Dynamic discrete choice models II

Thierry Kamionka

2021-2022

Indirect inference is an attractive method of estimation. However, for **discrete choice models**, the objective function is not continuous with respect to the parameters.

One cannot use then gradient based optimization methods in order to obtain the value of the estimator.

To work around this problem, it is possible to smooth the objective function with respect to the parameters (see Bruins et al., 2018, Frazier et al., 2019).

The idea is to use a **smoothed function** of the latent utility (or latent variable) as the dependent variable of the auxiliary model.

When the smoothing parameter **tends to zero**, the objective function tends to the one corresponding to indirect inference. The estimator is consistent and as the same limiting distribution as the indirect inference estimator. This estimator can be obtained using gradient based optimization methods.

The model: A dynamic discrete choice model.

The method can be used for other models that include at least one discrete dependent variables and, singularly, to models mixing **discrete and continuous outcomes.**

Let us assume we have panel data (n individuals and T periods of time, n is "large").

Each individual choose among J alternatives.

Let us denote $u_{ijt} = y^*_{ijt}$ the **latent utility** of the individual i for alternative j in period t. The utility associated to alternative J is fixed to 0.

For each period, the individual chooses the alternative associated to the highest utility.

Let y_{itj} denote the binary variable equal to 1 if the individual chooses the alternative j in period t and to 0 otherwise.

Then, one can write

$$y_{itj} = \mathbb{1}[\, u_{itj} \geq \max_{\substack{k \in \{1,\dots,J\} \\ k \neq j}} u_{itk} \,] = \begin{cases} 1, & \text{if } u_{itj} \geq \max_{k \neq j} u_{ijk}, \\ 0, & \text{otherwise.} \end{cases}$$

Let us assume that the latent utilities are such that

$$u_{it} = y_{it}^{*} = f(x_{it}, y_{i,t-1}, \dots, y_{i,t-\ell}, \epsilon_{it}; \beta),$$

where $t = 1, \dots, T$ and $\mathbf{u}_{it} = (u_{it1}, \dots, u_{it,J-1})$.

Let us assume that the error term ϵ_{it} follows a Markov process

$$\epsilon_{it} = g(\epsilon_{i,t-1}, \eta_{it}; \beta),$$

where $t = 1, \dots, T$ and η_{it} is an unobserved i.i.d. random vector. The distribution of η_{it} is assumed not to depend on β and η_{it} is independent of x_{it}.

Remark: The assumption on the error term η_{it} (see below) should be relaxed in order to introduce a random effect.

Let $y_{it} = (y_{it1}, \ldots, y_{it,J-1})$. The econometrician can observe y_{it} but cannot observe the **latent utilities** y^*_{it}. The vector of exogenous variables x_{it} can be observed.

Let us assume that the initial values $y^0_{t=1-\ell}$ (where $\ell \geq 1$ is fixed) and ϵ_{i0} are assumed to be exogeneous.

Example: Let us assume $J = 2$, $T > 1$ and the latent utility is

$$u_{it} = y^*_{it} = b_1\, x_{it} + b_2\, y_{i,t-1} + \epsilon_{it},$$

where $x_{it} \in \boldsymbol{R}$, $\epsilon_{it} = r\, \epsilon_{i,t-1} + \eta_{it}$, η_{it} i.i.d.$\sim N(0,1)$ and $\epsilon_{i0} = 0$.

Moreover, if the econometrician do not observe the first s realizations ($s < T$) of the choices, there is an **initial condition problem** (Heckman, 1981, An and Liu, 2000, Wooldridge 2005).

Let us assume that T is fixed and n is large (asymptotics is considered such that $n \longrightarrow \infty$).

Let $y_i = (y'_{i1}, \ldots, y'_{iT})'$ is the vector of outcomes for individual i.

Let us assume that vector of parameters of the **initial model** $\beta \in \boldsymbol{R}^{d_\beta}$ and that the vector of parameters of the **auxiliary model** $\theta \in \boldsymbol{R}^{d_\theta}$**, where** $d_\beta \leq d_\theta$.

Let us assume that θ is estimated maximizing the **quasi likelihood** associated to the auxiliary model with respect to θ (the initial model is assumed to be the true model and β_0 is the true value of β).

Then $\hat{\theta}_n$ is such that

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

where $\ell(y_i, x_i; \theta)$ is the log of the contribution of individual i to the "likelihood" function (indeed a quasi-likelihood).

Le us denote $\eta^r = (\eta_1^r, \ldots, \eta_n^r)$ a set of draws for the unobserved heterogeneity component, one per individual in the sample. These draws are assumed to be independent across $r = 1, \ldots, H$.

Given $x = (x_1', \ldots, x_n')'$ and β, the **initial model** can be used to obtain H sets of simulated choices $y^r(\beta) = (y_1^r(\beta), \ldots, y_n^r(\beta))$.

Let us remind that the distribution of η_i **does not depend** of β (the assumption could be relaxed).

We can the estimate β using the simulated data $y^r(\beta)$ and the **auxiliary model**

$$\hat{\theta}_n^r(\beta) = \underset{\theta \in \Theta}{argmax}\, \mathcal{L}_n(y^r(\beta), x; \theta).$$

Let us denote

$$\bar{\theta}_n(\beta) = \frac{1}{H} \sum_{r=1}^{H} \hat{\theta}_n^r(\beta).$$

Under regularity assumptions (Gourieroux, Monfort, Renault, 1993), as n becomes large, $\bar{\theta}_n(\beta)$ converges uniformly in probability to $\theta(\beta)$ (the **binding function**).

How to approximate the true value of β (namely, β_0) ? In order to obtain the value of the indirect inference estimator, we have to make the estimate of θ obtained on the observed data set using the auxiliary model (namely, $\hat{\theta}_n$) as close as possible to $\bar{\theta}_n(\beta)$.

We can use one of the following objective function

$Q_n^w(\beta) = || \bar{\theta}_n(\beta) - \hat{\theta}_n ||_{W_n}^2 = (\bar{\theta}_n(\beta) - \hat{\theta}_n)' W_n (\bar{\theta}_n(\beta) - \hat{\theta}_n)$, where W_n is a sequence of positive-definite matrices.

$Q_n^{LR}(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i; \bar{\theta}_n(\beta))$.

$Q_n^{LM}(\beta) = (\frac{1}{H} \sum_{r=1}^{H} \frac{\partial}{\partial \beta} \mathcal{L}_n(y^r(\beta), x; \hat{\theta}_n)' V_n (\frac{1}{H} \sum_{r=1}^{H} \frac{\partial}{\partial \beta} \mathcal{L}_n(y^r(\beta), x; \hat{\theta}_n))$, where V_n is a sequence of positive-definite matrices.

The estimators obtained by minimizing these objective functions with respect to β yield **consistent and asymptotically normal** estimators of β_0.

In the case such that $d_\beta = d_\theta$, these estimator are asymptotically equivalent.

If $d_\theta > d_\beta$ and if the weighting matrices W_n and V_n are chosen optimally, the indirect inference estimators obtained by minimizing $Q_n^w(\beta)$ and $Q_n^{LM}(\beta)$ are more efficient than the one obtained by minimizing $Q_n^{LR}(\beta)$.

If the likelihood function used for the **auxiliary model is correctly specified**, all these three indirect inference estimators are asymptotically equivalent.

If we consider a discrete choice model, the simulated outcome $y^r_{itj}(\beta)$ is discontinuous with respect to β ($y^r_{itj}(\beta) \in \{0, 1\}$), the random draws η^r being fixed.

Let us remind that $y_{itj} = 1$ iff $u_{itj} = \max_{k \neq j} u_{itk}$, where $u_{itj} = f(x_{it}, y_{i,t-1}, \dots, y_{i,t-\ell}, \epsilon_{it}; \beta)$.

Then, the estimation of the **binding function** $\bar{\theta}_n(\beta)$ is **discontinuous** with respect to β.

Bruins et al. (2018) propose a **generalization of indirect inference** consisting to replace $y^r_{itj}(\beta)$ by a smooth function of latent utilities.

Let us remind that

$$y^r_{itj}(\beta) = 1\!\!1[\, u^r_{itj}(\beta) \geq \max_{k \neq j} u^r_{itk}(\beta)\,] = \prod_{k \neq j} 1\!\!1[\, u^r_{itj}(\beta) - u^r_{itk}(\beta) \geq 0\,]$$

We replace the variables $y_{itj}^r(\beta)$ by

$$y_{itj}^r(\beta,\lambda) = K_\lambda[\, u_{itj}^r(\beta) - u_{it1}^r(\beta), \ldots, u_{itj}^r(\beta) - u_{itJ}^r(\beta)\,]$$

where J is the total number of alternatives, K: $\boldsymbol{R}^{J-1} \longrightarrow \boldsymbol{R}$ is a mean-zero multivariate cdf and $K_\lambda(v) = K(\frac{v}{\lambda})$. Let us remark that the alternative j in this definition is compared with all the other alternatives ($k \neq j$).

If the utilities are distinct, as λ goes to 0, then $y_{itj}^r(\beta,\lambda)$ converges to $y_{itj}^r(\beta,0) = y_{itj}^r(\beta)$.

Let us define

$$\hat{\theta}_n^r(\beta,\lambda) = \underset{\theta\in\Theta}{argmax}\, \mathcal{L}_n(y^r(\beta,\lambda), x, \theta),$$

and

$$\bar{\theta}_n(\beta,\lambda) = \frac{1}{H}\sum_{r=1}^{H} \hat{\theta}_n^r(\beta,\lambda).$$

$\bar{\theta}_n(\beta, \lambda)$ can be considered as a smoothed estimate of $\theta(\beta)$ (consistent for $n \to \infty$ and $\lambda \to 0$).

For instance, one can use $K(v) = \frac{1}{1+\sum_{j=1}^{J-1} e^{-v_j}}$ or $K(v) = \prod_{j=1}^{J-1} \Phi(v_j)$.

One of the three objective functions can be used in order to obtain the **Generalized indirect inference** estimator replacing $y_{itj}^r(\beta)$ by $y_{itj}^r(\beta, \lambda_n)$ and $\bar{\theta}_n(\beta)$ by $\bar{\theta}_n(\beta, \lambda_n)$.

Singularly, $y_{itj}^r(\beta, \lambda_n)$ is generated using the **initial model** and smoothing the latent utilities using a **smoothing parameter** λ_n.

If $\lambda_n \longrightarrow 0$ too quickly, the derivates of the objective function will vary a lot and derivative-based optimization algorithm will have difficulties to locate the minimum of the function. Smoothing introduces a bias to the binding function. This bias will be dominated by the variance of the estimator if $n^{1/2}\lambda_n \longrightarrow 0$ (bias will be relatively small). These two **constraints are conflicting**.

Bruin et al. (2018) proposes to use a bias reduction techniques. They propose to use **Jackknifing** (or "**Richardson extrapolation**").

The *kth* order **Jackknifed sample binding function** is

$$\bar{\theta}^k(\beta, \lambda_n) = \sum_{m=0}^{k} \gamma_{mk}\, \bar{\theta}_n(\beta, \delta^m \lambda_n),$$

where $\delta \in (0, 1)$, the weights $(\gamma_{0k}, \ldots, \gamma_{kk})$ are such that $\sum_{m=0}^{k} \gamma_{mk} = 1$ (γ_{mk} can be negative) and can be calculated using Sidi (2003), algorithm 1.3.1 ($H(k+1)$ estimations).

Remark : The un-jackknifed estimator can be obtained by taking $k = 0$.

Let us denote

$$\frac{\partial}{\partial\theta}\mathcal{L}_n^{rk}(\beta, \lambda; \hat{\theta}_n) \equiv \sum_{m=0}^{k} \gamma_{mk}\, \frac{\partial}{\partial\theta}\mathcal{L}_n(y^r(\beta, \delta^m \lambda), x, \hat{\theta}_n),$$

the **jackknifed score function.**

k is the order of **the extrapolation**, r is a random draw (a vector $y^r(\beta,\lambda)$).

Remark: The extrapolation of the order $(k+1)th$ involves H more optimizations compared to extrapolation of the order kth.

The jackknifed GII estimators of the order k are the minimizers of

$Q^w_{nk}(\beta,\lambda_n) = \| \bar{\theta}^k_n(\beta,\lambda_n) - \hat{\theta}_n \|^2_{W_n} = (\bar{\theta}^k_n(\beta,\lambda_n) - \hat{\theta}_n)' \, W_n \, (\bar{\theta}^k_n(\beta,\lambda_n) - \hat{\theta}_n)$, where W_n is a sequence of positive-definite matrices,

$Q^{LR}_{nk}(\beta,\lambda_n) = -\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x_i; \bar{\theta}^k_n(\beta,\lambda_n))$,

$Q^{LM}_{nk}(\beta,\lambda_n) = (\frac{1}{H}\sum_{r=1}^{H} \frac{\partial}{\partial\beta}\mathcal{L}^{rk}_n(\beta,\lambda_n;\hat{\theta}_n)' \, V_n \, (\frac{1}{H}\sum_{r=1}^{H} \frac{\partial}{\partial\beta}\mathcal{L}^{rk}_n(\beta,\lambda_n;\hat{\theta}_n))$, where V_n is a sequence of positive-definite matrices,

for $k \in \mathbb{N}_0$.

"Newton-Raphson" estimator:

In practice, a large value of the total number of simulations H allows to select smaller value of the smoothing parameter λ (similar to have a sample size equal to nH).

But, as H becomes large, it is relatively costly to evaluate the objective function Q_{nk} (indeed, Q^{e}_{nk} for $e =$ W, LR or LM).

A **less costly** approach may consists to fix $H = 1$ and a large value of $\lambda = \lambda^{(0)}$. It is then possible to obtain a initial GII estimator, namely $\beta^{(0)}$.

Then we choose a large value of $H = H^{(1)}$ and a smaller value of $\lambda = \lambda^{(1)}$. Then we calculate a new value of the GII estimator using **one Newton-Raphson step**:

$$\hat{\beta}^{NR} = \hat{\beta}^{(0)} - [\,\frac{\partial^2}{\partial\beta\partial\beta'} Q_{nk}(\hat{\beta}^{(0)}, \lambda^{(1)}; H^{(1)})\,]^{-1}\,\frac{\partial}{\partial\beta} Q_{nk}(\hat{\beta}^{(0)}, \lambda^{(1)}; H^{(1)}),$$

where the objective function $Q_{nk}(\hat{\beta}^{(0)}, \lambda^{(1)}; H^{(1)})$ is indexed now by the number of draws ($H^{(1)}$) and the parameter ($\lambda^{(1)}$).

Remark : There is some evidence that this NR estimator ($k = 0$) **performs comparably** to the **jackknifed estimator** (Monte Carlo simulations).

It is then possible to iterate : let $\hat{\beta}_n(\lambda^{(i-1)})$ the GII estimator obtained selecting $\lambda = \lambda^{(i-1)}$.

We then minimize $Q_{nk}(\beta, \lambda^{(i)})$ with respect to β using $\hat{\beta}_n(\lambda^{(i-1)})$ as a starting value, where $\lambda^{(i)} = \rho\lambda^{(i-1)}$, for $\rho \in (0, 1)$.

For instance, one can stop once $\| \hat{\beta}_n(\lambda^{(i)}) - \hat{\beta}_n(\lambda^{(i-1)}) \| < 10^{-5}$.

Remark : you can increase H each time you reduce λ.

Notations:

Let us remind that $y^r(\beta, \lambda)$ is list of smoothed outcomes (one per period, per individual, per alternative) corresponding to the *rth* draw of the **initial model** such that the vector of parameters is fixed to β. λ is the smoothing parameter.

Let us remind that

$$\mathcal{L}_n(y, x; \theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i; \theta)$$

is the average log-likelihood corresponding to the **auxiliary model.** $\ell(y_i, x_i; \theta)$ is the contribution of individual to the log-likelihood (quasi-likelihood).

Let us define

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(y, x; \theta) \text{ and } \ell_i^r(\beta, \lambda; \theta) = \ell(y_i^r(\beta, \lambda), x_i; \theta).$$

Let

$$\phi_n^r = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ell_i^r(\beta_0, 0; \theta_0)$$

denote the **standardized score** vector for the *rth* draw calculated for $\lambda = 0$ (no smoothing).

The observed outcomes vector y in the sample can be denoted $y^0(\beta_0, 0)$ (i.e. the 0*th* simulation of the model for the true value of the vector of parameters without any smoothing).

Similarly, we can write $\hat{\theta}_n = \hat{\theta}_n^0(\beta_0, 0)$, the QMLE estimator of θ obtained using the 0*th* simulated sample (the observed one so), for the auxiliary model and without smoothing.

Let us assume that $W_n \xrightarrow{p} W$, where W is positive definite.

Let $D = E[\frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\theta_0)]$ denote the matrix of the expectation of the Hessian of $\mathcal{L}_n$ and $\theta_0 = \theta(\beta_0, 0)$ is often called the **pseudo true value** corresponding the auxiliary model.

Let $\Sigma{=}E[\phi_n^{r_1}\ \phi_n^{r_1\prime}]$ and $R{=}E[\phi_n^{r_1}\ \phi_n^{r_2\prime}]$, for every $r_1, r_2 \in \{0, 1, \ldots, H\}$ (two sample draws for the true value of the parameters and **without smoothing**).

We assume that $d_\theta \geq d_\beta$.

Let the matrix G denote the Jacobian of the **binding function** evaluated at $(\beta_0, 0)$

$$G = \frac{\partial}{\partial \beta'} \theta(\beta_0, 0),$$

is a $d_\theta \times d_\beta$ matrix. Let us remark that Σ, D, R and W are $d_\theta \times d_\theta$ matrices.

Let $\hat{\beta}^e_{nk}$ denote a minimizer of the objective function $Q^e_{nk}(\beta, \lambda_n)$, where $e = W, LR, LM$.

The **asymptotic variance** of the estimators has the usual sandwich expression

$$\Omega(U, V) = (G'UG)^{-1} G'UD^{-1} VD^{-1} UG(G'UG)^{-1}$$

where U and V are symmetric matrices (see below).
Let us remark that if $U = W^* = DV^{-1}D$ then $\Omega(U, V) = (G'DV^{-1}DG)^{-1}$ (this suggest a weighting matrix for $\hat{\beta}^W_{nk}$).

Theorem (Bruins et al. (2018))

Let us assume some regularity conditions hold then

$$\sqrt{n}(\hat{\beta}^e_{nk} - \beta_0) \xrightarrow{D} N(0, \Omega(U_e, V))$$

where

$$U_e = \begin{cases} W, & \text{if } e = W, \\ D, & \text{if } e = LR, \end{cases}$$

and $V = (1 + \frac{1}{H})(\Sigma - R)$.

Remarks :

For the estimator $\hat{\beta}^W_{nk}$, $\Omega(W^*, V) = (1+\frac{1}{H})(G'D(\Sigma-R)^{-1}DG)^{-1}$.

The regularity conditions include restrictions on the sequence λ_n.

The property holds singularly if $\lambda_n = 0$ (no smoothing).

Notations:

Let us define

$$A' = [\, I_{d_\theta} \;\; -\frac{1}{H} I_{d_\theta} \ldots - \frac{1}{H} I_{d_\theta}],$$

where I_{d_θ} is a $d_\theta \times d_\theta$ matrix and A' is $d_\theta \times ((H+1)\, d_\theta)$.

Let us denote

$$s'_{ni} = \left[\frac{\partial}{\partial \theta} \ell_i^0(\hat{\theta}_n)' \;\; \frac{\partial}{\partial \theta} \ell_i^1(\hat{\beta}^e_{nk}, \lambda_n; \hat{\theta}^1_n)' \;\ldots\; \frac{\partial}{\partial \theta} \ell_i^H(\hat{\beta}^e_{nk}, \lambda_n; \hat{\theta}^H_n)' \right],$$

where $\hat{\theta}^r_n = \hat{\theta}^r_n(\hat{\beta}^e_{nk}, \lambda_n)$ $(r = 1, \ldots, H)$ and $\ell_i^0(\theta) = \ell(y_i, x_i; \theta)$. s'_{ni} is a $1 \times ((H+1)\, d_\theta)$ matrix.

Theorem (Bruins et al. (2018) Variance estimation)

Let us assume some regularity conditions hold then

(i) $\hat{D}_n = \frac{\partial^2}{\partial\theta\partial\theta'}\mathcal{L}_n(\hat{\theta}_n) \xrightarrow{p} D,$

(ii) $\hat{V}_n = A' \ \left(\frac{1}{n}\sum_{i=1}^{n} s_{ni}\, s_{ni}'\right) \ A \xrightarrow{p} V,$

(iii) $\hat{G}_n = \frac{\partial}{\partial\beta}\bar{\theta}_n(\hat{\beta}_{nk}^{e}, \lambda_n) \xrightarrow{p} G, \text{ for } e = W, LR.$

Dynamic models : smoothing procedure

In this case, the lagged value(s) of the dependent variable can enter in the expression of the latent variables (or utilities).

Example: Let us assume that we have two alternatives to consider ($J = 2$) and $u_{it} = x_{it}\beta_1 + y_{i,t-1}\beta_2 + \epsilon_{it}$, where $\epsilon_{it} = \rho\epsilon_{i,t-1} + \eta_{it}$, $\epsilon_{i0} = 0$ and $\eta_{it} \sim N(0,1)$. $y_{it} = \mathbb{1}[u_{it} \geq 0]$. This is a **dynamic Probit model.**

Let us remark that

$$y_{it}(\beta) = \mathbb{1}[v^r_{it0}(\beta)\geq 0]\ (1-y_{i,t-1}(\beta)) + \mathbb{1}[v^r_{it1}(\beta)\geq 0]\ y_{i,t-1}(\beta),$$

where $v^r_{itk}(\beta) = x_{it}\beta_1 + \mathbb{1}[k=1]\beta_2 + \epsilon^r_{it}$.

The smoothed value of the dependent should be constructed using the expression

$$y^r_{it}(\beta,\lambda) = K_\lambda(v^r_{it0}(\beta))\ (1-y^r_{i,t-1}(\beta,\lambda)) + K_\lambda(v^r_{it1}(\beta))\ y^r_{i,t-1}(\beta,\lambda),$$

and $y^r_{i0}(\beta,\lambda) = 0$. We obtain an approximation of $y_{it}(\beta)$.

This part is dedicated to the presentation of the approach proposed by Keane and Sauer (Econometrica, 2009), that is a simulation based estimation algorithm presented in Keane and Sauer (IER, 2010). See, Keane and Sauer, 2006, for a synthesis.

The interest of the approach is to take into account **classification error** associated to the dependent variable. An application is made to study **female labor supply behavior**.

The methodology is used by Kean and Sauer (2006) in order to asses the exogeneity of fertility.

We are going to consider a **dynamic panel probit model**.

The estimation procedure was developed by Keane and Wolpin (2001) and Keane and Sauer (2006).

The model:

The dependent variable (a binary variable) is

$$y_{it} = 1\!\!1[\, x_{it}'\beta + \gamma\, y_{i,t-1} + u_{it} > 0\,], \tag{1}$$

where $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

In the application considered by Keane and Saurer, 2006, y_{it} represents **labor market participation** choice of women i at time t.

x_{it} is a vector of covariates (non labor income, number of children, age, education).

In the **random effect** specification (RE) of the model we consider that

$$u_{it} = \alpha_i + \epsilon_{it},$$

where α_i is an individual effet, $\alpha_i \sim N(0, \sigma_\alpha^2)$, and it is assumed to be conditionally independent of x_{it}.

In the correlated effect specification (CRE) of the model we consider that

$$\alpha_i = \sum_{t=0}^{T} z_{it}'\delta_t + \eta_i,$$

where $\eta_i \sim N(0, \sigma_\eta^2)$ and it is assumed to be conditionally independent of z_{it} and x_{it}.

ϵ_{it} follow an AR(1) process:

$$\epsilon_{it} = \rho\,\epsilon_{i,t-1} + v_{it},$$

where $v_{it} \sim N(0, \sigma_v^2)$ and it is conditionally independent of $\epsilon_{i,t-1}$. The process is assumed to be stationary so that $\sigma_\epsilon^2 = \frac{\sigma_v^2}{1-\rho^2}$.

Let us assume that $\sigma_u^2 = \sigma_\eta^2 + \sigma_\epsilon^2 = 1$ for identification (we observe the realization of a binary variable). Then, $\sigma_u^2 = \sigma_\eta^2 + \frac{\sigma_v^2}{1-\rho^2} = 1$. Consequently $\sigma_v^2 = (1-\rho^2)(1-\sigma_\eta^2)$ (σ_u^2 and σ_v^2 can be deduced from ρ and σ_η^2).

Initial conditions: (Heckman, 1981)

$$y_{i0} = \mathbb{1}[\, x_{i0}'\beta_0 + u_{i0} > 0\,]$$

$$\rho_t = corr(u_{i0}, u_{it}),$$

for $t \geq 1$. $t = 0$ is the time of the first observation of the data (it is different from the start time of the process).

Let us assume $u_{i0} \sim N(0, 1)$. We can assume that $\rho_t = \rho_0$ for all $t \geq 1$.

Remark: Keane and Sauer (2010) propose an alternative way to take into account the initial conditions problem. They simulate histories from the theoretical start of the process ($\tau < 0$). If the variables x_{it} are also missing for $t < 0$ these explanatory variables must also be simulated.

Let $\bar{u}_i = (u_{i1}, \cdots, u_{iT})'$ denote the vector of error terms for the periods distinct from the initial time. We obtain that

$$\bar{u}_i \ \sim \ \mathrm{N}(0, \Omega_{\bar{u}}),$$

where

$$\Omega_{\bar{u}} = var(\bar{u}_i | x_i, z_i) = \sigma_\eta^2 \ \mathit{1}_T \ \mathit{1}_T' + \frac{\sigma_v^2}{(\mathit{1} - \rho^2)} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{T-2} & \rho^{T-3} & \dots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{bmatrix}.$$

It is necessary to assume that $0 < \sigma_\eta^2 < 1$ (because $\sigma_\eta^2 + \sigma_\epsilon^2 = 1$).

Let us consider the vector $u_i = (u_{i0}, \ldots, u_{iT})'$. Then we have $u_i \sim N(0, \Omega)$, with

$$\Omega = \begin{bmatrix} 1 & \rho_0 & \cdots & \rho_0 \\ \rho_0 & & & \\ \vdots & & \Omega_{\bar{u}} & \\ \rho_0 & & & \end{bmatrix},$$

where ρ_0 represents the correlation between the error term of the initial time and the error terms of the next periods. This correlation is identifiable.

Classification error:

Let y^*_{it} denote the **reported choice** by the individual. Let us define four probabilities

Let us define four probabilities

$$\begin{aligned}
\pi_{11t} = &\quad \text{Prob}[y_{it}^* = 1 \mid y_{it} = 1], \\
\pi_{01t} = &\quad \text{Prob}[y_{it}^* = 1 \mid y_{it} = 0], \\
\pi_{10t} = &\quad 1 - \pi_{11t}, \\
\pi_{00} = &\quad 1 - \pi_{01t} \qquad\qquad\qquad ,
\end{aligned}$$

For instance, π_{01t} is the conditional probability that the choice 1 is reported ($y_{it}^* = 1$) given that the true choice is 0 ($y_{it} = 0$).

We are going to consider a simple specification of the classification error process. We are going to assume that classification error is biaised (Keane and Sauer, 2010, consider also an unbiased classification error).

A biased classification error means that the probability an individual is observed to choose an alternative is not equal to the true probability that this person chooses this alternative, i.e. $\text{Prob}[y_{it}^* = 1] \neq \text{Prob}[y_{it} = 1]$.

The biased process y_{it}^* is defined by the **latent variable**

$$l_{it} = \gamma_0 + \gamma_1\, y_{it} + \gamma_2\, y_{i,t-1}^* + \omega_{it},$$

where y_{it}^* is the **reported choice** and ω_{it} is an error term.

$$y_{it}^* = \begin{cases} 1 & \text{if } l_{it} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let us remark that the greater is γ_2 and the greater is the **persistence** in classification error (dynamics).

Let us assume ω_{it} is **logistic**. Consequently, we have

$$\pi_{11t} = \text{Prob}[y_{it}^* = 1 | y_{it} = 1] = \frac{\exp(\gamma_0 + \gamma_1 + \gamma_2\, y_{i,t-1}^*)}{1 + \exp(\gamma_0 + \gamma_1 + \gamma_2\, y_{i,t-1}^*)} \quad (2)$$

$$\pi_{01t} = \text{Prob}[y_{it}^* = 1 | y_{it} = 0] = \frac{\exp(\gamma_0 + \gamma_2\, y_{i,t-1}^*)}{1 + \exp(\gamma_0 + \gamma_2\, y_{i,t-1}^*)}$$

Let $Y_i^* = \{y_{it}^*\}_{t=0}^T$ denote the **history of reported choices** for the person i, where $i = 1, \ldots, n$ (n is the sample size). Let $x_i = \{x_{it}\}_{t=0}^T$ denote the history of the exogenous variables for the individual i.

Keane and Sauer's algorithm (SMLE):

1) For each individual i we generate H sequences of errors terms $u_i = (u_{i0}, \ldots, u_{iT})'$ from the joint distribution $\phi(u \mid \Omega)$ (the pdf of the normal distribution with mean zero and varcov Ω). We obtain the sequences $u_i^r = (u_{i0}^r, \ldots, u_{iT}^r)'$, for $r = 1, \ldots, H$.

2) Given the vector of parameters θ, using $\{u_{it}^r\}_{t=0}^T$ and x_i, for $r = 1, \ldots, H$ and $i = 1, \ldots, n$, we can construct H simulated choice histories $\{y_{it}^r\}_{t=0}^T$ for each individual given the decision model (1).

3) Construct the probabilities $\{\hat{\pi}^r_{jkt}\}_{t=0}^T$, for each draw r and each individual i, where j is the **simulated choice** and k is the reported choice. The probabilities are given in (2). We assume **there is no missing observation** (otherwise the y^*_{it} should be simulated).

4) Then we can calculate the value of the simulator of the **contribution** of a given individual to the likelihood function:

$$P\hat{r}ob[Y^*_i \mid x_i; \theta] = \frac{1}{H} \sum_{r=1}^{H} \prod_{t=0}^{T} \left(\sum_{j=0}^{1} \sum_{k=0}^{1} \hat{\pi}^r_{jkt} \, 1\!\!1[y^r_{it} = j; y^*_{it} = k] \right) \tag{3}$$

The SML estimator is **consistent and asymptotically normal** if $\frac{H}{\sqrt{n}} \longrightarrow +\infty$ as $n \longrightarrow +\infty$.

Importance Sampling:

The contribution to the simulated likelihood function (3) is not smooth with respect to the parameters. For a given draw of the error term $\{u_{it}^r\}_{t=0}^T$ a modification of the value of the parameters θ can result in discrete changes of the sequence $\{y_{it}^r\}_{t=0}^T$.

In order to smooth the simulated likelihood function with respect to θ we start by constructing simulated histories $\{y_{it}^r\}_{t=0}^T$ for a starting value θ_1 of the parameters. Let $\{y_{it}^r(\theta_1)\}_{t=0}^T$ denote these **simulated histories.**

These histories then are not modified as we make θ vary.

We then construct a weight associated to the parameters θ and the sequence r

$$W_{ir}(\theta) = \frac{\text{Prob}[y_{i0}^r(\theta_1), \ldots, y_{iT}^r(\theta_1) \mid x_i, \theta]}{\text{Prob}[y_{i0}^r(\theta_1), \ldots, y_{iT}^r(\theta_1) \mid x_i, \theta_1]}.$$

The numerator is the conditional probability of the simulated sequence $\{y^r_{it}(\theta_1)\}_{t=0}^T$ for a value of the parameter θ and the denominator is the conditional probability of the same simulated history calculated for the value θ_1 of the vector of parmeters.

This **conditional probability** at numerator of the weight $W_{ir}(\theta)$ is

$$\int_{D(\{y^r_{it}(\theta_1)\}_{t=0}^T,\{x_{it}\}_{t=0}^T;\theta)} \phi(u_{i0},\ldots,u_{iT} \mid \Omega(\theta))\, d\, u_{i0} \ldots d\, u_{iT}$$

where $\phi(. \mid \Omega)$ is the joint density of the normal distribution with mean 0 and var-cov Ω. $\Omega(\theta)$ is the value of Ω when the elements of the matrix are constructed using the vector θ of parameters.

Let us assume that $D(\{y^r_{it}(\theta_1)\}_{t=0}^T, \{x_{it}\}_{t=0}^T; \theta)=$

$$\{u_i \in \boldsymbol{R}^{T+1} : \mathbb{1}[u_{it} > -x'_{it}\beta - \gamma y^r_{i,t-1}(\theta_1) - z'_i\delta] = y^r_{it}(\theta_1), \forall t \geq 1 \text{ and } \mathbb{1}[u_{i0} > -x'_{i0}\beta_0] = y^r_{i0}(\theta_1)\}$$

The weight can be difficult to evaluate. There exists an **alternative way to smooth the likelihood.** We construct the simulated choices $\{y_{it}^r(\theta_1)\}_{t=1}^T$ for an initial value of the vector of parameters θ_1.

We then save the **latent variables**, namely $\{V_{it}^r(\theta_1)\}_{t=0}^T$, that corresponds to the simulations, where

$$V_{it}^r(\theta_1) = x_{it}'\beta + \gamma y_{i,t-1}^r(\theta_1) + z_i'\delta + u_{it}^r,$$

where u_{it}^r is a draw of the residual $\eta_i + \epsilon_{it}$, for $t \geq 1$ and

$$V_{i0}^r(\theta_1) = x_{i0}'\beta_0 + u_{i0}^r,$$

$(\beta_0, \beta, \gamma, \delta)$ **are elements** of the vector θ_1.

Then we keep the histories $\{V_{it}^r(\theta_1)\}_{t=0}^T$ and $\{y_{it}^r(\theta_1)\}_{t=0}^T$ fixed when θ varies.

Let us remark that: $y_{it}^r(\theta_1) = \mathbb{1}[V_{it}^r(\theta_1) > 0]$, for $t \geq 0$.

For this alternative way to smooth the contributions, the weight as the expression

$$W_{ir}(\theta) = \frac{g(V^r_{i0}(\theta_1), \ldots, V^r_{iT}(\theta_1) \mid x_i, z_i; \theta)}{g(V^r_{i0}(\theta_1), \ldots, V^r_{iT}(\theta_1) \mid x_i, z_i; \theta_1)} \tag{4}$$

where $g(.)$ is the joint density of the simulated history $(V^r_{i0}(\theta_1), \ldots, V^r_{iT}(\theta_1))'$:

$$g(V^r_{i0}(\theta_1), \ldots, V^r_{iT}(\theta_1) \mid x_i, z_i; \theta)$$

$$= \phi(V^r_{i0}(\theta_1) - x'_{i0}\beta_0,$$

$$V^r_{i1}(\theta_1) - x'_{i1}\beta - \gamma y^r_{i,0}(\theta_1) - z'_i\delta, \ldots, V^r_{iT}(\theta_1) - x'_{iT}\beta - \gamma y^r_{i,T-1}(\theta_1) - z'_i\delta | \Omega(\theta))$$

where $\phi(. \mid \Omega(\theta))$ is the pdf of the multivariate normal distribution with mean 0 and var-cov $\Omega(\theta)$ (i.e. the matrix Ω calculated **using** θ), $(\beta_0, \beta, \gamma, \delta)$ are elements of the vector θ.

The smooth simulated contribution to the likelihood function of individual i is

$$\hat{Prob}[Y_i^* \mid x_i; \theta] = \frac{1}{H}\sum_{r=1}^{H} W_{ir}(\theta) \prod_{t=0}^{T} \left(\sum_{j=0}^{1} \sum_{k=0}^{1} \hat{\pi}_{jkt}^{r} \, 1\!\!1[y_{it}^{r}(\theta_1) = j; y_{it}^{*} = k] \right) \tag{5}$$

where the weight are given by (4).

Remark: An advantage of this version of the contribution is that the **choice histories have to be generated only one time** per individual for $\theta = \theta_1$.

An, M.Y., Liu, M., 2000, Using indirect inference to solve the initial-conditions problems, Revue of Economics and Statistics, 82(4), 656-667.

Bruins, M., Duffy, J.A., Keane M., Smith, A.A. Jr., 2018, Generalized indirect inference for discrete choice models, Journal of Econometrics, 205, 177-203.

Gouriéroux, Ch., Monfort, A., Renault, E., 1993, Indirect inference, Journal of Applied Econometrics, 8(S1), S85-S118.

Heckman, 1981, The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process, in C. Manski and D. McFadden, eds., The StructuralAnalysis of Discrete Data. Cambridge: MIT Press.

Keane, M., Sauer, R.M., 2006, Classification Error in Dynamic Discrete Choice Models: Implications for Female Labor Supply Behavior. IZA Working Paper 2332.

Keane, M., Sauer, R.M., 2009, Classification error in dynamic discrete choice models: Implications for female labor supply behavior. Econometrica, 77, 975-991.

Keane, M., Sauer, R.M., 2010, A computationally practical simulation estimation algorithm for dynamic panel data models with endogenous state variables. International Economic Review, 51(4), 925-958.

Keane, M., Wolpin, K., 2001, The effects of parental transfers and borrowing contraints on educational attainment, International Economic Review, 42, 1051-1103.

Quenouille, M., 1956, Notes on bias estimation, Biometrika, 43, 353-360.

Sidi, A., 2003, Practical extrapolation methods : Theory and applications. Cambridge University Press, Cambridge (UK).

Wooldridge, 2005, Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. Journal of Applied Econometrics

Jackknife and bias reduction (a short introduction)

Let us consider an i.i.d. sample $\{y_1, \ldots, y_n\}$, where n is the size of the sample. Let us consider a parameter θ of the distribution of Y_i.

Let $\hat{\theta}$ denote a **biased estimator** of θ. The objective is to reduce the bias of this estimator.

Let us assume that the bias of the estimator has a power series expansion in n^{-1}

$$E(\hat{\theta} - \theta_0) = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \ldots$$

Let $\hat{\theta}_{(i)}$ denote the estimator $\hat{\theta}$ when the observation y_i is dropped from the sample.

The expression of the **jackknife estimator** is

$$J(\hat{\theta}) = n\,\hat{\theta} - (n{-}1)\,\hat{\theta}_{.}$$

where $\hat{\theta}_{.} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$.

Quenouille (1956) has proved that the jackknife estimator has eliminated the 1/n term in the power series expansion

$$E(J(\hat{\theta}) - \theta_0) = -\frac{a_2}{n(n-1)} - \frac{a_3(2n-1)}{n^2(n-1)^2} - \dots$$

$\longrightarrow$ we obtain a bias reduction.

Remark: We can construct an example such that, for instance, $\theta_0 > 0$ and $\hat{\theta} > 0$, but $J(\hat{\theta})$ can be negative.