary model #4, $q = 5$ and there are 41 parameters: relative to #3, $z_{i,s+1}, z_{i,s+2}$ and $z_{i,s+3}$ are augmented by an additional lag of x_{is}, and

$$z_{i,s+4} = (1, x_{i,s+4}, y_{i,s+3}, x_{i,s+3}, y_{i,s+2}, x_{i,s+2}, y_{i,s+1}, x_{i,s+1})$$
$$\quad z_{it} = (1, x_{it}, y_{i,t-1}, x_{i,t-1}, y_{i,t-2}, x_{i,t-2}, y_{i,t-3}, x_{i,t-3}, y_{i,t-4}), \ t \in \{s+5, \ldots, T\}.$$

Table 7 presents the results of four sets of Monte Carlo experiments, each with 1000 replications. There are two key findings: First, as with Models 1 and 2, GII generates estimates with very little bias. Second, increasing the "richness" of the auxiliary model leads to large efficiency gains relative to auxiliary model #1, particularly when the errors are persistent. However, auxiliary model #4 provides few efficiency gains relative to auxiliary model #3.

5.2.5. Results for Model 4

Model 4 is a (static) three-alternative probit model with eight unknown parameters. We described it in detail in Section 5.1.3, so we will not repeat that description here. Recall that the auxiliary model is a pair of linear probability models, one for each of the first two alternatives:

$$y_{i1} = z_i^T \alpha_1 + \xi_{i1}$$
$$y_{i2} = z_i^T \alpha_2 + \xi_{i2},$$

Table 7
Monte Carlo results for Model 3.

	Mean			Std. dev.			Time (s)
	b_1	r	b_2	b_1	r	b_2	
$b_1 = 1, r = 0, b_2 = 0.2$							
GII #1	0.997	−0.000	0.200	0.0272	0.0532	0.0387	3.91
GII #2	0.994	−0.001	0.200	0.0271	0.0387	0.0347	4.01
GII #3	0.993	−0.001	0.199	0.0272	0.0385	0.0345	4.81
GII #4	0.991	−0.001	0.199	0.0275	0.0389	0.0347	5.38
$b_1 = 1, r = 0.4, b_2 = 0.2$							
GII #1	0.994	0.397	0.198	0.0361	0.0518	0.0493	3.99
GII #2	0.991	0.397	0.197	0.0309	0.0363	0.0430	4.00
GII #3	0.990	0.396	0.196	0.0306	0.0317	0.0399	4.80
GII #4	0.987	0.395	0.196	0.0302	0.0318	0.0400	5.35
$b_1 = 1, r = 0.85, b_2 = 0.2$							
GII #1	0.993	0.851	0.184	0.0936	0.0403	0.1289	4.41
GII #2	0.986	0.851	0.191	0.0546	0.0249	0.0905	4.37
GII #3	0.987	0.850	0.189	0.0430	0.0140	0.0598	4.93
GII #4	0.984	0.849	0.185	0.0411	0.0136	0.0597	5.56

Table 8
Monte Carlo results for Model 4 ($b_{10} = 0, b_{11} = 1, b_{12} = 1, b_{20} = 0, b_{21} = 1, b_{22} = 1, c_1 = 0, c_2 = 1$).

	SML		GII				$\sigma_{GII}/\sigma_{SML}$			
	#1	#2	#1	#2	#3	#4	#1	#2	#3	#4
Mean										
b_{10}	0.007	0.005	0.003	0.002	0.002	0.002	–	–	–	–
b_{11}	1.000	1.001	0.995	0.994	0.992	0.990	–	–	–	–
b_{12}	1.000	1.003	0.998	0.997	0.995	0.992	–	–	–	–
b_{20}	−0.001	−0.003	−0.006	−0.004	−0.004	0.004	–	–	–	–
b_{21}	1.006	1.007	1.001	0.999	0.997	0.996	–	–	–	–
b_{22}	1.005	1.007	1.004	1.000	0.998	0.996	–	–	–	–
c_1	0.020	0.010	0.007	0.005	0.005	0.006	–	–	–	–
c_2	1.004	1.003	1.006	1.001	1.001	1.002	–	–	–	–
Std. dev.										
b_{10}	0.0630	0.0628	0.0720	0.0666	0.0656	0.0665	1.15	1.06	1.04	1.06
b_{11}	0.0686	0.0686	0.0872	0.0764	0.0741	0.0743	1.27	1.11	1.08	1.08
b_{12}	0.0572	0.0574	0.0719	0.0667	0.0632	0.0646	1.25	1.16	1.10	1.13
b_{20}	0.0663	0.0657	0.0745	0.0686	0.0677	0.0676	1.13	1.04	1.04	1.03
b_{21}	0.1065	0.1050	0.1395	0.1128	0.1095	0.1099	1.33	1.07	1.04	1.05
b_{22}	0.1190	0.1174	0.1593	0.1285	0.1249	0.1244	1.36	1.09	1.06	1.06
c_1	0.1091	0.1107	0.1303	0.1276	0.1224	0.1265	1.18	1.15	1.11	1.14
c_2	0.1352	0.1325	0.1991	0.1509	0.1439	0.1421	1.50	1.14	1.09	1.07
Time	11.5	23.1	7.1	10.4	16.4	34.1	–	–	–	–

where $\xi_i \sim_{\text{i.i.d.}} N[0, \Sigma_\xi]$. We conduct Monte Carlo experiments using four nested versions of the auxiliary model. In auxiliary model #1, $z_i = (1, x_{i1}, x_{i2}, x_{i3})$, giving a total of 11 parameters (including 3 error covariance parameters). This is the auxiliary model that we used in the analysis of Section 5.1.3.

Auxiliary model #2 adds all the second-order products of these variables, as well as one third-order product to z_i, i.e.

$$z_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i1}^2, x_{i2}^2, x_{i3}^2, x_{i1}x_{i2}, x_{i1}x_{i3}, x_{i2}x_{i3}, x_{i1}x_{i2}x_{i3}),$$

for a total of 25 parameters. In auxiliary model #3, z_i contains all third-order products (for a total of 43 parameters) and in auxiliary model #4, z_i contains all fourth-order products (for a total of 67 parameters).

Tables 8 and 9 present the results of six sets of Monte Carlo experiments, each with 1000 replications; the labels SML #i and GII #i are to be interpreted exactly as for Table 5. The key findings are qualitatively similar to those for Models 1, 2, and 3. First, both SML and GII generate estimates with very little bias. Second, auxiliary model #1, which contains only linear terms, leads to large efficiency losses relative to SML (as large as 50%). But auxiliary model #2, which contains terms up to second order, reduces the efficiency losses substantially (to no more than 15% when the errors are uncorrelated, and to no more than 26% when $c = 1.33$). Auxiliary model #3, which contains terms up to third order, provides additional small efficiency gains (the largest efficiency loss is reduced to 20%), while auxiliary model #4, which contains fourth-order terms, provides few, if any, efficiency gains relative to auxiliary model #3. Finally, computing estimates using GII with auxiliary model #3 takes about 30% less time than computing estimates using SML with 50 draws.

Table 9
Monte Carlo results for Model 4 ($b_{10} = 0$, $b_{11} = 1$, $b_{12} = 1$, $b_{20} = 0$, $b_{21} = 1$, $b_{22} = 1$, $c_1 = 1.33$, $c_2 = 1$).

	SML		GII				$\sigma_{GII}/\sigma_{SML}$			
	#1	#2	#1	#2	#3	#4	#1	#2	#3	#4
Mean										
b_{10}	−0.031	−0.017	0.000	−0.001	−0.000	−0.001	–	–	–	–
b_{11}	0.998	1.000	0.993	0.993	0.991	0.989	–	–	–	–
b_{12}	1.016	1.011	0.998	0.998	0.996	0.994	–	–	–	–
b_{20}	−0.011	−0.010	−0.011	−0.007	−0.007	−0.006	–	–	–	–
b_{21}	0.992	0.999	1.000	0.997	0.995	0.991	–	–	–	–
b_{22}	1.004	1.008	1.006	1.001	0.999	0.995	–	–	–	–
c_1	1.269	1.306	1.347	1.338	1.335	1.330	–	–	–	–
c_2	1.025	1.011	0.993	0.993	0.995	0.997	–	–	–	–
Std. dev.										
b_{10}	0.0693	0.0698	0.0789	0.0776	0.0758	0.0757	1.13	1.11	1.09	1.08
b_{11}	0.0587	0.0588	0.0696	0.0658	0.0632	0.0636	1.18	1.12	1.07	1.08
b_{12}	0.0745	0.0737	0.0883	0.0801	0.0781	0.0782	1.20	1.09	1.06	1.06
b_{20}	0.0766	0.0764	0.0900	0.0801	0.0786	0.0780	1.18	1.05	1.03	1.02
b_{21}	0.0884	0.0886	0.1140	0.0969	0.0952	0.0943	1.29	1.09	1.07	1.06
b_{22}	0.1106	0.1103	0.1471	0.1204	0.1176	0.1153	1.34	1.09	1.07	1.05
c_1	0.1641	0.1707	0.2454	0.2152	0.2049	0.2041	1.44	1.26	1.20	1.20
c_2	0.1229	0.1206	0.1599	0.1387	0.1338	0.1311	1.33	1.15	1.11	1.09
Time	12.7	25.6	7.4	10.8	17.1	34.4	–	–	–	–

6. Conclusion

Discrete choice models play an important role in many fields of economics, from labor economics to industrial organization to macroeconomics. Unfortunately, these models are usually quite challenging to estimate (except in special cases like MNL where choice probabilities have closed forms). Simulation-based methods like SML and MSM have been developed that can be used for more complex models like MNP. But in many important cases (models with initial conditions problems and Heckman selection models being leading cases) even these methods are very difficult to implement.

In this paper we develop and implement a new simulation-based method for estimating models with discrete or mixed discrete/continuous outcomes. The method is based on indirect inference. But the traditional II approach is not easily applicable to discrete choice models because one must deal with a non-smooth objective surface. The key innovation here is that we develop a generalized method of indirect inference (GII), in which the auxiliary models that are estimated on the actual and simulated data may differ (provided that the estimates from both models share a common probability limit). This allows us to chose an auxiliary model for the simulated data such that we obtain an objective function that is a smooth function of the structural parameters. This smoothness renders GII practical as a method for estimating discrete choice models. Our theoretical analysis shows that the GII estimator enjoys the same limiting distribution as the unsmoothed II estimator. Inferences based on the GII estimates may thus be drawn in the standard manner, via the usual Wald statistics. Moreover, the GII estimator can be computed using standard derivative-based optimizers, provided that the smoothing is done in such a way that the first (and possibly second) derivatives of the smoothed GII criterion remain consistent for their population counterparts.

We also provide a set of Monte Carlo experiments to illustrate the practical usefulness of GII. In addition to being fast and straightforward to compute, GII yields estimates with good properties in small samples. In particular, the estimates display very little bias and are nearly as efficient as maximum likelihood (in those cases where simulated versions of maximum likelihood can be used) provided that the auxiliary model is chosen judiciously.

GII could potentially be applied to a wide range of discrete and discrete/continuous outcome models beyond those we consider in our Monte Carlo experiments. Indeed, GII is sufficiently flexible to accommodate almost any conceivable model of discrete choice, including, discrete choice dynamic programming models, discrete dynamic games, etc. We hope that applied economists from a variety of fields find GII a useful and easy-to-implement method for estimating discrete choice models.

Appendix A. Extensions and refinements

A.1. A modified smoothing procedure for dynamic models

For models in which latent utilities depend on past choices (as distinct from past *utilities*, which are already smooth), such as Models 2 and 3, the performance of GII may be improved by making a further adjustment to the smoothing proposed in Section 3.3. The nature of this adjustment is best illustrated in terms of the example provided by Model 2. In this case, it is clear that setting

$$y_{it}^m(\beta, \lambda) := K_\lambda[b_1 x_{it} + b_2 y_{i,t-1}^m(\beta) + \epsilon_{it}^m],$$

where $y_{i,t-1}^m(\beta)$ denotes the *unsmoothed* choice made at date $t-1$, will yield unsatisfactory results, insofar as the $y_{it}^m(\beta, \lambda)$ so constructed will remain discontinuous in β. To some extent, this may be remedied by modifying the preceding to

$$y_{it}^m(\beta, \lambda) := K_\lambda[b_1 x_{it} + b_2 y_{i,t-1}^m(\beta, \lambda) + \epsilon_{it}^m], \tag{A.1}$$

with $y_{i0}^m(\beta, \lambda) := 0$, as per the specification of the model. However, while the $y_{it}^m(\beta, \lambda)$'s generated through this recursion will indeed be smooth (i.e., twice continuously differentiable), the nesting of successive approximations entailed by (A.1) implies that for large t, the derivatives of $y_{it}^m(\beta, \lambda)$ may be highly irregular unless a relatively large value of λ is employed.

This problem may be avoided by instead computing $y_{it}^m(\beta, \lambda)$ as follows. Defining $v_{itk}^m(\beta) := b_1 x_{it} + b_2 \mathbf{1}\{k = 1\} + \epsilon_{it}^m$, we see that the *unsmoothed* choices satisfy

$$y_{it}(\beta) = \mathbf{1}\{v_{it0}^m(\beta) \geq 0\} \cdot [1 - y_{i,t-1}(\beta)] + \mathbf{1}\{v_{it1}^m(\beta) \geq 0\} \cdot y_{i,t-1}(\beta),$$

which suggests using the following recursion for the smoothed choices,

$$y_{it}^m(\beta, \lambda) := K_\lambda[v_{it0}^m(\beta)] \cdot [1 - y_{i,t-1}^m(\beta, \lambda)] + K_\lambda[v_{it1}^m(\beta)] \cdot y_{i,t-1}^m(\beta, \lambda), \tag{A.2}$$

with $y_{i0}^m(\beta, \lambda) := 0$. This indeed yields a valid approximation to $y_{it}(\beta)$, as $\lambda \to 0$. The smoothed choices computed using (A.2) involve no nested approximations, but merely sums of products involving terms of the form $K_\lambda[v_{isk}^m(\beta)]$. The derivatives of these are well-behaved with respect to λ, even for large t, and are amenable to the theoretical analysis of Section 4.

Nonetheless, we find that even if smoothing is done by simply using (A.1), GII appears to work well in practice; this is shown in the simulation exercises reported in Section 5.

A.2. Smoothing parameter selection

Estimating bias and variance. Let $\beta^* \in B$ be an fixed value for the structural parameters, which may come e.g. from an initial optimization of Q_{nk}, with a suitably large $\lambda^{(0)}$; suppose for simplicity that no jackknifing is used. Ideally, we would like to compute both the bias and variance with which

$$\overline{\theta}_n(\beta^*, \lambda) := \frac{1}{M} \sum_{m=1}^M \hat{\theta}_n^m(\beta^*, \lambda) \tag{A.3}$$

estimates the (unsmoothed) binding function $\theta(\beta^*, 0)$. In practice, since $\theta(\beta^*, 0)$ is unknown, we will settle for computing the bias of $\overline{\theta}_n(\beta^*, \lambda)$ relative to the closely related quantity

$$h_n(\beta^*) := \frac{1}{M^*} \sum_{m=1}^{M^*} \hat{\theta}_n^m(\beta^*, 0) = \mathbb{E}_\eta[\hat{\theta}_n(\beta^*, 0)] + o_p(1)$$

where the second equality holds as $M^* \to \infty$, and $\mathbb{E}_\eta$ denotes an expectation computed only with respect to the unobserved components η of the structural model.

We may then proceed as follows. Choose a 'large' value of M^*, on the order of $M^* = 500$ or so. (This will typically be much larger than the value of M appearing in (A.3).) The bias in $\overline{\theta}_n(\beta^*, \lambda)$, relative to h_n, may be estimated by

$$\hat{b}(\beta^*, \lambda) := \left\| \frac{1}{M^*} \sum_{m=1}^{M^*} [\hat{\theta}_n^m(\beta^*, \lambda) - h_n(\beta^*)] \right\| = \left\| \frac{1}{M^*} \sum_{m=1}^{M^*} [\hat{\theta}_n^m(\beta^*, \lambda) - \hat{\theta}_n^m(\beta^*, 0)] \right\|,$$

where, by analogy with cross-validation, independent draws may be used to compute $\hat{\theta}_n^m(\beta^*, \lambda)$ and $\hat{\theta}_n^m(\beta^*, 0)$ for each m. An estimate of the total variance is

$$\hat{\sigma}^2(\beta^*, \lambda) = \frac{1}{M^*} \sum_{m=1}^{M^*} \left\| \hat{\theta}_n^m(\beta^*, \lambda) - \frac{1}{M^*} \sum_{m=1}^M \hat{\theta}_n^m(\beta^*, \lambda) \right\|^2.$$

Given that M replications are used to compute $\overline{\theta}_n(\beta^*, \lambda)$ itself, the mean squared error of the binding function could be computed as

$$\widehat{\mathrm{MSE}}_n(\beta^*, \lambda) := \hat{b}^2(\beta^*, \lambda) + \frac{1}{M} \hat{\sigma}^2(\beta^*, \lambda). \tag{A.4}$$

A similar approach may be taken to estimate the bias and variance of the *derivatives* $\partial_\beta \overline{\theta}_n(\beta^*, \lambda)$. In this case, the preceding formulas remain valid, except that $\partial_\beta \hat{\theta}_n^m(\beta^*, \lambda)$ takes the place of $\hat{\theta}_n^m(\beta^*, \lambda)$, and h_n must be replaced by some finite-differencing approximation to $\partial_\beta h_n(\beta^*)$. (If M^* is taken large enough, the problems posed, for finite-differencing, by the non-smoothness of $h_n(\beta)$ are eventually overcome, even though $\lambda = 0$.)

Automated smoothing parameter selection. Given a choice of β^*, we could choose the λ that minimizes the estimated mean squared error (A.4), either for the level of the binding function, or its first derivatives. One would expect –e.g. by

analogy with cdf and density estimation –that the former will generally give much smaller 'optimal' values of λ than will the latter. Thus if a conservative initial value of λ is sought, minimization of the derivatives' MSE may be more suitable. On the other hand, the Monte Carlo simulations reported in Section 5.1.1 indicate that the λ that minimizes the MSE of (level of) the binding function may yield estimates with relatively little bias.

Termination. The alternative stopping rule proposed in Section 3.5 may be implemented simply by checking whether

$$\hat{b}(\hat{\beta}_n(\lambda^{(i)}), \lambda^{(i)}) \le \delta \hat{\sigma}(\hat{\beta}_n(\lambda^{(i)}), \lambda^{(i)})$$

where $\delta \in (0, 1)$ is some pre-specified quantity, e.g. $\delta = 0.05$.

Appendix B. Low-level conditions

This appendix provides some low-level conditions (Assumption L) which are sufficient for the high-level conditions (Assumption H) given in H. We subsequently verify that these conditions are satisfied by each of Models 1–5, when the auxiliary model is a Gaussian SUR; more generally, it should be straightforward to verify our conditions for any dynamic discrete choice model satisfying (2.4)–(2.5).

B.1. A general framework for models with smoothed outcomes

We first introduce the following framework, which is sufficiently general to encompass both the dynamic discrete choice models of Section 2, as well as models with mixed discrete/continuous regressors (such as Model 5).

Data. Individual i is described by vectors $x_i \in \mathbb{R}^{d_x}$ and $\eta_i \in \mathbb{R}^{d_\eta}$ of observable and unobservable characteristics; x_i collects *all* the covariates appearing in the structural and auxiliary models. η_i is a vector of independent variates that are also independent of x_i, and normalized to have unit variance. Their marginal distributions are fully specified by the model, allowing them to be simulated. Collect $z_i := (x_i^\mathsf{T}, \eta_i^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{d_z}$, and define the projections $[x(\cdot), \eta(\cdot)]$ so that $(x_i, \eta_i) = [x(z_i), \eta(z_i)]$. Individual i has a vector $y(z_i; \beta, \lambda) \in \mathbb{R}^{d_y}$ of smoothed outcomes, parametrized by $(\beta, \lambda) \in B \times \Lambda$, with $\lambda = 0$ corresponding to the true, unsmoothed outcomes under β. At this level of abstraction, we need not make any notational distinction between choices made by an individual at the same date (over competing alternatives), vs. choices made at distinct dates; we note simply that each corresponds to some element of $y(\cdot)$. With this notation, the mth simulated choices may be written as $y(z_i^m; \beta, \lambda)$; since the same x_i's are used across all simulations, we have $x(z_i^m) = x(z_i^{m'})$ but $\eta(z_i^m) \neq \eta(z_i^{m'})$ for $m' \neq m$.

Auxiliary model. We shall assume that the auxiliary model takes the form of a system of seemingly unrelated regressions (SUR; see e.g. Section 10.2 in Greene, 2008)

$$y_r(z_i; \beta, \lambda) = \alpha_{xr}^\mathsf{T} \Pi_{xr} x(z_i) + \alpha_{yr}^\mathsf{T} \Pi_{yr} y(z_i; \beta, \lambda) + \xi_{ri}, \tag{B.1}$$

where $\xi_i := (\xi_{1i}, \ldots, \xi_{d_y i})^\mathsf{T} \sim_{\text{i.i.d.}} N[0, \Sigma_\xi]$, and Π_{xr} and Π_{yr} are selection matrices (i.e. matrices that take at most one unit value along each row, and have zeros everywhere else); let $\alpha_r := (\alpha_{xr}^\mathsf{T}, \alpha_{yr}^\mathsf{T})^\mathsf{T}$. Typically, Σ_ξ will be assumed block diagonal: for example, we may only allow those ξ_{ri}'s pertaining to alternatives from the same period to be correlated. The auxiliary parameter vector θ collects a subset of the elements of $(\alpha_1^\mathsf{T}, \ldots, \alpha_{d_y}^\mathsf{T})^\mathsf{T}$ and those of Σ_ξ^{-1}. (For the calculations involving the score vector in Section E of the Supplementary Material, it shall be more convenient to treat the model as being parametrized in terms of Σ_ξ^{-1}, than Σ_ξ.)

Several estimators of θ are available, most notably OLS, feasible GLS, and maximum likelihood, all of which agree only under certain conditions.[20] For concreteness, we shall assume that both the data-based and simulation-based estimates of θ are produced by maximum likelihood. However, our results easily extend to the case where these estimates are computed using OLS or feasible GLS. (In those cases, the auxiliary estimator can be still be written as a function of a vector of sufficient statistics, a property that greatly facilitates the proofs of our results.)

Smoothed indices. We shall also need to restrict the manner in which $y(\cdot)$ is parametrized. To that end, we introduce the following collections of linear indices

$$\nu_r(z; \beta) := z^\mathsf{T} \Pi_{\nu r} \gamma(\beta) \qquad r \in \{1, \ldots, d_\nu\} \tag{B.2a}$$
$$\omega_r(z; \beta) := z^\mathsf{T} \Pi_{\omega r} \gamma(\beta) \qquad r \in \{1, \ldots, d_\omega\}, \tag{B.2b}$$

where $\gamma : B \to \Gamma$, and $\Pi_{\nu r}$ and $\Pi_{\omega r}$ are selection matrices. Let $d_c \ge d_\omega$; for each $r \in \{1, \ldots, d_c\}$, let $\mathcal{S}_r \subseteq \{1, \ldots, d_\nu\}$ and define

$$\tilde{y}_r(\beta, \lambda) := \omega_r(\beta) \cdot \prod_{s \in \mathcal{S}_r} K_\lambda[\nu_s(\beta)] \tag{B.3}$$

---

[20] In Section 5, exact numerical agreement between these estimators is ensured by requiring the auxiliary model equations referring to alternatives from the same period to have the same set of regressors.

collecting these in the vector $\tilde{y}(\beta, \lambda)$; these are products of both smoothed and unsmoothed linear indices. We shall require that the smoothed choices y are themselves linear combinations of elements of $\tilde{y}(\beta, \lambda)$; see L3. $K : \mathbb{R} \to [0, 1]$ is a smooth univariate cdf, and $K_\lambda(v) := K(\lambda^{-1}v)$. Note that $d_c \geq d_\omega$, and that we have defined

$$\omega_r(z; \beta) := 1 \qquad r \in \{d_\omega + 1, \dots, d_c\}. \tag{B.4}$$

Low-level conditions. Let $\eta_{\omega i} := \Pi_{\eta \omega} \eta_i$ select the elements of η_i upon which ω actually depends (as determined by the $\Pi_{\omega r}$ matrices), and let $W_r \geq 1$ denote an envelope for ω_r, in the sense that $|\omega_r(z; \beta)| \leq W_r(z)$ for all $\beta \in$ B. Let $\varrho_{\min}(A)$ denote the smallest eigenvalue of a symmetric matrix A.

Assumption L (*Low-level Conditions*).

L1 η_i^m and x_i are mutually independent, and i.i.d. over i and m;

L2 K in (B.3) is a twice continuously differentiable cdf, for a distribution having integer moments of all orders, and density $\dot{K}$ symmetric about the origin;

L3 $y(\beta, \lambda) = D\tilde{y}(\beta, \lambda)$ for some $D \in \mathbb{R}^{d_y \times d_c}$, for $\tilde{y}$ as in (B.3);

L4 $\gamma : $ B $\to \Gamma$ in (B.2) is twice continuously differentiable;

L5 for each $k \in \{1, \dots, d_\eta\}$, $\text{var}(\eta_{ki}) = 1$, and η_{ki} has a density f_k with

$$\sup_{u \in \mathbb{R}} (1 + |u|^4) f_k(u) < \infty;$$

L6 there exists an $\epsilon > 0$ such that, for every $r \in \{1, \dots, d_v\}$ and $\beta \in$ B,

$$\text{var}(\nu_r(z_i; \beta) \mid \eta_{\omega i}, x_i) \geq \epsilon;$$

L7 there exists a $p_0 \geq 2$ such that for each $r \in \{1, \dots, d_c\}$, $\mathbb{E}(W_r^4 + \|z_i\|^4) < \infty$, $\mathbb{E}|W_r\|z_i\|^3|^{p_0} < \infty$ and $\mathbb{E}|W_r^2\|z_i\|^2|^{p_0} < \infty$;

L8 $\inf_{(\beta, \lambda) \in \text{B} \times \Lambda} \varrho_{\min}[\mathbb{E}\bar{y}(z_i; \beta, \lambda)\bar{y}(z_i; \beta, \lambda)^\mathsf{T}] > 0$, where $\bar{y}(\beta, \lambda) := [y(\beta, \lambda)^\mathsf{T}, x^\mathsf{T}]^\mathsf{T}$; and

L9 the auxiliary model is a Gaussian SUR, as in (B.1).

Remark B.1. (B.2) entails that the estimator criterion function Q_n depends on β only through $\gamma(\beta)$, i.e. $Q_n(\beta) = \tilde{Q}_n(\gamma(\beta))$ for some $\tilde{Q}_n$. Since the derivatives of $\tilde{Q}_n$ with respect to γ take a reasonably simple form, in Section E of the Supplementary Material we establish the convergence of $\partial_\beta^l Q_n$ to $\partial_\beta^l Q$, for $l \in \{1, 2\}$, by first proving the corresponding result for $\partial_\gamma^l \tilde{Q}_n$ and then applying the chain rule.

Remark B.2. Assumption L is least restrictive in models with purely discrete outcomes, for which we may take $d_\omega = 0$. In particular, L7 reduces to the requirement that $\mathbb{E}\|z_i\|^{3p_0} < \infty$.

Remark B.3. As the examples discussed immediately below illustrate, except in the case where current (discrete) choices depend on past choices, it is generally possible to take $D = I_{d_y}$ in L3, so that $y(\beta, \lambda) = \tilde{y}(\beta, \lambda)$.

B.2. Verification for Models 1–5

We may verify that each of the models from Section 2 satisfies L3–L7. Note that L2 will be satisfied for many standard choices of K, such as the Gaussian cdf, and many smooth, compactly supported kernels. In all cases, x_i collects all the (unique) elements of $\{x_{it}\}_{t=1}^T$, together with any additional exogenous covariates used to estimate the auxiliary model; while η_i collects the elements of $\{\eta_{it}\}_{t=1}^T$. Note that for the discrete choice Models 1–4, since the η_i are Gaussian L7 will be satisfied if $\mathbb{E}\|x_i\|^{3p_0} < \infty$. L8 is a standard non-degeneracy condition.

Model 1. $u_{it} = bx_{it} + \sum_{s=1}^t r^{t-s}\eta_{is}$ by backward substitution. So we set $(d_v, d_\omega) = (T, 0)$, with

$$\nu_t(z_i; \beta) = x_t(z_i)b(\beta) + \sum_{s=1}^t \eta_s(z_i)d_{ts}(\beta),$$

where $\beta = (b, r)$, $b(\beta) = b$ and $d_{ts}(\beta) = r^{t-s}$; while $x_t(z_i)$ and $\eta_s(z_i)$ select the appropriate elements of z_i, which collects $\{x_{it}\}$, $\{\eta_{it}\}$, and any other exogenous covariates used in the auxiliary model. Thus L3 and L4 hold (formally, take $\gamma(\beta) = (b(\beta), \{d_{ts}(\beta)\})$). L6 follows from the $\eta_t(z_i)$'s being standard Gaussian.

Model 2. As per the discussion in Appendix A.1, and (A.2) in particular, we define

$$\nu_{tk}(z_i; \beta) := x_t(z_i)b_1(\beta) + b_2(\beta)\mathbf{1}\{k = 1\} + \sum_{s=1}^t \eta_s(z_i)d_{ts}(\beta)$$

where the right-hand side quantities are defined by analogy with the preceding example. Setting

$$y_t(\beta, \lambda) := K_\lambda[\nu_{t0}(\beta)] \cdot [1 - y_{t-1}(\beta, \lambda)] + K_\lambda[\nu_{t1}(\beta)] \cdot y_{t-1}(\beta, \lambda) \tag{B.5}$$

with $y_0(\beta, \lambda) := 0$ thus yields smoothed choices having the form required by L3 and L4, as may be easily verified by backwards substitution. L6 again follows from Gaussianity of $\eta_t(z_i)$.

An identical recursion to (B.5) also works for Model 3. Model 4 may be handled in a similar way to Model 1, but it is in certain respects simpler, because the errors are not serially dependent.

Model 5. From the preceding examples, it is clear that $\omega(z_i; \beta) = w_i$ and $\nu(z_i; \beta) = u_i$ can be written in the linear index form (B.2). The observable outcomes are the individual's decision to work, and also his wage if he decides to work. These may be smoothly approximated by:

$$y_1(\beta, \lambda) := K_\lambda[\nu(\beta)] \qquad\qquad y_2(\beta, \lambda) := \omega(\beta) \cdot K_\lambda[\nu(\beta)]$$

respectively. Thus L3–L6 hold just as in the other models. L7 holds, in this case, if $\mathbb{E}\|z_i\|^{4p_0} < \infty$.

Appendix C. Details of optimization routines

We consider two popular line-search optimization methods – Gauss–Newton, and quasi-Newton with BFGS updating –and a trust-region algorithm. When applied to a criterion Q, each of these routines proceeds as follows: given an iterate $\beta^{(s)}$, locally approximate $Q(\beta)$ by the following quadratic model,

$$f_{(s)}(\beta) := Q(\beta^{(s)}) + \nabla_{(s)}^\mathsf{T}(\beta - \beta^{(s)}) + \tfrac{1}{2}(\beta - \beta^{(s)})^\mathsf{T}\Delta_{(s)}(\beta - \beta^{(s)}), \tag{C.1}$$

where $\nabla_{(s)} := \partial_\beta Q(\beta^{(s)})$. A new iterate $\beta^{(s+1)}$ is then generated by approximately minimizing $f_{(s)}$ with respect to β. The main differences between these procedures concern the choice of approximate Hessian $\Delta_{(s)}$, and the manner in which $f_{(s)}$ is (approximately) minimized.

Both line-search methods (Gauss–Newton and quasi-Newton) involve the use of a positive definite Hessian $\Delta_{(s)}$ in the approximating model (C.1), and so the problem solved at step $s + 1$ reduces to that of "approximately" solving

$$\min_{\alpha \in \mathbb{R}} Q(\beta^{(s)} + \alpha p_{(s)}), \tag{C.2}$$

where $p_{(s)} := -\Delta_{(s)}^{-1}\nabla_{(s)}$. We do not require that $\alpha_{(s)}$ solves (C.2) exactly; we require only that it satisfies the strong Wolfe conditions,

$$Q(\beta^{(s)} + \alpha_{(s)}p_{(s)}) \le Q(\beta^{(s)}) + c_1\alpha_{(s)}\nabla_{(s)}^\mathsf{T}p_{(s)}$$
$$|\dot{Q}(\beta^{(s)} + \alpha_{(s)}p_{(s)})^\mathsf{T}p_{(s)}| \le c_2|\nabla_{(s)}^\mathsf{T}p_{(s)}|$$

for $0 < c_1 < c_2 < 1$, where $\dot{Q} := \partial_\beta Q$ (cf. (3.7) in Nocedal and Wright, 2006). For some such $\alpha_{(s)}$, we set $\beta^{(s+1)} = \beta^{(s)} + \alpha_{(s)}p_{(s)}$. For the Hessians $\Delta_{(s)}$, the Gauss–Newton method is only applicable to criteria of the form $Q(\beta) = \frac{1}{2}\|g(\beta)\|_W^2$, and uses

$$\Delta_{(s)} = G_{(s)}^\mathsf{T}WG_{(s)}, \tag{C.3}$$

where $G_{(s)} := [\partial_\beta g(\beta^{(s)})]^\mathsf{T}$. The Quasi-Newton method with BFGS updating starts with some initial positive definite $\Delta_{(0)}$, and updates it according to,

$$\Delta_{(s+1)} = \Delta_{(s)} - \frac{\Delta_{(s)}x_{(s)}x_{(s)}^\mathsf{T}\Delta_{(s)}}{x_{(s)}^\mathsf{T}\Delta_{(s)}x_{(s)}} + \frac{d_{(s)}d_{(s)}^\mathsf{T}}{d_{(s)}^\mathsf{T}x_{(s)}},$$

where $x_{(s)} := \alpha_{(s)}p_{(s)}$ and $d_{(s)} = \nabla^{(s+1)} - \nabla^{(s)}$ (cf. (6.19) in Nocedal and Wright, 2006).

The trust region method considered here sets $\Delta_{(s)} = \partial_\beta^2 Q(\beta^{(s)})$, which need not be positive definite. The procedure then attempts to approximately minimize (C.1), subject to the constraint that $\|\beta\| \le \delta_{(s)}$, where $\delta_{(s)}$ defines the size of the trust region, which is adjusted at each iteration depending on the value of

$$\rho_{(s)} := \frac{Q(\beta^{(s)}) - Q(\beta^{(s+1)})}{f_{(s)}(0) - f_{(s)}(\beta^{(s+1)})},$$

which measures the proximity of the true reduction in Q at step s, with that predicted by the approximating model (C.1); the adjustment is made in accordance with Algorithm 4.2 in Moré and Sorensen (1983). Various algorithms are available for approximately solving (C.1) in this case, but we shall assume that Algorithm 3.14 from that paper is used.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.03.010.

References

Altonji, J.G., Segal, L.M., 1996. Small-sample bias in GMM estimation of covariance structures. J. Bus. Econom. Statist. 14 (3), 353–366.

Altonji, J.G., Smith, A.A., Vidangos, I., 2013. Modeling earnings dynamics. Econometrica 81 (4), 1395–1454.

An, M.Y., Liu, M., 2000. Using indirect inference to solve the initial-conditions problem. Rev. Econ. Stat. 82 (4), 656–667.

Cassidy, H., 2012. Skills, tasks, and occupational choice, University of Western Ontario.

Chernozhukov, V., Hong, H., 2003. An MCMC approach to classical estimation. J. Econometrics 115 (2), 293–346.

Diggle, P.J., Gratton, R.J., 1984. Monte Carlo methods of inference for implicit statistical models. J. Roy. Statist. Soc. Ser. B 46 (2), 193–227.

Eisenhauer, P., Heckman, J.J., Mosso, S., 2015. Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments. Internat. Econom. Rev. 56 (2), 331–357.

Fermanian, J.-D., Salanié, B., 2004. A nonparametric simulated maximum likelihood estimation method. Econometric Theory 20 (4), 701–734.

Gallant, A.R., Tauchen, G., 1996. Which moments to match? Econometric Theory 12 (4), 657–681.

Gan, L., Gong, G., 2007. Estimating interdependence between health and education in a dynamic model, Working Paper 12830, National Bureau of Economic Research.

Genton, M.G., Ronchetti, E., 2003. Robust indirect inference. J. Amer. Statist. Assoc. 98 (461), 67–76.

Geweke, J., Keane, M.P., 2001. Computationally intensive methods for integration in econometrics. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, Vol. 5. Elsevier.

Gourieroux, C., Monfort, A., Renault, E., 1993. Indirect inference. J. Appl. Econometrics 8 (S1), S85–S118.

Greene, W.H., 2008. Econometric Analysis, sixth ed.. Pearson Prentice Hall, New Jersey (USA).

Heckman, J.J., 1981. In: Manski, C.F., McFadden, D. (Eds.), The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–discrete Data Stochastic Process. MIT Press, Cambridge, MA, pp. 179–195.

Horowitz, J.L., 1992. A smoothed maximum score estimator for the binary response model. Econometrica 60 (3), 505–531.

Horowitz, J.L., 1998. Bootstrap methods for median regression models. Econometrica 66 (6), 1327–1351.

Kaplan, D.M., Sun, Y., 2012. Smoothed estimating equations for instrumental variables quantile regression, University of California, San Diego.

Keane, M.P., 1994. A computationally practical simulation estimator for panel data. Econometrica 62, 95–116.

Keane, M.P., Sauer, R.M., 2010. A computationally practical simulation estimation algorithm for dynamic panel data models with unobserved endogenous state variables. Internat. Econom. Rev. 51 (4), 925–958.

Keane, M., Smith, A.A., 2003. Generalized indirect inference for discrete choice models, Yale University.

Kim, J., Pollard, D., 1990. Cube root asymptotics. Ann. Statist. 18 (1), 191–219.

Kristensen, D., Salanié, B., 2016. Higher-order properties of approximate estimators, Columbia University.

Kristensen, D., Shin, Y., 2012. Estimation of dynamic models with nonparametric simulated maximum likelihood. J. Econometrics 167 (1), 76–94.

Lee, L.-F., 1997. Simulated maximum likelihood estimation of dynamic discrete choice statistical models: some Monte Carlo results. J. Econometrics 82 (1), 1–35.

Lerman, S., Manski, C.F., 1981. In: Manski, C.F., McFadden, D. (Eds.), On the use of Simulated Frequencies to Approximate Choice Probabilities. MIT Press, Cambridge, MA, pp. 305–319.

Li, T., Zhang, B., 2015. Affiliation and entry in first-price auctions with heterogeneous bidders: an analysis of merger effects. Amer. Econ. J.: Microecon. 7 (2), 188–214.

Lopez Garcia, I., 2015. Human capital and labor informality in Chile: a life-cycle approach Working Paper WR-1087, RAND Corporation.

Lopez-Mayan, C., 2014. Microeconometric analysis of residential water demand. Environ. Resour. Econ. 59 (1), 137–166.

Magnac, T., Robin, J.-M., Visser, M., 1995. Analysing incomplete individual employment histories using indirect inference. J. Appl. Econometrics 10 (1), S153–S169.

Magnus, J.R., Neudecker, H., 2007. Matrix Differential Calculus with Applications in Statistics and Econometrics, third ed.. Wiley, Chichester (UK).

Manski, C.F., 1985. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. J. Econom. 27 (3), 313–333.

McFadden, D.L., 1989. A method of simulated moments for estimation of discrete response models without numerical integration. Econometrica 57, 995–1026.

Moré, J.J., Sorensen, D.C., 1983. Computing a trust region step. SIAM J. Sci. Stat. Comput. 4 (3), 553–572.

Morten, M., 2013. Temporary migration and endogenous risk sharing in village india, Stanford University.

Nagypál, É., 2007. Learning by doing vs. learning about match quality: Can we tell them apart? Rev. Econom. Stud. 74 (2), 537–566.

Nocedal, J., Wright, S.J., 2006. Numerical Optimization, secnd ed.. Springer.

Otsu, T., 2008. Conditional empirical likelihood estimation and inference for quantile regression models. J. Econometrics 142 (1), 508–538.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57 (5), 1027–1057.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1993. Numerical Recipes: the art of scientific computing, second ed.. Cambridge University Press, Cambridge (UK).

Sauer, R.M., Taber, C., 2013. Indirect inference with importance sampling, Royal Holloway, University of London.

Sidi, A., 2003. Practical Extrapolation Methods: Theory and Applications. Cambridge University Press, Cambridge (UK).

Skira, M.M., 2015. Dynamic wage and employment effects of elder parent care. Internat. Econom. Rev. 56 (1), 63–93.

Smith Jr., A.A., 1990. Three Essays on the Solution and Estimation of Dynamic Macroeconomic Models (Ph.D. thesis), Duke University.

Smith Jr., A.A., 1993. Estimating nonlinear time-series models using simulated vector autoregressions. J. Appl. Econometrics 8 (S1), S63–S84.

Whang, Y.-J., 2006. Smoothed empirical likelihood methods for quantile regression models. Econometric Theory 22 (2), 173–205.

Ypma, J.Y., 2013. Dynamic models of continuous and discrete outcomes; methods and applications (Ph.D. thesis), University College London.

Further reading

Billingsley, P., 1968. Convergence of Probability Measures. Wiley, New York (USA).

Einmahl, U., Mason, D.M., 2005. Uniform in bandwidth consistency of kernel-type function estimators. Ann. Statist. 33 (3), 1380–1403.

Engle, R.F., McFadden, D.L. (Eds.), 1994. Handbook of Econometrics, Vol. IV. Elsevier.

Kosorok, M.R., 2008. Introduction to Empirical Processes and Semiparametric Inference. Springer.

Manski, C.F., McFadden, D. (Eds.), 1981. Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA.

Newey, W.K., McFadden, D.L., 1994. In: Engle, R.F., McFadden, D.L. (Eds.), Large Sample Estimation and Hypothesis Testing, Vol. IV. Elsevier, pp. 2111–2245.

Nolan, D., Pollard, D., 1987. *U*-processes: rates of convergence. Ann. Statist. 15 (2), 780–799.

Pollard, D., 1984. Convergence of Stochastic Processes. Springer, New York (USA).

van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: with applications to statistics. Springer, New York (USA).