Discrete Choice

J.R. Walker

E718: Topics in Applied Econometrics

Classical consumer theory assumes that products are infinitely divisible and consumers may purchase fractional quantities as dictated by their preferences and budget sets. However, for many goods and services individuals purchase only a few and sometimes at most one. Thus classical models of discrete choice in economics were developed to extend consumer choice theory to model demand of discrete goods.

One particular applied problem which received attention of McFadden et al. was to forecast the demand for a new good, add a good to the choice set. That is, live in regime with preferences defined over goods \((X_1, \ldots, X_N)\) and now add a new good \(\tilde{U}(X_1, \ldots, X_N, X_{N+1})\). The problem is that we have no data in which includes the demand for \(X_{N+1}\). This is a situation that requires out-of-sample forecasting and according to Marschak, structural modeling. We specify preferences, opportunity sets (the primitives). Estimate preference parameters and then add new product to the opportunity set. McFadden (1974) is a first attempt at this problem.

The key to the solution, of course, is that we must be able to relate the old goods and new goods. One approach is the Lancaster (Becker) approach in which consumers have preferences defined over the characteristics of goods, not the goods themselves. In Lancaster \(X = AZ\) where \(X\) are goods and \(Z\) are attributes of the goods. Preferences are defined \(U^*(Z)\). In this set up, to merge the new good with the old requires that we specify the attributes of the new good, i.e., goods are packages of attributes. This leads to a hedonic pricing equation for attributes. A common application is consumer demand for durable goods such as automobiles and houses. When goods are packaged with fixed attributes and it is not possible to purchase multiple combinations of goods then the pricing function of attributes will be nonlinear (i.e., no arbitrage of characteristics).

We consideration to alternative approaches to specifying discrete choice systems with direct linkages to economic theory.

**Random parameter models**

Since these models were developed within consumer theory that is the natural framework in which to present them.

**Notation**

Index alternatives by \(j\), and the choice set has \(m\) alternatives. An alternative is characterized completely by a vector \(x^j\) of dimension \(q\). The characteristics may be quantities of goods, quantities
of characteristics, dummy variables representing characteristics in qualitative form, or if preferences represented by an indirect utility function, $x_j$ may include prices divided by the individual’s income. Some characteristics may depend on the alternative (and hence the same across individuals). Others may vary across people, e.g., could represent a shadow price of a non-market good.

Preferences are represented by a utility function $u = V(x; \beta)$ where $\beta$ is a vector of unobserved preference shift variables. $\beta$ will depend in some way on the observed personal characteristics of $n$, $z_n$ of the individual. The content of the model is that that the function $V$ is common to all individuals and all alternatives, and displays generic properties of a utility function for admissible values of $\beta$.

A difference between modeling strategies described below is whether stochastic elements enter into the description of alternatives or in the comparison of the $m$ utilities. The stochastic elements are needed to explain why two observational equivalent individuals make different choices. Thus, choices will be presented by conditional choice probabilities. Let $\zeta = (x_1^1, \ldots, x_m^m | z)$ represent the opportunity set. Then the choice probability conditional on opportunity set is:

$$\Pr(\text{Select } j | \zeta) = P(j | \zeta, \theta)$$

**Random Parameter Models (Quandt)**

Assume that preferences can be universally represented by a utility function of the form $u = V(x; \beta)$, where $\beta$ is a vector of substitution parameters or unobserved shift variables that vary randomly between individuals. Thus, a typical individual $n$ will derive from alternative $j$ utility:

$$u_j^i = V(x_j^i; \beta_n)$$

$x_j^i$ is a vector of characteristics of alternative $j$ for individual $n$. Included in the vector can be the market price of alternative $j$ or for non market goods, $x_j^i$ could include the shadow price of alternative $j$ for person $n$. As we will see, characteristics can depend only on alternative $j$ (and not vary by $n$) but alternatives with whose characteristics depend only on $n$ are not viable (or not identified). The vector $\beta_n$ will be random with a joint c.d.f. $G_\beta(\beta | z_n; \theta)$ whose form depends on the characteristics of the individual, $z_n$, and on the model’s parameters $\theta$. We have considerable latitude in our modeling as we put parameters $\theta$ in $U$ or in $G$. We will assume that $V \perp G$.

Interpret the model that the functional form of preferences is stable and the same across people (e.g., CES, Cobb-Douglas). However, $\beta$ indexes heterogeneity in preferences across the population. Hence, $G$ and $\theta$ describe consumer heterogeneity, as each person preferences represent a draw from $G_\beta$.

The econometrics problem is to estimate $\theta$, which of course requires the choice probabilities, $\Pr(j | \zeta_n; \theta)$. Rationality implies the individual select the alternative that yields the maximum utility payoff $u_n^* = \max u_n^1, \ldots, u_n^m$. The ordering of utilities depends on the vector $\beta_n$, so the probability of selection for any particular alternative can be regarded as the probability that $\beta_n$ falls within the region of $\beta$–space that corresponds to the optimality of that alternative. The choice sets, $B_n^j$, are defined as:

$$B_n^j = \{\beta : V(x_j^i; \beta) > V(x_j^j; \beta), \forall i \neq j\}$$
When $\beta_n \in B^j_n$ alternative $j$ is selected with probability equal to:

$$P(j \mid \xi_n; \theta) = \int_{\beta^j_n} dG_{\beta}(\beta; z_n, \theta).$$

Define a set of indicator variables $d^j_n = 1$ if alternative $j$ is selected and zero otherwise, and assuming the alternatives are mutually exclusive and collectively exhaustive $\sum_j d^j_n = 1$. Then the contribution of individual $n$ to the log–likelihood is

$$\ell_n(\theta; d_n, \xi_n) = \sum_{k=1}^m d^k_n \ln \left[ Pr(k \mid \xi_n, \theta) \right].$$

This representation has been little used in applications because of the significant computation demands. Advances in computers and simulation estimators (particularly MCMC) have reduced the computational burden significantly, but the computational burden remains high. The computational burden comes from two sources:

1. The computational burden increases exponentially in $q$ the dimension of the characteristics vector. Numerical algorithms (e.g., Clark’s method for the multinomial normal) works for four, at most five, choices. Simulation methods whose computational burden is independent (essentially) of the number of choices may offer a solution.

2. Irregular and complicated choice sets. The limits of integration need not be connected or otherwise well–behaved. Again, simulation may offer a solution.

**Alternative Specific Random Errors**

The more common specification is:

$$u^j_n = V^*(x^j_n, z_n; \alpha) + \epsilon^j_n, \quad j = 1, \ldots, m$$  \hspace{1cm} (2)

where $\alpha$ is a vector of parameters and a subvector of $\theta$ (the full set of parameters). Note, however, $V^*$ is not a utility function in the usual sense. It is true that $V^* + \epsilon^j_n$ is invariant to monotonic transformations, but behavior is not invariant to transformations of $V^*$ alone. It is important to realize that $V^*$ is a *cardinal* (as opposed to an *ordinal*) representation. A change such as $\tilde{V} = \log(V^*)$ is a different model.

According to consumer theory, utility functions should be invariant to monotonic transformations of $V^*$.

However equation (2) could arise from the random utility framework (1) if

$$V^*(x^j_n, z_n; \alpha) = E \left[ V(x^j_n; \beta_n) \mid \xi_n \right]$$

$$\epsilon^j_n = V(x^j_n; \beta_n) - V^*(x^j_n, z_n; \alpha)$$

Yet, a more common interpretation is that the error in equation (2) represents unobserved (random) attributes of alternative $j$. Write the vectors of observable and unobservable characteristics as $x^j_n$ and $\nu^j_n$, and represent preferences by a utility function

$$u^j_n = V(x^j_n, \nu^j_n; \beta_n)$$
Where $\beta_n$ depends the characteristics $z_n$ in some way and may be stochastic. Let $\omega^j_n = (\nu^j_n, \beta_n)$ and write its conditional cdf as $G^j_\omega(\omega^j \mid \zeta_n)$. Then expected utility of alternative $j$ for person $n$ is

$$E(u^j_n \mid \zeta_n) = \int V(x^j_n, \nu^j_n, \beta_n) dG^j_\omega(\omega^j \mid \zeta_n) = V^*_j(x^j_n, z_n; \alpha)$$

This implies we need to define $\epsilon^j_n$ as

$$\epsilon^j_n = u^j_n - E[u^j_n \mid \zeta_n]$$

(3)

$$\epsilon^j_n = u^j_n - V^*_j(x^j_n, z_n; \alpha)$$

(4)

So that

$$u^j_n = V^*_j(x^j_n, z_n; \alpha) + \epsilon^j_n$$

(5)

This yields equation (5) with the form of equation (2) but with first term now deterministic and dependent on $j$. Only in the special cases will the $V^*_j$ be identical for all $j$, for example if all the $\omega^1, \ldots, \omega^m$ share a common distribution conditional on $\zeta_n$. This is implausible in most cases.\(^1\)

And one has to accept that the non–stochastic component of the utility function may not be invariant.

In most applications the alternative specific choice functions are additive:

$$V^*_j = V^*_j(x^j_n, z_n; \alpha^0) + \alpha^j$$

(6)

A normalization is needed and frequently it’s assumed that $\alpha^1 = 0$, with $\alpha^2, \ldots, \alpha^m$ random, and $\alpha^0$ are parameters of the deterministic choice component to be estimated (a vector).

The chief disadvantage of equations (2,5,) is that the additive structure combined with independence of the $\epsilon^j$ across alternatives has the undesirable consequence of restricting the utility trade–off among alternatives. And more general assumptions bring additional covariance parameters that are specific to each alternative (always true for models of equations (??, ??)). (This makes use of these forms to predict the impact of a new product difficult. [adding unknown parameters].)

Alternative specific error component models based on equations (5, ) have choice probabilities with a common representation in terms of the joint distribution of $\epsilon^1, \ldots, \epsilon^m$. Alternative $j$ is chosen if

$$u^j_n - u^i_n > 0 \forall i \neq j$$

This implies

$$\epsilon^i_n < \epsilon^j_n + V^*_j(x^j_n, z_n; \alpha) - V^*_i(x^i_n, z_n; \alpha)$$

Defining $\psi^i_n(\alpha)$ as the difference of the non–stochastic utilities yields

$$\epsilon^i_n < \epsilon^j_n + \psi^i_n$$

Assuming that $\epsilon^1, \ldots, \epsilon^m$ have a continuous distribution with joint pdf equal to $g(\cdot)$ the conditional probability that alternative $j$ is chosen is:

\(^1\)Pudney notes it is hard to believe that individual preferences of the level of comfort of cars and buses fluctuate randomly about a common mean.
If the $e_1, \ldots, e_m$ have a MVN(0, $\Sigma$) distribution working with disturbance differences is more informative. Define

$$\eta_{ij} = e_i - e_j, \quad i = 1, \ldots, m; \quad i \neq j$$

the probability of selecting alternative is

$$P(j \mid \zeta_n, \theta) = \Pr \left( \eta_{ij} = \psi_n^{ij}(\alpha), \text{ all } i \neq j \mid \zeta_n \right)$$

$$= G_n(\psi_n^{ij}, \ldots, \psi_n^{i-1j}, \psi_n^{ij+1}, \ldots, \psi_n^{mj})$$

This is an $m - 1$ dimension integral. This is an important reduction in computation but the task of evaluating these integrals can be formidable unless the number of alternatives is small. For this reason, it is common to work with distributions for which $G_e$ or $G_\eta$ have closed form solutions and do not require numerical integrations. However, we consider next the most computationally demanding set of choice probabilities. Importantly, it is also the most flexible for specifying substitution possibilities.

**Multinomial Probit**

Let $e_n = (e_n^1, \ldots, e_n^m) \sim MVN(0, \Sigma)$. The variance–covariance matrix, $\Sigma$, is $m \times m$ positive definite with elements $\{\sigma_{ij}\}$. There is a scale normalization required, e.g., $\sigma_{11} = 1$.

Since the elements of $e_n$ are normal, differences of the elements are also normal random variables. Hence, $\eta_{ij}^n$ is normal with mean 0, and variance equal to $\sigma_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$. Then the probability of alternative $j$ is

$$P(j \mid \zeta_n, \theta) = \Phi(\psi_n^{ij}(\alpha); 0, \Omega^j)$$

The $\psi_n^{ij}(\alpha)$ is an $m - 1$ vector containing $\psi_n^{ij}(\alpha)$.

The advantage of the MVN is its flexible (unrestricted) substitution pattern. The price of this flexibility is the computational burden of evaluating the $m - 1$ dimensional integrals.

Setting $m = 2$ yields the binomial probit:

$$P(j = 2 \mid \zeta_n, \theta) = \Phi \left( \frac{V_2^*(\chi^2_n, \zeta_n, \alpha) - V_1^*(\chi^1_n, \zeta_n, \theta)}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}} \right)$$

with the scale normalization is $\sigma_{11} + \sigma_{22} - 2\sigma_{12} = 1$. If the choice specific utility functions are linear then

$$V_2^*(\chi^2_n, \zeta_n, \theta) - V_1^*(\chi^1_n, \zeta_n, \theta) = \gamma' \zeta_n$$

So that

$$P(1 \mid \zeta_n, \theta) = 1 - \Phi(\gamma' \zeta_n)$$

$$P(2 \mid \zeta_n, \theta) = \Phi(\gamma' \zeta_n)$$
Computation of MNP

Let me follow the notation of Geweke and Keane (chapter 56, Vol 5 Handbook of Econometrics) and denote the discrete choice model as

\[ y_{ij}^* = Z_{ij}\gamma + X_i\beta_j + \epsilon_{ij}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, N. \]

Notice that to retain individual-specific characteristics in the model it is necessary to include choice specific coefficients \( \beta_j \). The idea is that \( X_i \) are taste shifters (e.g., if \( X_i \) is age, then \( \beta_j \) captures differences in preferences by age). \( y_{ij}^* \) are latent random variables.

Define alternative \( m \) as the base alternative and define

\[
y_{ij} = y_{ij}^* - y_{im} = (Z_{ij} - Z_{im})\gamma + X_i(\beta_j^* - \beta_m^*) + (\epsilon_{ij}^* - \epsilon_{im}^*) = (Z_{ij} - Z_{im})\gamma + X_i\beta_j + \epsilon_{ij}, \quad j = 1, \ldots, m - 1
\]

\[ y_{im} = 0 \]

where \( \epsilon_i \equiv (\epsilon_{i1}, \ldots, \epsilon_{im-1}) \sim N(0, \Sigma) \) and \( \Sigma \) is a \((m - 1) \times (m - 1)\) covariance matrix obtained from \( \Sigma^* \). A further scale normalization is usually imposed by setting \( \Sigma_{11} = 1 \).

It is convenient to write \( y_{ij} = \bar{y}_{ij} + \epsilon_{ij} \) for \( j = 1, \ldots, m \) and adopt the convention that \( \bar{y}_{im} = \epsilon_{im} = 0 \). Then the agent chooses option \( j \) if and only if \( y_{ik} - y_{ij} \leq 0 \ \forall k \), which generates the \( m - 1 \) dimension partition of the \( \epsilon \) space \( \epsilon_{ik} - \epsilon_{ij} \leq \bar{y}_{ij} - \bar{y}_{ik} \ \forall k \). Define \( \bar{\epsilon}_{ik} = \epsilon_{ik} - \epsilon_{ij} \) for \( k = 1, \ldots, m \) and further define \( \bar{\epsilon}_i = (\bar{\epsilon}_{i1}, \ldots, \bar{\epsilon}_{ij-1}, \bar{\epsilon}_{ij+1}, \ldots, \bar{\epsilon}_{im}) \sim N(0, \bar{\Sigma}) \).

Then the probability that agent \( i \) chooses option \( j \) can be written as the \( m - 1 \) dimension integral:

\[
p(j | Z_{ii}, X_i, \gamma, \beta, \Sigma) = \int_{-\infty}^{g_{ij} - g_{i1}} \cdots \int_{-\infty}^{g_{ij} - g_{i1}} p\left(\bar{\epsilon}_1, \ldots, \bar{\epsilon}_j | \bar{\Sigma}^j\right) d\bar{\epsilon}_j \cdots d\bar{\epsilon}_1
\]

\[ = P\left(\bar{y}_{ij} - \bar{y}_{i1}, \ldots, \bar{y}_{ij} - \bar{y}_{ij} | \bar{\Sigma}^j\right). \]

Letting \( d = (d_1, \ldots, d_m) \) the MNP likelihood function is:

\[
p(d | \gamma, \beta, \Sigma) = \prod_{i=1}^{N} \prod_{j=1}^{J} P\left(\bar{y}_{ij} - \bar{y}_{i1}, \ldots, \bar{y}_{ij} - \bar{y}_{im} | \bar{\Sigma}^j\right)^{1[d_j = j]}.
\]

The computationally demanding task of MNP estimation is the evaluation of the \( m - 1 \) dimension integrals that comprise the likelihood function.

Geweke and Keane state that deterministic methods (e.g., quadrature) are useful only for very low dimensions, \( J = 3, 4 \). Because the likelihood has to be evaluated at a large number of trial values
of the $K \times 1$ parameter vector $\theta = (\gamma, \beta, \Sigma)$. At each trial value $\theta_T$ the $J - 1$ dimensional integrals must be evaluated for all $N$ agents in the population. Moreover, if using a derivative based search algorithm to maximize the likelihood function, it is necessary to evaluate the likelihood at several different values of the parameter vector. These include the (1) initial value of $\theta_T$, and (2) $K$ incremented parameter values $\theta_T + h \Delta_k$, where $h$ is the increment size, and $\Delta_k = 1$ for $k = 1, \ldots, K$ and zero otherwise (used to evaluate the derivatives), and (3) the new trial parameter vector $\theta_T'$ of which a line search algorithm will always try at least two. Thus, each iteration involves at least $K + 3$ evaluations of the likelihood.

Geweke and Keane note that in a small problem with $N = 500$ and $J = 4$, with twelve parameters, and 5 free elements of $\Sigma$, optimization may require 50 iterations. Then approximately $50 \cdot (12 + 3) \cdot 500 = 375000$ three dimensional integrals must be evaluated. At 100 evaluations per second (which is probably not so fast today), would require more than an hour of computation.

It is this curse of dimensionality that stimulated researchers to consider simulation based procedures. Simulated Maximum Likelihood (SML) applies fast simulation methods to approximate $p(j|Z_i, X_i, \gamma, \beta, \Sigma)$ and to insert these approximations into the likelihood function.

Early attempts (e.g., Manski and Lerman in ?) used crude frequency simulators

$$\hat{p}(j|Z_i, X_i, \gamma, \beta, \Sigma) = M^{-1} \sum_{m=1}^{M} \begin{cases} 1 & \text{if } \tilde{\epsilon}_j^{(m)} \leq \tilde{y}_ij - \tilde{y}_i1, \ldots, \tilde{\epsilon}_j^{(m)} \leq \tilde{y}_ij - \tilde{y}_iJ, \\ 0 & \text{otherwise} \end{cases},$$

where $\left\{ \tilde{\epsilon}_k^{(m)} \right\}_{k=1}^{J}$ for $m = 1, \ldots, M$ are iid draws from the joint $\tilde{\epsilon}_j^{(m)} \sim N(0, \tilde{\Sigma}_j)$.

The frequency simulators are most intuitive and yet consensus is that frequency simulators did not perform satisfactory. These simulations require a huge number of replications ($M$) per case. Because with small $M$ low probabilities are simulated to be zero, and the objective function is not differentiable (or continuous) functions of $\theta$. (Frequency simulators renders the likelihood function a step function with jumps at parameter configurations that produce ties among alternatives.)

Moreover, the non-differentiable objective function requires the use of a non–gradient based optimizer (such as the Nelder–Mead simplex method). These optimizers are robust but S–L–O–W.

And from a theoretical econometric perspective, notice that the simulation error enters the log likelihood function nonlinearly (b/c of the log). The nonlinearity requires that both $N \to \infty$ and $M \to \infty$. But this means for fixed $N$ requires $M$ to be very large so the (simulated) to make the simulation error negligible. Frequency simulators are crude.

McFadden (1989) proposed the use of smoothed simulators. Smooth simulator can be constructed to solve the deficiencies of the frequency simulators noted above. Thus smooth simulators can be unbiased and the objective function is differentiable in the parameters (allowing for the use of (relatively fast) gradient based optimizers). Plus, smoother estimators are more accurate than frequency simulators for a fixed number of draws.
The consensus is that among smooth simulators the procedure proposed (independently) by Geweke, Hajivassiliou, and Keane (GHK) is the preferred and performs well in a broad class of problems. More on this later in the semester.

In the 1989 paper McFadden introduced the Method of Simulated Moments (MSM). It held great promise because of the linearity of the moment condition (and thus averaging [smoothing] across $N$) consistency required $N \to \infty$ for fixed $M$. To see this, write the simulated moment conditions as

$$\sum_{i=1}^{M} \sum_{j=1}^{I} W_{ij} [1[d_j = j] - \hat{p}(j|Z_i, X_i, \hat{\theta})] = 0,$$

where $W_{ij}$ is a set of instruments and $\hat{p}$ is any unbiased simulator (satisfying certain regularity conditions). The simulation errors need to be asymptotically uncorrelated with the instruments, however.

The last twenty or so years of research has not born this out because the finite sample properties deteriorate rapidly when imprecise simulators are used (such as the frequency simulator). Still need to introduce another form of smoothing beyond sample averaging.

Geweke and Keane note: one exception MSM preferred to SML is in panel data with correlated errors.

Geweke and Keane’s summary:

1. Except as noted above, the choice of between MSM or SML as the estimation method is not important.
2. Essential to use a highly accurate smooth probability simulator like GHK.
3. Care must be taken to use sufficiently large simulation sizes. For small choice sets ($J = 3$ or 4, GHK with only 10 to 30 draws appears to work, while much larger simulation sizes needed for more complex models.
4. In MNP models alternative specific covariates (i.e., exclusion restrictions) are critical if covariance parameters are to be identified. Even so, covariance parameters are poorly identified in models with large choice sets. And the number of covariance terms grows rapidly as the choice set increases. Numerical optimization procedures “have difficulty” (which I think means converge slowly and may have multi–optima) for high dimensional parameter vectors. These are problems with MNP not simulation methods.

The last item motivates the use of factor structures, $\epsilon_{ij} = \phi_{i} \alpha_{j} + \delta_{ij}$, with $\delta_{ij}$ as mutually independent mean zero shocks independent of $\phi_{i}$. This is a generalization of the more common error component model $\epsilon_{ij} = \kappa_{ij} + v_{ij}$ where $\kappa_{ij}$ is independent of $v_{ij}$ and $v_{ij}$ is frequently assumed to be iid across people and alternatives, but could include some dependence across alternatives.
Multinomial Logit

If \( m = 2 \) and \( \epsilon^1, \epsilon^2 \) are independent, identically distributed Extreme Value Type 1 random variables,\(^2\) then \( \eta_n = \epsilon^i_n - \epsilon^j_n \) has a logistic distribution, its CDF is

\[ G_\eta(\psi) = \left[ 1 + \exp(-\psi) \right]^{-1}. \]

The choice probability becomes

\[ P(2 \mid \zeta_n, \theta) = \Pr[\eta_n < \psi^{1,2}_n(\alpha) \mid \zeta_n] \]
\[ = \left( 1 + \exp\left[-\psi^{1,2}_n(\alpha)\right]\right)^{-1} \]
\[ = \frac{\exp\left[V^*2(x^2_n, z_n; \alpha)\right]}{\exp\left[V^*1(x^1_n, z_n; \alpha)\right] + \exp\left[V^*2(x^2_n, z_n; \alpha)\right]} \]

If \( \psi^{1,2}_n = \alpha' \zeta_n \) then the binary logit is sometimes referred to as logistic regression. Be careful of terminology as “MNL” means

\[ P(j \mid z_n) = \frac{\exp (z_n a_j)}{\sum_{k=1}^m \exp (z_n a_k)} \]

Observe in the random utility with additive errors, we estimate one vector of length \( q, \alpha \). In the “usual” MNL estimates \( m - 1 \) length–\( q \) vectors of estimated parameters \( \alpha_k \). The idea that we only have characteristics on the individual \( z_n \) and not alternative–specific. For example, in STATA the command \texttt{logic} estimates the common logistic regression. McFadden’s model is estimated by the STATA command \texttt{asclogit} — alternative–specific conditional logit.

Example of the MNL

To see the limitations of induced by independence of the \( \epsilon^l \) consider again the characteristics of the MNL. The MNL exhibits “independence from irrelevant alternatives” (IIA): the odds ratio between any two alternatives is independent of the characteristics or even existence of any other alternative. This implies that the introduction of a new alternative reduces all the selection probabilities proportionately. Yet, it is more natural to believe that the selection probabilities of close substitutes will decline more.

The IIA property imposes a very restrictive pattern of elasticities of the choice probabilities with respect to the characteristics \( x^i \) (characteristics of alternative \( i \)). One can show that

\[ \frac{\partial \ln P(j \mid \zeta_n)}{\partial \ln x^i} = -P(i \mid \zeta_n) \frac{\partial V^*_i(x^i, z_n)}{\partial \ln x^i}. \]

Importantly the elasticity depends only on \( i \) and not on \( j \). Hence, choice probabilities share a uniform set of cross–elasticities, and this imposes a severe restriction on the type of substitution responses (among alternatives) than can be modeled.

This is the restrictiveness purchased by the computational simplicity of the MNL.

\(^2\chi^* \text{ distributed as Extreme Value Type 1 if its CDF is } \Pr \{ \chi \leq x \} = \exp(-\exp(-x)). \)
McFadden’s GEV generates a family of distributions that include the MNL, and the less restrictive Nested-MNL as members, and many less restrictive members as well.

**Framework Supporting GEV**

For any model based on (1) or (2) the implied selection probabilities have a common representation in terms of the joint distribution assumed for $\epsilon_1, \ldots, \epsilon_J$. Alternative $j$ is chosen by individual $n$ if $u^j_n > u^i_n$, $\forall i \neq j$. This happens if and only if,

$$
\epsilon^j_n < \epsilon^i_n + V^*_j - V^*_i
$$

or, defining $\phi^i_n(\alpha)$ as the difference in the non–stochastic utilities,

$$
\epsilon^i_n < \epsilon^j_n + \phi^i_n(\alpha).
$$

Since the inequalities are defined only in terms of the utility differences rather than the utilities $V^*_j$ themselves, any additive component of the $V^*_j$ that is independent of $j$ can not be identified. Assuming that $\epsilon^1, \ldots, \epsilon^J$ have a continuous distribution, the conditional probability of the event of selecting $j$ is

$$
P(j|\zeta_n; \theta) = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\epsilon^j_n + \phi^i_n(\alpha)} \cdots \int_{-\infty}^{\epsilon^j_n + \phi^i_n(\alpha)} g(\epsilon^1, \ldots, \epsilon^J) \prod_{i \neq j} d\epsilon^i \right] d\epsilon^j
$$

$g$ is the joint p.d.f of the $\epsilon^1, \ldots, \epsilon^J$.

**GEV and Nested Logit**

Define $G$ to be the c.d.f. of the corresponding density, $g$ and differentiate w.r.t. to its $j^{th}$ argument,

$$
Pr[\epsilon^j \in (E^j, E^j + d E^j), \epsilon^j \leq E^j, \forall i \neq j] = G^j(E^1, \ldots, E^J) dE^j
$$

$G^j = \partial G / \partial E^j$. Setting $E^j$ equal to $\epsilon^j + \phi^i_n(\alpha)$ and integrating over the whole range of $\epsilon^j$ yeilds:

$$
Pr[j|\zeta_n, \theta] = \int_{-\infty}^{\infty} G^j(\epsilon^j + \phi^1_n(\alpha), \ldots, \epsilon^j + \phi^J_n(\alpha)) d\epsilon^j
$$

This form is sometimes convenient as a starting point for the derivation of choice probabilities (esp. GEV).
GEV Defined

Consider a function $Y(\cdot)$ with the following properties:

1. $Y(v_1, \ldots, v_m)$ is a non-negative function with non-negative arguments;
2. $Y(v_1, \ldots, v_m)$ is linearly homogenous in its $J$ arguments;
3. $\lim_{v_i \rightarrow \infty} Y(v_1, \ldots, v_m) = +\infty$ for all $i = 1, \ldots, m$;
4. The partial derivative $\frac{\partial^k Y}{\partial v_{i_1} \partial v_{i_2} \ldots \partial v_{i_k}}$ is non-negative for $k$ odd and non-positive for $k$ even, provided that $i_1, \ldots, i_k$ are distinct.

The random variables $\epsilon_1, \ldots, \epsilon_m$ are said to have a generalized extreme value (GEV) distribution if their c.d.f. is expressible in the form:

$$F_\epsilon(E_1, \ldots, E_m) = \exp(-Y(\exp(-E_1), \ldots, \exp(-E_m)))$$

and the joint p.d.f. is found in the usual way by partial differentiating with respect to each of the $\epsilon_i$. (p.d.f. awkward and not informative.)

It is called the GEV because its marginal distributions are all of displaced univariate extreme value form.

An important special case of the GEV corresponds to linearity in the function $Y$. Can show that a linear $Y$ produces the joint distribution for $J$ independent extreme value variables. Thus, $Y$ must be nonlinear to introduce non-independent distributions.

McFadden shows that the choice probabilities are

$$P(j | \zeta_n; \theta) = \frac{\partial \ln Y(\exp(V^*_{n1}), \ldots, \exp(V^*_{nj}))}{\partial V^*_{nj}}$$

$$= \frac{\exp(V^*_{nj}) Y_j[\exp(V^*_{n1}), \ldots, \exp(V^*_{nj})]}{Y[\exp(V^*_{n1}), \ldots, \exp(V^*_{nj})]}$$

where $Y_j$ is the partial derivative of $Y$ with respect to its $j^{th}$ argument.

The MNL is a member of the GEV with $Y(v_1, \ldots, v_J) = (\sum v_i^\rho)^{1/\rho}$ and $\rho \geq 1$. Note: this parameterization differs from the usual one, say expressed in terms of $\sigma$, with $\rho = \frac{1}{1-\sigma}$. Thus, $\sigma = 0$, corresponds to $\rho = 1$.

The choice probabilities above imply elasticities of the form

$$\frac{\partial \ln P(j | \zeta_n)}{\partial \ln x^i} = \left[ \frac{Y_{ji}}{Y_j} \exp(V^*_{ni}) \right] \frac{\partial V^*_{ji}(x^i, \zeta_n)}{\partial \ln x^i}$$

where $Y_j$ and $Y_{ji}$ are the first and second derivatives of $Y(\cdot)$ evaluated at $\exp(V^*_{n1}), \ldots, \exp(V^*_{nj})$. The elasticities are independent of $j$ only if the cross-derivatives $Y_{jk}$ are proportional to $Y_j$. 


Consider specification put forth by ?:

\[
Y(\nu_1, \ldots, \nu_m) = \sum_{i=1}^{m} \sum_{k=1}^{m} A_{ik}^n (v_i^\rho + v_k^\rho)^{1/\rho}
\]

where, \( \rho \geq 1 \) is a parameter that controls the general scale of correlation and \( A_{ik}^n \) is a quantity reflecting the similarity between alternatives \( i \) and \( k \) for individual \( n \).

The implied choice probabilities are

\[
P(j|\zeta_n; \theta) = \frac{\exp(\rho V_{nj}^* \sum_{i=1}^{j} (A_{ij}^n + A_{ji}^n)) \left(\exp(\rho V_{nj}^*) + \exp(\rho V_{nk}^*)\right)^{(1-\rho)/\rho}}{\sum_{i=1}^{j} \sum_{k=1}^{j} A_{ik}^n \left(\exp(\rho V_{ni}^*) + \exp(\rho V_{nk}^*)\right)^{1/\rho}}
\]

**Nested MNL**

The Nested MNL obtained from

\[
Y(\nu_1, \ldots, \nu_m) = \sum_{C \in F_B} \sum_{i \in C} a(Z_c) \left( \sum_{i \in C} v_i^{\rho(z_c)} \right)^{1/(\rho(z_c))}
\]

where \( a(z_c) \geq 0 \) and \( \rho(z_c) = 1/(1 - \sigma(z_c)) \) with \( 0 \leq \sigma(z_c) < 1 \). Where \( F_B \) is a subset of alternatives from the choice set \( B = \{1, \ldots, m\} \). \( z_c \) corresponds to the vector of attributes of all the choices in subset \( C \), \( z_c = \{z_i\}_{i \in C} \). The choice probabilities can be written as

\[
P(i|B) = \sum_{C \in F_B} P(i|C)Q(C|B)
\]

where

\[
P(i|C) = \frac{\exp(V_i^{*})/(1 - \sigma(z_c))}{\sum_{k \in C} \exp(V_k^{*})/(1 - \sigma(z_c))} \quad \text{for} \quad i \in C
\]

\[
Q(C|B) = \frac{a(z_c) \exp((1 - \sigma(z_c))h_c)}{\sum_{A \in F_B} a(z_c) \exp((1 - \sigma(z_c))h_A)}
\]

with \( h_c = \ln \sum_{k \in C} \exp(V_k^{*}) \) where I have dropped the explicit conditioning on person \( n \).

The system can be interpreted as a choice system in which decision makers invoke a subset of alternatives \( C \) from (the complete choice set) \( B \) and then select an alternative from \( C \). So the

\footnote{The summation \( C \) in \( F_B \) that appears below stems from assumption that alternatives are defined by their characteristics. So for the example given in class on neighborhoods (branch) and houses (twigs). A house with the same observable attributes is considered to be the same alternative. Thus, “alternative \( j \)” can appear in more than one subset. That is the subsets do not have to be mutually exclusive.}
$P(i|C)$ represents the selection at the lower level of the system (“selection of the twig”). While $Q(C|B)$ is the choice of the subset (“selection of the branch”). Notice that $P(i|C)$ has a **MNL** form. Associated with each subset $C$ is an **inclusive value**, $h_C$. Choice probabilities for the invoked set $C$ are multinomial logit functions of the inclusive values. The function $\sigma(z_C)$ is a measure of the similarity of alternatives with $C$. When alternatives in $C$ are very similar $\sigma(z_C)$ is near one, the conditional choice probability $P(i|C)$ selects with high probability the alternative with the highest value of $C$ of $V_j^\ast$. Then $h_C$ is approximately $\max_{j\in C} V_j^\ast$ and the choice of an invoked set using $Q(C|B)$, the set $C$ is assessed approximately as if it contained a single alternative with a scale value of

$$\max_{j\in C} V_j^\ast.$$  

More complicated models possible with nesting of higher levels. Inclusive values arise at each level.