has cumulative distribution:

$$exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in B_k}e^{-\varepsilon_{nj}/\lambda_k}\right)^{\lambda_k}\right). \tag{4.1}$$

This distribution is a type of generalized extreme value (GEV) distribution. It is a generalization of the distribution that gives rise to the logit model. For logit, each ε_{nj} is independent with a univariate extreme value distribution. For this GEV, the marginal distribution of each ε_{nj} is univariate extreme value. However, the ε_{nj}'s are correlated within nests. For any two alternatives j and m in nest B_k, ε_{nj} is correlated with ε_{nm}. For any two alternatives in different nests, the unobserved portion of utility is still uncorrelated: $Cov(\varepsilon_{nj}, \varepsilon_{nm}) = 0$ for any $j \in B_k$ and $m \in B_\ell$ with $\ell \neq k$.

The parameter λ_k is a measure of the degree of independence in unobserved utility among the alternatives in nest k. A higher value of λ_k means greater independence and less correlation. The statistic $(1 - \lambda_k)$ is a measure of correlation, in the sense that as λ_k rises, indicating less correlation, this statistic drops. As McFadden (1978) points out, the correlation is actually more complex than $(1 - \lambda_k)$, but $(1 - \lambda_k)$ can be used as an indication of correlation. A value of $\lambda_k = 1$ indicates complete independence within nest k, that is, no correlation. When $\lambda_k = 1$ for all k, representing independence among all the alternatives in all nests, the GEV distribution becomes the product of independent extreme value terms, whose distribution is given in (3.2). In this case, the nested logit model reduces to the standard logit model.

As shown by the authors cited above, this distribution for the unobserved components of utility gives rise to the following choice probability for alternative $i \in B_k$:

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k-1}}{\sum_{\ell=1}^{K}\left(\sum_{j\in B_\ell} e^{V_{nj}/\lambda_\ell}\right)^{\lambda_\ell}}. \tag{4.2}$$

We can use this formula to show that IIA holds within each subset of alternatives but not across subsets. Consider alternatives $i \in B_k$ and $m \in B_\ell$. Since the denominator of (4.2) is the same for all alternatives, the ratio of probabilities is the ratio of numerators:

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k} \left(\sum_{j \in B_k} e^{V_{nj}/\lambda_k} \right)^{\lambda_k - 1}}{e^{V_{nm}/\lambda_\ell} \left(\sum_{j \in B_\ell} e^{V_{nj}/\lambda_\ell} \right)^{\lambda_\ell - 1}}.$$

If $k = \ell$ (i.e., i and m are in the same nest) then the terms in parentheses cancel out and we have

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k}}{e^{V_{nm}/\lambda_\ell}}$$

This ratio is independent of all other alternatives. For $k \neq \ell$ (i.e., i and m are in different nests), the terms in parentheses do not cancel out. The ratio of probabilities depends on the attributes of all alternatives in the nests that contain i and m. Note, however, that the ratio does not depend on the attributes of alternatives in nests other those containing i and m. A form of IIA holds, therefore, even for alternatives in different nests. This form of IIA can be loosely described as "independence from irrelevant nests" or IIN. With a nested logit model, IIA holds over alternatives in each nest and IIN holds over alternatives in different nests. This property of nested logit models is reinforced in the next section when we decompose the nested logit probability into two standard logit probabilities.

When $\lambda_k = 1$ for all k (and hence $1 - \lambda_k = 0$), indicating no correlation among the unobserved components of utility for alternatives within a nest, the choice probabilities become simply logit. The nested logit model is a generalization of logit that allows for a particular pattern of correlation in unobserved utility.

The parameter λ_k can differ over nests, reflecting different correlation among unobserved factors within each nest. The researcher can constrain the λ_k's to be the same for all (or some) nests, indicating that the correlation is the same in each of these nests. Hypothesis testing can be used to determine whether constraints on the λ_k's are reasonable. Testing the constraint $\lambda_k = 1 \ \forall k$ is equivalent to testing whether the standard logit model is a reasonable specification against the more general nested logit. These tests are performed most readily with the likelihood ratio statistic described in section (3.8.2).

The value of λ_k must be within a particular range for the model to be consistent with utility maximizing behavior. If $\lambda_k \ \forall k$ is between zero and one, the model is consistent with utility maximization for all possible values of the explanatory variables. For λ_k greater than one, the model is consistent with utility maximizing behavior for some range of the explanatory variables but not for all values. Kling and Herriges (1995) and Herriges and Kling (1996) provide tests of consistency of nested logit with utility maximization when $\lambda_k > 1$, and Train, McFadden and Ben-Akiva (1987) and Lee (1999) provide examples of models for which $\lambda_k > 1$. A value of λ_k below zero is inconsistent with utility maximization and implies that improving the attributes of an alternative (such as lowering its price) can decrease the probability of the alternative being chosen. With positive λ_k, the nested logit approaches the "elimination by aspects" model of Tversky (1972) as $\lambda_k \to 0$.

In the notation that we have been using, each λ_k is a fixed parameter, which implies that all decision-makers have the same correlations among unobserved factors. In reality, correlations might differ over decision-makers based on their observed characteristics. To accommodate this possibility, each λ_k can be specified to be a parametric function of observed demographics or other variables, as long as the function maintains a positive value. For example, Bhat (1997) specifies $\lambda = exp(\alpha z_n)$, where z_n is a vector of characteristics of decision-maker n and α is a vector of parameters to be estimated along with the parameters that enter representative utility. The exponential transformation assures that λ is positive.

4.2.3 Decomposition into two logits

Expression (4.2) is not very illuminating as a formula. However, the choice probabilities can be expressed in an alternative fashion that is quite simple and readily interpretable. Without loss of generality, the observed component of utility can be decomposed into two parts: (1) a part labeled W that is constant for all alternatives within a nest, and (2) a part labeled Y that varies over alternatives within a nest. Utility is written as:

$$U_{nj} = W_{nk} + Y_{nj} + \varepsilon_{nj} \tag{4.3}$$

for $j \in B_k$, where:

W_{nk} depends only on variables that describe nest k. These variables differ over nests but not over alternatives within each nest.

Y_{nj} depends on variables that describe alternative j. These variables vary over alternatives within nest k.

Note that this decomposition is fully general since for any W_{nk}, Y_{nj} is defined as $V_{nj} - W_{nk}$.

With this decomposition of utility, the nested logit probability can be written as the product of two standard logit probabilities. Let the probability of choosing alternative $i \in B_k$ be expressed as the product of two probabilities, namely: the probability that an alternative within nest B_k is chosen and the probability that the alternative i is chosen given that an alternative in B_k is chosen. This is denoted as

$$P_{ni} = P_{ni|B_k} P_{nB_k},$$

where $P_{ni|B_k}$ is the conditional probability of choosing alternative i given that an alternative in nest B_k is chosen, and P_{nB_k} is the marginal probability of choosing an alternative in nest B_k (with the marginality being over all alternatives in B_k.) This equality is exact since any probability can be written as the product of a marginal and a conditional probability.

The reason for decomposing P_{ni} into a marginal and a conditional probability is that, with the nested logit formula for P_{ni}, the marginal and conditional probabilities take the form of logits. In particular, the marginal and conditional probabilities can be expressed as

$$P_{nB_k} = \frac{e^{W_{nk}+\lambda_k I_{nk}}}{\sum_{\ell=1}^{K} e^{W_{n\ell}+\lambda_\ell I_{n\ell}}} \tag{4.4}$$

$$P_{ni|B_k} = \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j \in B_k} e^{Y_{nj}/\lambda_k}} \tag{4.5}$$

where

$$I_{nk} = ln \sum_{j \in B_k} e^{Y_{nj}/\lambda_k}$$

The derivation of these expressions from the choice probability (4.2) simply involves algebraic rearrangement. For interested readers, it is given in section (4.2.5).

Stated in words, the probability of choosing an alternative in B_k takes the form of the logit formula, as if it were a model for a choice among nests. This probability includes variables W_{nk} that vary over nests but not over alternatives within each nest. It also includes a term labeled I_{nk} whose meaning we elucidate below. The conditional probability of choosing i given that an alternative in B_k is chosen is also a logit formula, as if it were a model for the choice among the alternatives within the nest. This conditional probability includes variables Y_{nj} that vary over alternatives within the nest. Note that these terms are divided by λ_k such that, when Y_{nj} is linear in parameters, the coefficients that enter this conditional probability are the original coefficients divided by λ_k. It is customary to refer to the marginal probability (choice of nest) as the "upper model" and the conditional probability (choice of alternative within the nest) as the "lower model," reflecting their relative positions in Figure 4.1.

The term I_{nk} links the upper and lower models by bringing information from the lower model into the upper model. Ben-Akiva (1972) first identified the correct formula for this link. In particular, I_{nk} is the log of the denominator of the lower model. This formula has an important meaning. Recall from the discussion of consumer surplus for a logit model (section 3.5) that the log of the denominator of the logit model is the expected utility that the decision-maker obtains from the choice situation, as shown by Williams (1977) and Small and Rosen (1981). The same interpretation applies here: $\lambda_k I_{nk}$ is the expected utility that decision-maker n receives from the choice among the alternatives in nest B_k. The formula for expected utility is the same here as for a logit model because, conditional on the nest, the choice of alternatives within the nest is indeed a logit, as given by equation (4.5). I_{nk} is often called the "inclusive value" or "inclusive utility" of nest B_k. It is also called the "log-sum term" because it is the log of a sum (of exponentiated representative utilities). The term "inclusive price" is sometimes used; however the negative of I_{nk} more closely resembles a price.

The coefficient of I_{nk} in the upper model is λ_k, which is often called the log-sum coefficient. As discussed above, λ_k reflects the degree of independence among the unobserved portions of utility for alternatives in nest B_k, with a lower λ_k indicating less indendendent (more correlation.)

It is appropriate that the inclusive value term enters as an explanatory variable in the upper model. Stated loosely, the probability of choosing nest B_k depends on the expected utility that the person receives from that nest. This expected utility includes the utility that he receives no matter which alternative he chooses in the nest, which is W_{nk}, plus the expected extra utility that he receives by being able to choose the best alternative in the nest, which is $\lambda_k I_{nk}$.

Recall that the coefficients that enter the lower model are divided by λ_k, as given in equation 4.5. Models have been specified and estimated without dividing by λ_k in the lower model. Daly (1987) and Greene (2000) describes such a model, and the computer software STATA includes it as their nested logit model in the nlogit command. The package NLOGIT allows either specification. If the coefficients in the lower model are not divided by λ_k, the choice probabilities are not the same as those given in equation 4.2. As shown in the derivation in section 4.2.5, the division by λ_k is needed for the product of the conditional and marginal probabilities to equal the nested logit probabilities given by equation 4.2. However, the fact that the model does not give the probabilities in equation 4.2 does not necessarily mean that the model is inappropriate. Koppelman and Wen (1998) and Hensher and Greene (forthcoming) compare the two approaches (dividing by λ_k versus not) and show that the latter model is not consistent with utility maximization when any coefficients are common across nests (such as a cost coefficient that is the same for bus and car modes.) Heiss (2002) points out the converse: if no coefficients are common over nests, then the latter model is consistent with utility maximization, since the necessary division by λ_k in each nest is accomplished implicitly (rather than explicitly) by allowing separate coefficients in each nests such that the scale of coefficients differs over nests. When coefficients are common over nests, she found that not dividing by λ_k provides counter-intuitive implications.

4.2.4 Estimation

The parameters of a nested model can be estimated by standard maximum likelihood techniques. Substituting the choice probabilities of expression (4.2) into the log likelihood function gives an explicit function of the parameters of this model. The value of the parameters that maximizes this function is, under fairly general conditions, consistent and efficient (Brownstone and Small, 1989).

Computer routines are available in commercial software packages for estimating nested models by maximum likelihood. Hensher and Greene (forthcoming) provide a guide for nested logits using available software. Numerical maximization is sometimes difficult since the log-likelihood function is not globally concave and even in concave areas is not close to a quadratic. The researcher might need to help the routines by trying different algorithms and/or starting values, as discussed in Chapter 8.

Instead of performing maximum likelihood, nested logit models can be estimated consistently (but not efficiently) in a sequential fashion, exploiting the fact that the choice probabilities can be decomposed into marginal and conditional probabilities that are logit. This sequential estimation is performed "bottom up." The lower models (for the choice of alternative within a nest) are estimated first. Using the estimated coefficients, the inclusive value term is calculated for each lower model. Then the upper model (for choice of nest) is estimated, with the inclusive value terms entering as explanatory variables.

Sequential estimation creates two difficulties that argue against its use. First, the standard errors of the upper-model parameters are biased downward, as Amemiya (1978) first pointed out. This bias arises because the variance of the inclusive value estimate that enters the upper model is not incorporated into the calculation of standard errors. With downwardly biased standard errors, smaller confidence bounds and larger t-statistics are estimated for the parameters than are true, and the upper model will appear to be better than it actually is. Ben-Akiva and Lerman (1985, p. 298) give a procedure for adjusting the standard errors to eliminate the bias.

Second, it is usually the case that some parameters appear in several submodels. Estimating the various upper and lower models separately provides separate estimations of whatever common parameters appear in the model. Simultaneous estimation by maximum likelihood assures that the common parameters are constrained to be the same wherever they appear in the model.

These two complications are symptoms of a more general circumstance, namely, that sequential estimation of nested logit models, while consistent, is not as efficient as simultaneous estimation by maximum likelihood. With simultaneous estimation, all information is utilized in the estimation of each parameter, and parameters that are common across components are necessarily constrained to be equal. Since commercial software is available for simultaneous estimation, there is little reason to estimate a nested logit sequentially. If problems arise in simultaneous estimation, then the researcher might find it useful to estimate the model sequentially and then use the sequential estimates as starting values in the simultaneous estimation. The main value of the decomposition of the nested logit into its upper and lower components comes not in its use as a estimation tool but rather as a heuristic device: the decomposition helps greatly in understanding the meaning and structure of the nested logit model.

4.2.5 Equivalence of nested logit formulas

We asserted in section (4.2.3) that the product of the marginal and conditional probabilities in (4.4) and (4.5) equals the joint probability in (4.2). We now verify this assertion.

$$\begin{aligned}
P_{ni} &= \frac{e^{V_{ni}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k-1}}{\sum_{\ell=1}^{K}\left(\sum_{j\in B_\ell} e^{V_{nj}/\lambda_\ell}\right)^{\lambda_\ell}} \text{ by eq (4.2)} \\
&= \frac{e^{V_{ni}/\lambda_k}}{\sum_{j\in B_k} e^{V_{nj}/\lambda_k}} \frac{\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k}}{\sum_{\ell=1}^{K}\left(\sum_{j\in B_\ell} e^{V_{nj}/\lambda_\ell}\right)^{\lambda_\ell}} \\
&= \frac{e^{(W_{nk}+Y_{ni})/\lambda_k}}{\sum_{j\in B_k} e^{(W_{nk}+Y_{nj})/\lambda_k}} \frac{\left(\sum_{j\in B_k} e^{(W_{nk}+Y_{nj})/\lambda_k}\right)^{\lambda_k}}{\sum_{\ell=1}^{K}\left(\sum_{j\in B_\ell} e^{(W_{n\ell}+Y_{nj})/\lambda_\ell}\right)^{\lambda_\ell}} \text{ using (4.3)} \\
&= \frac{e^{W_{nk}/\lambda_k} e^{Y_{ni}/\lambda_k}}{e^{W_{nk}/\lambda_k}\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}} \frac{e^{W_{nk}}\left(\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}\right)^{\lambda_k}}{\sum_{\ell=1}^{K} e^{W_{n\ell}}\left(\sum_{j\in B_\ell} e^{Y_{nj}/\lambda_\ell}\right)^{\lambda_\ell}} \\
&= \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}} \frac{e^{W_{nk}+\lambda_k I_{nk}}}{\sum_{\ell=1}^{K} e^{W_{n\ell}+\lambda_\ell I_{n\ell}}} \\
&= P_{ni|B_k} P_{nB_k}.
\end{aligned}$$

where the next to last equality is because $I_{nk} = ln \sum_{j\in B_k} e^{Y_{nj}/\lambda_k}$, recognizing that $e^x b^c = e^{x+cln(b)}$.

4.3 Three-Level Nested Logit

The nested logit model that we have discussed up this this point is called a two-level nested logit model because there are two levels of modeling: the marginal probabilities (upper model) and the conditional probabilities (lower models.) In the case of the mode choice, the two levels are the marginal model of auto versus transit and the conditional models of type of auto or transit (auto alone or carpool given auto, and bus or rail given transit).

In some situations, three- or higher level nested logit models are appropriate. Three-level models are obtained by partitioning the set of alternatives into nests and then partitioning each nest into subnests. The probability formula is a generalization of (4.2) with extra sums for the subnests within the sums for nests. See McFadden (1978) or Ben-Akiva and Lerman (1985) for the formula.

As with a two-level nested logit, the choice probabilities for a three-level model can be expressed as a series of logits. The top model describes the choice of nest; the middle models describe the choice of subnest within each nest; and the bottom models describe the choice of alternative within each subnest. The top model includes an inclusive value term for each nest. This term represents the expected utility that the decision-maker can obtain from the subnests within the nest. It is calculated as the log of the denominator of the middle model for that nest. Similarly, the middle models include an inclusive value term for each subnest that represents the expected utility that the decision-maker can obtain from the alternatives within the subnest. It is calculated as the log of the denominator of the bottom model for the subnest.

As an example, consider a household's choice of housing unit within a metropolitan area. The household has a choice among all the available housing units in the city. The housing units are available in different neighborhoods in the city and with different numbers of bedrooms. It is reasonable to assume that there are unobserved factors that are common to all units in the same neighborhood, such as the proximity to shopping and entertainment. The unobserved portion of utility is therefore expected to be correlated over all units in a given neighborhood. There are also unobserved factors that are common to all units with the same number of bedrooms, such as the convenience of working at home. We therefore expect the unobserved utility to be even more highly correlated among units of the same size in the same neighborhood than between units of different size in the same neighborhood. This pattern of correlation can be represented by nesting the units by neighborhood and then subnesting them by number of bedrooms. A tree diagram depicting this situation is given in Figure 4.2 for San Francisco. There are three levels of submodels: the probability for choice of neighborhood, the probability for choice of number of bedrooms given the neighborhood, and the choice of unit given the neighborhood and number of bedrooms.

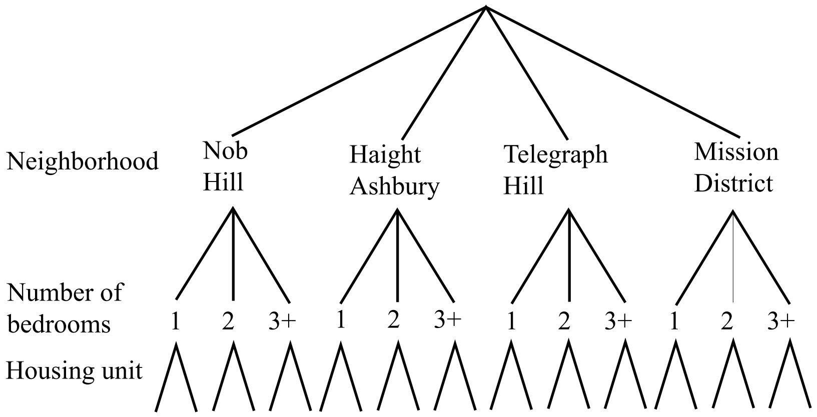

Figure 4.2: Three-level nested logit.

A nested logit model with this nesting structure embodies IIA in the following ways. (1) The ratio of probabilities of two housing units in the same neighborhood and with the same number of bedrooms is independent of the characteristics of all other units. For example, lowering the price of a two-bedroom apartment in Pacific Heights draws proportionately from all one-bedroom units on Russian Hill. (2) The ratio of probabilities of two housing units in the same neighborhood but with different numbers of bedrooms is independent of the characteristics of units in other neighborhoods but depends on the characteristics of units in the same neighborhood that have the same number of bedrooms as either of these units. Lowering the price of a two-bedroom apartment in Pacific Heights draws proportionately from one- and two-bedroom units on Russian Hill, but draws disproportionately from two-bedroom units in Pacific Heights relative to one-bedroom units in Pacific Heights. (3) The ratio of probabilities of two housing units in different neighborhoods depends on the characteristics of all the other housing units in those neighborhoods but not on the characteristics of units in other neighborhoods. Lowering the price of a two-bedroom apartment in Pacific Heights draws proportionately from all units outside of Pacific Heights but draws disproportionately from units in Pacific Heights relative to units outside of Pacific Heights.

Each layer of a nesting in a nested logit introduces parameters that represent the degree of correlation among alternatives within the nests. With the full set of alternatives partitioned into nests, the parameter λ_k is introduced for nest k, as described for two-level models. If the nests are further partitioned into subnests, then a parameter σ_{mk} is introduced for subnest m of nest k. Using the decomposition of the probability into a series of logit models, σ_{mk} is the coefficient of the inclusive value term in the middle model, and $\lambda_k \sigma_{mk}$ is the coefficient of the inclusive value term in the top model. Just as for a two-level nested logit, the value of these parameters must be in certain ranges to be consistent with utility maximization. If $0 < \lambda_k < 1$ and $0 < \sigma_{mk} < 1$, then the model is consistent with utility maximization for all levels of the explanatory variables. A negative value for either parameter is inconsistent with utility maximization. And values greater than one are consistent for a range of explanatory variables.

4.4 Overlapping Nests

For the nested logit models that we have considered, each alternative is a member of only one nest (and, for three-level models, only one subnest.) This aspect of nested logit models is a restriction that is sometimes inappropriate. For example, in our example of mode choice, we put carpool and auto alone into a nest because they have some similar unobserved attributes. However, carpooling also has some unobserved attributes that are similar to bus and rail, such as a lack of flexibility in scheduling (the worker cannot go to work whenever he wants each day but rather has to go at the time that the carpool has decided, similar to taking a bus or rail line with fixed departure times.) It would be useful to have a model in which the unobserved utility for the carpool alternative could be correlated with that of auto alone and also correlated, though to a different degree, with that of bus and rail. Stated equivalently, it would be useful for the carpool alternative to be in two nests: one with auto alone and another with bus and rail.

Several kinds of GEV models have been specified with overlapping nests, such that an alternative can be a member of more than one nest. Vovsha (1997), Bierlaire (1998), and Ben-Akiva and Bierlaire (1999) have proposed various models called cross-nested logits (CNL) that contain multiple overlapping nests. Small (1987) considered a situation where the alternatives have a natural order, such as the number of cars that a household owns (0, 1, 2, 3, and so on) or the destination for shopping trips, with the shopping areas ordered by distance from the household's home. He specified a model called ordered generalized extreme value (OGEV) in which the correlation in unobserved utility between any two alternatives depends on their proximity in the ordering. This model has overlapping nests like the cross-nested logits, but each nest consists of two alternatives and a pattern is imposed on the correlations (higher correlation for closer pairs). Small (1994) and Bhat (1998*b*) described a nested version of the OGEV, which is similar to a nested logit except that the lower models (for the alternatives given the nests) are OGEV rather than standard logit. Chu (1981, 1989) proposed a model called the paired combinatorial logit (PCL) in which each pair of alternatives constitutes a nest with its own correlation. With J alternatives, each alternative is a member of $J-1$ nests and the correlation of its unobserved utility with each other alternative is estimated. Wen and Koppelman (2001) have developed a "generalized nested logit" (GNL) model that includes the PCL and other cross-nested models as special cases. I describe below the PCL and GNL, the former because of its simplicity and the later because of its generality.

4.4.1 Paired combinatorial logit

Each pair of alternatives is considered to be a nest. Since each alternative is paired with each of the other alternative, each alternative is member of $J-1$ nests. A parameter labeled λ_{ij} indicates the degree of independence between alternatives i and j. Stated equivalently: $(1-\lambda_{ij})$ is a measure of the correlation between the unobserved utility of alternative i and that of alternative j. This parameter is analogous to the λ_k in a nested logit model, where λ_k indicates the degree of independence of alternatives within the nest and $1-\lambda_k$ is a measure of correlation within the nest. And as with nested logit, the PCL model becomes a standard logit when $\lambda_{ij}=1$ for all pairs of alternatives.

The choice probabilities for the PCL model are

$$P_{ni} = \frac{\sum_{j\neq i} e^{V_{ni}/\lambda_{ij}} \left(e^{V_{ni}/\lambda_{ij}} + e^{V_{nj}/\lambda_{ij}}\right)^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{\ell=k+1}^{J}\left(e^{V_{nk}/\lambda_{k\ell}} + e^{V_{n\ell}/\lambda_{k\ell}}\right)^{\lambda_{k\ell}}} \tag{4.6}$$

The sum in the numerator is over all $J-1$ nests that alternative i

is in. For each of these nests, the term being added is the same as the numerator of the nested logit probability (4.2). In this sense, the PCL is like the nested logit except that it allows i to be in more than one nest. The denominator in the PCL also take the same form as in a nested logit: it is the sum over all nests of the sum of the $exp(V/\lambda)$'s within the nest, raised to the appropriate power λ. If λ_{ij} is between zero and one for all ij pairs, then the model is consistent with utility maximization for all levels of the data. It is easy to verify that P_{ni} becomes the standard logit formula when $\lambda_{ij} = 1 \; \forall \, i, j$. In their application, Koppelman and Wen (2000) found PCL to perform better than nested logit or standard logit.

The researcher can test the hypothesis that $\lambda_{ij} = 1$ for some or all of the pairs, using the likelihood ratio test of section (3.8.2). Acceptance of the hypothesis for a pair of alternatives implies that there is no significant correlation in the unobserved utility for that pair. The researcher can also place structure on the pattern of correlation. For example, correlations can be assumed to be the same among a group of alternatives; this assumption is imposed by setting $\lambda_{ij} = \lambda_{k\ell}$ for all i, j, k, and ℓ in the group. Small's OGEV model is a PCL model in which λ_{ij} is specified to be a function of the proximity between i and j. With a large number of alternatives, the researcher will probably need to impose some form of structure on the λ_{ij}'s , simply to avoid the proliferation of parameters that arises with large J. This proliferation of parameters, one for each pair of alternatives, is what makes the PCL so flexible. The researcher's goal is to apply this flexibility meaningfully for his particular situation.

As discussed near the end of section (2.5), since the scale and level of utility is immaterial, at most $J(J-1)/2-1$ covariance parameters can be estimated in a discrete choice model. A PCL model contains $J(J-1)/2$ λ's: one for each alternative paired with each other alternative, recognizing that i paired with j is the same as j paired with i. The number of $\lambda's$ exceeds the number of identifiable covariance parameters by exactly one. The researcher must therefore place at least one constraint on the λ's. This can be accomplished by normalizing one of the λ to 1. If structure is placed on the pattern of correlation, as described in the previous paragraph, then this structure will usually impose the normalization automatically.

4.4.2 Generalized nested logit

Nests of alternatives are labeled $B_1, B_2, \ldots, B_K$. Each alternative can be a member of more than one nest. Importantly, an alternative can be in a nest to varying degrees. Stated differently, an alternative is allocated among the nests, with the alternative being in some nests more than other nests. An "allocation" parameter α_{jk} reflects the extent to which alternative j is a member of nest k. This parameter must be non-negative: $\alpha_{jk} \geq 0 \ \forall j, k$. A value of zero means that the alternative is not in the nest at all. Interpretation is facilitated by having the allocation parameters sum to one over nests for any alternative: $\sum_k \alpha_{jk} = 1 \ \forall j$. Under this condition, α_{jk} reflects the portion of the alternative that is allocated to each nest.

A parameter λ_k is defined for each nest and serves the same function as in nested logit models, namely to indicate the degree of independence among alternatives within the nest: higher λ_k translates into greater independence and less correlation.

The probability that person n chooses alternative i is

$$P_{ni} = \frac{\sum_k (\alpha_{ik} e^{V_{ni}})^{1/\lambda_k} \left(\sum_{j \in B_k} (\alpha_{jk} e^{V_{nj}})^{1/\lambda_k} \right)^{\lambda_k - 1}}{\sum_{\ell=1}^{K} \left(\sum_{j \in B_\ell} (\alpha_{j\ell} e^{V_{nj}})^{1/\lambda_\ell} \right)^{\lambda_\ell}}. \tag{4.7}$$

This formula is similar to the nested logit probability given in equation 4.2, except that the numerator is a sum over all the nests that contains alternative i, with weights applied to these nests. If each alternative enters only one nest, with $\alpha_{jk} = 1$ for $j \in B_k$ and zero otherwise, the model becomes a nested logit model. And if, in addition, $\lambda_k = 1$ for all nests, then the model becomes standard logit. Wen and Koppelman (2001) derive various cross-nested models as special cases of the GNL.

To facilitate interpretation, the GNL probability can be decomposed as:

$$P_{ni} = \sum_k P_{ni|B_k} P_{nk},$$

where the probability of nest k is

$$P_{nk} = \frac{\sum_j (\alpha_{jk} e^{V_{nj}})^{1/\lambda_k}}{\sum_{\ell=1}^{K} \left(\sum_{j \in B_\ell} (\alpha_{j\ell} e^{V_{nj}})^{1/\lambda_\ell} \right)^{\lambda_\ell}}$$

and the probability of alternative i given nest k is

$$P_{ni|B_k} = \frac{(\alpha_{ik} e^{V_{ni}})^{1/\lambda_k}}{\sum_j (\alpha_{jk} e^{V_{nj}})^{1/\lambda_k}}.$$

4.5 Heteroskedastic Logit

Instead of capturing correlations among alternatives, the researcher may simply want to allow the variance of unobserved factors to differ over alternatives. Steckel and Vanhonacker (1988), Bhat (1995), and Recker (1995) describe a type of GEV model, called "heteroskedastic extreme value" (HEV), that is the same as logit except with different variance for each alternative. Utility is specified as $U_{nj} = V_{nj} + \varepsilon_{nj}$ where ε_{nj} is distributed independently extreme value with variance $(\theta_j \pi)^2/6$. There is no correlation in unobserved factors over alternatives; however, the variance of the unobserved factors is different for different alternatives. To set the overall scale of utility, the variance for one alternative is normalized to $\pi^2/6$, which is the variance of the standardized extreme value distribution. The variances for the other alternatives are then estimated relative to the normalized variance.

The choice probabilities for this heteroskedastic logit are ((Bhat, 1995)):

$$P_{ni} = \int \left[\prod_{j \neq i} e^{-e^{-(V_{ni} - V_{nj} + \theta_i w)/\theta_j}} \right] e^{-e^{-w}} e^{-w} dw$$

where $w = \varepsilon_{ni}/\theta_i$. The integral does not take a closed form; however, it can be approximated by simulation. Note that $exp(-exp(-w))exp(-w)$ is the extreme value density, given in section 3.1. P_{ni} is therefore the integral of the term in square brackets over the extreme value density. It can be simulated as follows. (1) Take a draw from the extreme value distribution, using the procedure described in section 9.2.3. (2) For this draw of w, calculate the term in brackets, namely: $\prod_{j \neq i} exp(-exp(-(v_{ni} - V_{nj} + \theta_i w)/\theta_j))$. (3) Repeat steps 1 and 2 many times and average the results. This average is an approximation to P_{ni}. Bhat (1995) shows that, since the integral is only one-dimensional, the heteroskedastic logit probabilities can be calculated effectively with quadrature rather than simulation.

4.6 The GEV family

We now describe the processs that McFadden (1978) developed to generate GEV models. Using this process, the researcher is able to develop new GEV models that best fit the specific circumstances of his choice situation. As illustration, we show how the procedure is used to generate models that we have already discussed, namely logit, nested logit, and paired combinatorial logits. The same procedure can be applied by a researcher to generate new models with properties that meet his research needs.

For notational simplicity, we will omit the subscript n denoting the decision-maker. Also, since we will be using $exp(V_j)$ repeatedly, let's denote it more compactly by Y_j. That is, let $Y_j \equiv exp(V_j)$. Note that Y_j is necessarily positive.

Consider a function G that depends on Y_j for all j. We denote this function $G = G(Y_1, \ldots, Y_J)$. Let G_i be the derivative of G with respect to Y_i: $G_i = \partial G / \partial Y_i$. If this function meets certain conditions, then a discrete choice model can be based upon it. In particular, if G satisfies the conditions that are listed below, then

$$P_i = \frac{Y_i G_i}{G} \tag{4.8}$$

is the choice probability for a discrete choice model that is consistent with utility maximization. Any model that can be derived in this way is a GEV model. This formula therefore defines the family of GEV models.

The properties that the function must exhibit are the following.

1. $G \geq 0$ for all positive values of Y_j $\forall j$.

2. G is homogeneous of degree one. That is, if each Y_j is raised by some proportion ρ, G rises by proportion ρ also: $G(\rho Y_1, \ldots, \rho Y_J) = \rho G(Y_1, \ldots, Y_J)$. Actually, Ben-Akiva and Francois (1983) showed that this condition can be relaxed to allow any degree of homogeneity. We retain the usage of degree one since doing makes the condition easier to interpret and is consistent with McFadden's original description.

3. $G \rightarrow \infty$ as $Y_j \rightarrow \infty$ for any j.

4. The cross partial derivatives of G change signs in a particular way. That is: $G_i \geq 0$ for all i, $G_{ij} = \partial G_i / \partial Y_j \leq 0$ for all $j \neq i$,

$G_{ijk} = \partial G_{ij}/\partial Y_k \geq 0$ for any distinct i, j, and k, and so on for higher order cross-partials.

There is little economic intuition to motivate these properties, particularly the last one. However, it is easy to verify whether a function exhibits these properties. The lack of intuition behind the properties is a blessing and a curse. The disadvantage is that the researcher has little guidance on how to specify a G that provides a model that meets the needs of his research. The advantage is that the purely mathematical approach allows the researcher to generate models that he might not have developed while relying only on his economic intuition. Karlstrom (2001) provides an example: he arbitrarily specified a G (in the sense that it was not based on behavioral concepts) and found that the resulting probability formula fit his data better than logit, nested logit, and PCL.

We can now show how logit, nested logit, and PCL models are obtained under appropriate specifications of G.

Logit

Let $G = \sum_{j=1}^{J} Y_j$. This G exhibits the four required properties. (1) The sum of positive Y_j's is positive. (2) If all Y_j's are raised by a factor ρ, G rises by that same factor. (3) If any Y_j rises without bound, then G does also. (4) The first partial derivative is $G_i = \partial G/\partial Y_i = 1$ which meets the criterion that $G_i \geq 0$. And the higher-order derivatives are all zero, which clearly meets the criterion since they are ≥ 0 or ≤ 0 as required.

Inserting this G and its first derivative G_i into (4.8), the resulting choice probability is:

$$\begin{aligned} P_i &= \frac{Y_i G_i}{G} \\ &= \frac{Y_i}{\sum_{j=1}^{J} Y_j} \\ &= \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}} \end{aligned}$$

which is the logit formula.

Nested Logit

The J alternatives are partitioned into K nests labeled $B_1, \ldots, B_K$. Let

$$G = \sum_{\ell=1}^{K} \left(\sum_{j \in B_\ell} Y_j^{(1/\lambda_\ell)} \right)^{\lambda_\ell}.$$

with each λ_k between zero and one. The first three properties are easy to verify. For the fourth property, we calculate the first partial derivative

$$\begin{aligned} G_i &= \lambda^k \left(\sum_{j \in B_k} Y_j^{(1/\lambda_k)} \right)^{\lambda_k - 1} \frac{1}{\lambda_k} Y_i^{(1/\lambda_k)-1} \\ &= Y_i^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{(1/\lambda_k)} \right)^{\lambda_k - 1}. \end{aligned}$$

for $i \in B_k$. Since $Y_j \geq 0 \; \forall j$, $G_i \geq 0$, as required. The second cross-partial derivative is

$$\begin{aligned} G_{im} &= \frac{\partial G_i}{\partial Y_m} \\ &= (\lambda_k - 1) Y_i^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{(1/\lambda_k)} \right)^{\lambda_k - 2} \frac{1}{\lambda_k} Y_m^{(1/\lambda_k)-1} \\ &= \left(\frac{\lambda_k - 1}{\lambda_k} \right) (Y_i Y_m)^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{(1/\lambda_k)} \right)^{\lambda_k - 2} \end{aligned}$$

for $m \in B_k$ and $m \neq i$. With $\lambda_k \leq 1$, $G_{ij} \leq 0$, as required. For j in a different nest than i, $G_{ij} = 0$ which also meets the criterion. Higher cross-partials are calculated similarly; they exhibit the required property if $0 < \lambda_k \leq 1$.

The choice probability becomes:

$$\begin{aligned} P_i &= \frac{Y_i G_i}{G} \\ &= \frac{Y_i Y_i^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{(1/\lambda_\ell)} \right)^{\lambda_k - 1}}{\sum_{\ell=1}^{K} \left(\sum_{j \in B_\ell} Y_j^{(1/\lambda_\ell)} \right)^{\lambda_\ell}} \end{aligned}$$

$$\begin{aligned}
&= \frac{Y_i^{(1/\lambda_k)}\left(\sum_{j\in B_k} Y_j^{(1/\lambda_\ell)}\right)^{\lambda_k-1}}{\sum_{\ell=1}^K\left(\sum_{j\in B_\ell} Y_j^{(1/\lambda_\ell)}\right)^{\lambda_\ell}} \\
&= \frac{(e^{V_i})^{(1/\lambda_k)}\left(\sum_{j\in B_k} (e^{V_j})^{(1/\lambda_\ell)}\right)^{\lambda_k-1}}{\sum_{\ell=1}^K\left(\sum_{j\in B_\ell} (e^{V_j})^{(1/\lambda_\ell)}\right)^{\lambda_\ell}} \\
&= \frac{e^{V_i/\lambda_k}\left(\sum_{j\in B_k} e^{V_j/\lambda_\ell}\right)^{\lambda_k-1}}{\sum_{\ell=1}^K\left(\sum_{j\in B_\ell} e^{V_j/\lambda_\ell}\right)^{\lambda_\ell}}
\end{aligned}$$

which is the nested logit formula (4.2).

Paired combinatorial logit

Let

$$G = \sum_{k=1}^{J-1}\sum_{\ell=k+1}^{J}\left(Y_k^{(1/\lambda_{k\ell})} + Y_\ell^{(1/\lambda_{k\ell})}\right)^{\lambda_{k\ell}}.$$

The required properties are verified in the same way as for the nested logit. We have

$$\begin{aligned}
G_i &= \sum_{j\neq i}\lambda_{ji}\left(Y_i^{(1/\lambda_{ij})} + Y_j^{(1/\lambda_{ij})}\right)^{\lambda_{ij}-1}\frac{1}{\lambda_{ij}}Y_i^{(1/\lambda_{ij})-1} \\
&= \sum_{j\neq i} Y_i^{(1/\lambda_{ij})-1}\left(Y_i^{(1/\lambda_{ij})} + Y_j^{(1/\lambda_{ij})}\right)^{\lambda_{ij}-1}.
\end{aligned}$$

And so the choice probability is:

$$\begin{aligned}
P_i &= \frac{Y_i G_i}{G} \\
&= \frac{Y_i\sum_{j\neq i} Y_i^{(1/\lambda_{ij})-1}\left(Y_i^{(1/\lambda_{ij})} + Y_j^{(1/\lambda_{ij})}\right)^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{\ell=k+1}^{J}\left(Y_k^{(1/\lambda_{k\ell})} + Y_\ell^{(1/\lambda_{k\ell})}\right)^{\lambda_{k\ell}}} \\
&= \frac{\sum_{j\neq i} Y_i^{(1/\lambda_{ij})}\left(Y_i^{(1/\lambda_{ij})} + Y_j^{(1/\lambda_{ij})}\right)^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{\ell=k+1}^{J}\left(Y_k^{(1/\lambda_{k\ell})} + Y_\ell^{(1/\lambda_{k\ell})}\right)^{\lambda_{k\ell}}} \\
&= \frac{\sum_{j\neq i} e^{V_i/\lambda_{ij}}\left(e^{V_i/\lambda_{ij}} + e^{V_j/\lambda_{ij}}\right)^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{\ell=k+1}^{J}\left(e^{V_k/\lambda_{k\ell}} + e^{V_\ell/\lambda_{k\ell}}\right)^{\lambda_{k\ell}}}
\end{aligned}$$

which is the PCL formula (4.6).

Generalized nest logit

The reader can verify that the GNL probabilities in equation 4.7 are derived from

$$G = \sum_{k=1}^{K} \left(\sum_{j \in B_k} (\alpha_{jk} Y_j)^{(1/\lambda_k)} \right)^{\lambda_k} .$$

Using the same process, researchers can generate other GEV models.

Chapter 5

Probit

5.1 Choice probabilities

The logit model is limited in three important ways. It cannot represent random taste variation. It exhibits restrictive substitution patterns due to the IIA property. And it cannot be used with panel data when unobserved factors are correlated over time for each decision-maker. GEV models relax the second of these restrictions, but not the other two. Probit models solve all three issues. They can handle random taste variation, they allow any pattern of substitution, and they are applicable to panel data with temporally correlated errors.

The only limitation of probit models is that they require normal distributions for all unobserved components of utility. In many, perhaps most situations, normal distributions provide an adequate representation of the random components. However, in some situations, normal distributions are inappropriate and can lead to perverse forecasts. A prominent example relates to price coefficients. For a probit model with random taste variation, the coefficient of price is assumed to be normally distributed in the population. Since the normal distribution has density on both sides of zero, the model necessarily implies that some people have a positive price coefficient. The use of a distribution that has density only on one side of zero, such as the log-normal, is more appropriate and yet cannot be accommodated within probit. Other than this restriction, the probit model is quite general.

The probit model is derived under the assumption of jointly normal unobserved utility components. The first derivation, by Thurstone (1927) for a binary probit, used the terminology of psychological stimuli, which Marschak (1960) translated into economic terms as utility. Hausman and Wise (1978) and Daganzo (1979) elucidated the generality of the specification for representing various aspects of choice behavior. Utility is decomposed into observed and unobserved parts: $U_{nj} = V_{nj} + \varepsilon_{nj}\ \forall j$. Consider the vector composed of each ε_{nj}, labeled $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$. We assume that ε_n is distributed normal with a mean vector of zero and covariance matrix Ω. The density of ε_n is

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{\frac{J}{2}}|\Omega|^{\frac{1}{2}}} e^{-\frac{1}{2}\varepsilon_n'\Omega^{-1}\varepsilon_n}.$$

The covariance Ω can depend on variables faced by decision-maker n, such that Ω_n is the more appropriate notation; however, we omit the subscript for the sake of simplicity.

The choice probability is

$$\begin{aligned} P_{ni} &= Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i) \\ &= \int I(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i)\phi(\varepsilon_n)d\varepsilon_n \qquad (5.1) \end{aligned}$$

where $I(\cdot)$ is an indicator of whether the statement in parentheses holds and the integral is over all values of ε_n. This integral does not have a closed form. It must be evaluated numerically through simulation.

The choice probabilities can be expressed in a couple of other ways that are useful for simulating the integral. Let B_{ni} be the set of error terms ε_n that result in the decision-maker choosing alternative i: $B_{ni} = \{\varepsilon_n \text{ s.t. } V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i\}$. Then

$$P_{ni} = \int_{\varepsilon_n \in B_{ni}} \phi(\varepsilon_n)d\varepsilon_n \qquad (5.2)$$

which is an integral over only some of the values of ε_n rather than all possible values, namely, the ε_n's in B_{ni}.

Expressions (5.1) and (5.2) are J dimensional integrals over the J errors ε_{nj}, $j = 1, \ldots, J$. Since only differences in utility matter, the choice probabilities can be equivalently expressed as $J-1$ dimensional integrals over the differences between the errors. Let us difference against alternative i, the alternative for which we are calculating the probability. Define $\tilde{U}_{nji} = U_{nj} - U_{ni}$, $\tilde{V}_{nji} = V_{nj} - V_{ni}$, and $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$. Then $P_{ni} = Prob(\tilde{U}_{nji} < 0\ \forall j \neq i)$. That is, the probability of choosing alternative i is the probability that all the utility differences, when differenced against i, are negative. Define the vector $\tilde{\varepsilon}_{ni} = \langle \tilde{\varepsilon}_{n1i}, \ldots, \tilde{\varepsilon}_{nJ1} \rangle$ where the "..." is over all alternatives except i, such that $\tilde{\varepsilon}_{ni}$ has dimension $J-1$. Since the difference between two normals is normal, the density of the error differences is

$$\phi(\tilde{\varepsilon}_{ni}) = \frac{1}{(2\pi)^{-\frac{1}{2}(J-1)}|\tilde{\Omega}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}\tilde{\varepsilon}'_{ni}\tilde{\Omega}_i\tilde{\varepsilon}_{ni}}$$

where $\tilde{\Omega}_i$ is the covariance of $\tilde{\varepsilon}_{ni}$, derived from Ω. Then the choice probability expressed in utility differences is:

$$P_{ni} = \int I(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i)\phi(\tilde{\varepsilon}_{ni})d\tilde{\varepsilon}_{ni} \tag{5.3}$$

which is a $J-1$ dimensional integral over all possible values of the error differences. An equivalent expression is:

$$P_{ni} = \int_{\tilde{\varepsilon}_{ni} \in \tilde{B}_{ni}} \phi(\tilde{\varepsilon}_{ni})d\tilde{\varepsilon}_{ni} \tag{5.4}$$

where $\tilde{B}_{ni} = \{\tilde{\varepsilon}_{ni} \text{ s.t. } \tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i\}$, which is a $J-1$ dimensional integral over the error differences in $\tilde{B}_{ni}$.

Expressions (5.3) and (5.4) utilize the covariance matrix $\tilde{\Omega}_i$ of the error differences. There is a straightforward way to derive $\tilde{\Omega}_i$ from the covariance of the errors themselves, Ω. Let M_i be the $J-1$ identity matrix with an extra column of -1's added as the i-th column. The extra column makes the matrix have size $J-1$ by J. For example, with $J=4$ alternatives and $i=3$, M_i is

$$M_i = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

This matrix can be used to transform the covariance matrix of errors into the covariance matrix of error differences: $\tilde{\Omega}_i = M_i \Omega M_i'$. Note that $\tilde{\Omega}_i$ is $(J-1) \times (J-1)$ while Ω is $J \times J$, since M_i is $(J-1) \times J$. As illustration, consider a three-alternative situation with errors $\langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle$ that have covariance

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}.$$

Suppose we takes differences against alternative 2. We know from first principles that the error differences $\langle \tilde{\varepsilon}_{n12}, \tilde{\varepsilon}_{n32} \rangle$ have covariance

$$\tilde{\Omega}_2 = Cov \begin{pmatrix} \varepsilon_{n1} - \varepsilon_{n2} \\ \varepsilon_{n3} - \varepsilon_{n2} \end{pmatrix} = \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}.$$

This covariance matrix can also be derived by the transformation $\tilde{\Omega}_2 = M_2 \Omega M_2'$:

$$\begin{aligned} \tilde{\Omega}_n &= \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} & -\sigma_{22} + \sigma_{23} & -\sigma_{23} + \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} - \sigma_{12} - \sigma_{12} + \sigma_{22} & -\sigma_{12} + \sigma_{22} + \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} + \sigma_{22} - \sigma_{23} & \sigma_{22} - \sigma_{23} - \sigma_{23} + \sigma_{33} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix} \end{aligned}$$

As we will see, this transformation by M_i comes in handy when simulating probit probabilities.

5.2 Identification

As described in section (2.5), any discrete choice model must be normalized to account for the fact that the level and scale of utility is irrelevant. The level of utility is immaterial since a constant can be added to the utility of all alternatives without changing which alternative has the highest utility: the alternative with the highest utility before the constant is added still has the highest utility after the constant is added to all utilities. Similarly, the scale of utility doesn't matter since the utility of each alternative can be multiplied by a (positive) constant without changing which alternative has the highest utility. In logit and nested logit models, the normalization for scale and level occurs automatically with the distributional assumptions that are placed on the error terms. As a result, normalization does not need to be considered explicitly for these models. With probit models, however, normalization for scale and level does not occur automatically. The researcher must normalize the model directly.

Normalization of the model is related to parameter identification. A parameter is "identified" if it can be estimated and is "unidentified" if it cannot be estimated. An example of an unidentified parameter is k in the utility specification $U_{nj} = V_{nj} + k + \varepsilon_{nj}$. While the researcher might write utility in this way, and might want to estimate k to obtain a measure of the overall level of utility, doing so is impossible. The behavior of the decision-maker is unaffected by k, and so the researcher cannot infer the value of k from the choices that decision-makers have made. Stated directly: parameters that do not affect the behavior of decision-makers cannot be estimated. In an unnormalized model, parameters can appear that are not identified; these parameters relate to the scale and level of utility, which do not affect behavior. Once the model is normalized, these parameters disappear. The difficulty arises because it is not always obvious which parameters relate to scale and level. In the above example, the fact that k is unidentified is fairly obvious. In many cases, it is not at all obvious which parameters are identified. Bunch and Kitamura (1989) have shown that the probit models in several published articles are not normalized and contain unidentified parameters. The fact that neither the authors nor the reviewers of these articles could tell that the models were un-normalized is a testament to the complexity of the issue.

I provide below a procedure that can always be used to normalize a probit model and assure that all parameters are identified. It is not the only procedure that can be used; see, for example, Bunch (1991). In some cases a researcher might find other normalization procedures more convenient. However, the procedure I give can always be used, either by itself or as a check on whatever other procedure the researcher uses for normalization.

I describe the procedure in terms of a four-alternative model. Generalization to more alternatives is obvious. As usual, utility is expressed as $U_{nj} = V_{nj} + \varepsilon_{nj}$ $j = 1, \ldots, 4$. The vector of errors is $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$. The error vector is normally distributed with