zero mean and a covariance matrix that can be expressed explicitly as:

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{pmatrix} \tag{5.5}$$

where the dots refer to the corresponding elements on the upper part of the matrix. Note that there are 10 elements in this matrix, that is, 10 distinct $\sigma$'s representing the variances and covariances among the 4 errors. In general, a model with $J$ alternatives has $J(J+1)/2$ distinct elements in the covariance matrix of the errors.

To account for the fact that the level of utility is irrelevant, we take utility differences. In my procedure, I always take differences with respect to the first alternative, since that simplifies the analysis in a way that we will see. Define error differences as $\tilde{\varepsilon}_{nj1} = \varepsilon_{nj} - \varepsilon_{n1}$ for $j = 2, 3, 4$, and define the vector of error differences as $\tilde{\varepsilon}_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}, \tilde{\varepsilon}_{n41} \rangle$. Note that the subscript 1 in $\tilde{\varepsilon}_{n1}$ means that the error differences are against the first alternative, rather than that the errors are for the first alternative. The covariance matrix for the vector of error differences takes the form

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \cdot & \theta_{33} & \theta_{34} \\ \cdot & \cdot & \theta_{44} \end{pmatrix}$$

where the $\theta$'s relate to the original $\sigma$'s as follows:

$$\begin{aligned} \theta_{22} &= \sigma_{22} + \sigma_{11} - 2\sigma_{12} \\ \theta_{33} &= \sigma_{33} + \sigma_{11} - 2\sigma_{13} \\ \theta_{44} &= \sigma_{44} + \sigma_{11} - 2\sigma_{14} \\ \theta_{23} &= \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13} \\ \theta_{24} &= \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14} \\ \theta_{34} &= \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}. \end{aligned}$$

Computationally, this matrix can be obtained using the transformation matrix $M_i$ defined in section (5.1), as $\tilde{\Omega}_1 = M_1 \Omega M_1'$.

To set the scale of utility, one of the diagonal elements is normalized. I set the top-left element of $\tilde{\Omega}_1$, which is the variance of $\tilde{\varepsilon}_{n21}$, to 1.

This normalization for scale gives us the following covariance matrix:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* \\ \cdot & \cdot & \theta_{44}^*. \end{pmatrix} \tag{5.6}$$

The $\theta^*$'s relate to the original $\sigma$'s as follows:

$$\begin{aligned}
\theta_{33}^* &= \frac{\sigma_{33} + \sigma_{11} - 2\sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\theta_{44}^* &= \frac{\sigma_{44} + \sigma_{11} - 2\sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\theta_{23}^* &= \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\theta_{24}^* &= \frac{\sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}} \\
\theta_{34}^* &= \frac{\sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}}.
\end{aligned}$$

There are 5 elements in $\tilde{\Omega}_1^*$. These are the only identified parameters in the model. This number is less than the 10 elements that enter $\Omega$. Each $\theta^*$ is a function of the $\sigma$'s. Since there are 5 $\theta^*$'s and 10 $\sigma$'s, it is not possible to solve for all the $\sigma$'s from estimated values of the $\theta^*$'s. It is therefore not possible to obtain estimates of all the $\sigma$'s.

In general, a model with $J$ alternatives and an unrestricted covariance matrix will have $[(J-1)J/2]-1$ covariance parameters when normalized, compared to the $J(J+1)/2$ parameters when un-normalized. Only $[(J-1)J/2]-1$ parameters are identified. This reduction in the number of parameters is *not* a restriction. The reduction in the number of parameters is a normalization that simply eliminates irrelevant aspects of the original covariance matrix, namely the scale and level of utility. The 10 elements in $\Omega$ allow for variance and covariance that is due simply to scale and level, which has no relevance for behavior. Only the 5 elements in $\tilde{\Omega}_1^*$ contain information about the variance and covariance of errors independent of scale and level. In this sense, only the 5 parameters have economic content, and only the 5 parameters can be estimated.

Suppose now that the researcher places structure on the covariance matrix. That is, instead of allowing a full covariance matrix for the errors, the researcher believes that the errors follow a pattern that implies particular values for, or relations among, the elements in the

covariance matrix. The researcher restricts the covariance matrix to incorporate this pattern. Imposing these restrictions is called "placing structure on the covariance matrix."

The structure can take various forms, depending on the application. Yai, Iwakura and Morichi (1997) estimate a probit model of route choice where the covariance between any two routes depends only on the length of shared route segments; this structure reduces the number of covariance parameters to only one, which captures the relation of the covariance to shared length. Bolduc, Fortin and Fournier (1996) estimate a model of physicians' choice of location where the covariance among locations is a function of their proximity to one another, using what Bolduc (1992) has called a "generalized autoregressive" structure. Haaijer *et al.* (1998) place a factor-analytic structure that arises from random coefficients of explanatory variables; this type of structure is described in detail in section 5.3 below. Elrod and Keane (1995) place a factor-analytic structure, but one that arises from error components rather than random coefficients *per se*.

Often the structure that is placed will be sufficient to normalize the model. That is, the restrictions that the researcher imposes on the covariance matrix to fit her beliefs about the way the errors relate to each other will also serve to normalize the model. However, this is not always the case. The examples cited by Bunch and Kitamura (1989) are cases where the restrictions that the researcher placed on the covariance matrix seemed sufficient to normalize the model but actually were not.

The procedure that I give above can be used to determine whether the restrictions that a researcher places on the covariance matrix are sufficient to normalize the model. The researcher specifies $\Omega$ with her restrictions on its elements. Then the procedure above is used to derive $\tilde{\Omega}_1^*$, which is normalized for scale and level. We know that each element of $\tilde{\Omega}_1^*$ is identified. If each of the restricted elements of $\Omega$ can be calculated from the elements of $\tilde{\Omega}_1^*$, then the restrictions are sufficient to normalize the model. In this case, each parameter in the restricted $\Omega$ is identified. On the other hand, if the elements of $\Omega$ cannot be calculated from the elements of $\tilde{\Omega}_1^*$, then the restrictions are not sufficient to normalize the model and the parameters in $\Omega$ are not identified.

To illustrate this approach, suppose the researcher is estimating a four-alternative model and assumes that the covariance matrix for the

errors has the following form:

$$\Omega = \begin{pmatrix} 1+\rho & \rho & 0 & 0 \\ \cdot & 1+\rho & 0 & 0 \\ \cdot & \cdot & 1+\rho & \rho \\ \cdot & \cdot & \cdot & 1+\rho \end{pmatrix}.$$

This covariance matrix allows the first and second errors to be correlated, the same as the third and fourth alternatives, but allows no other correlation. The correlation between the appropriate pairs is $\rho/(1+\rho)$. Note that by specifying the diagonal elements as $1+\rho$, the researcher assures that the correlation is between -1 and 1 for any value of $\rho$, as required for a correlation. Is this model, as specified, normalized for scale and level? To answer the question, we apply the procedure described above. First, we take differences with respect to the first alternative. The covariance matrix of error differences is:

$$\tilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \cdot & \theta_{33} & \theta_{34} \\ \cdot & \cdot & \theta_{44} \end{pmatrix}$$

where the $\theta$'s relate to the original $\sigma$'s as follows:

$$\begin{aligned} \theta_{22} &= 2 \\ \theta_{33} &= 2+2\rho \\ \theta_{44} &= 2+2\rho \\ \theta_{23} &= 1 \\ \theta_{24} &= 1 \\ \theta_{34} &= 1+2\rho. \end{aligned}$$

We then normalize for scale by setting the top-left element to 1. The normalized covariance matrix is

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* \\ \cdot & \cdot & \theta_{44}^* \end{pmatrix}$$

where the $\theta^*$'s relate to the original $\sigma$'s as follows:

$$\begin{aligned} \theta_{33}^* &= 1+\rho \\ \theta_{44}^* &= 1+\rho \end{aligned}$$

$$\begin{aligned}\theta^*_{23} &= 1/2\\ \theta^*_{24} &= 1/2\\ \theta^*_{34} &= 1/2+\rho.\end{aligned}$$

Note that $\theta^*_{33} = \theta^*_{44} = \theta^*_{34} - 1/2$ and that the other $\theta^*$'s have fixed values. There is one parameter in $\tilde{\Omega}^*_1$, as there is in $\Omega$. Define $\theta$=$1+\rho$. Then $\tilde{\Omega}^*_1$ is

$$\tilde{\Omega}^*_1 = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \cdot & \theta & \theta - \frac{1}{2} \\ \cdot & \cdot & \theta \end{pmatrix}.$$

The original $\rho$ can be calculated directly from $\theta$. For example, if $\theta$ is estimated to be 2.4, then the estimate of $\rho$ is $\theta - 1 = 1.4$ and the correlation is $1.4/2.4 = .58$. The fact that the parameters that enter $\Omega$ can be calculated from the parameters that enter the normalized covariance matrix $\tilde{\Omega}^*_1$ means that the original model is normalized for scale and level. That is, the restrictions that the researcher placed on $\Omega$ also provided the needed normalization.

Sometimes restrictions on the original covariance matrix can appear to be sufficient to normalize the model when in fact they do not. Applying our procedure will determine whether this is the case. Consider the same model as above, but now suppose that the researcher allows a different correlation between the first and second errors than between the third and fourth errors. The covariance matrix of errors is specified to be:

$$\Omega = \begin{pmatrix} 1+\rho_1 & \rho_1 & 0 & 0 \\ \cdot & 1+\rho_1 & 0 & 0 \\ \cdot & \cdot & 1+\rho_2 & \rho_2 \\ \cdot & \cdot & \cdot & 1+\rho_2 \end{pmatrix}.$$

The correlation between the first and second errors is $\rho_1/(1+\rho_1)$ and the correlation between the third and fourth errors is $\rho_2/(1+\rho_2)$. We can derive $\tilde{\Omega}_1$ for error differences and then derive $\tilde{\Omega}^*_1$ by setting the top-left element of $\tilde{\Omega}_1$ to 1. The resulting matrix is:

$$\tilde{\Omega}^*_1 = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \cdot & \theta & \theta - \frac{1}{2} \\ \cdot & \cdot & \theta \end{pmatrix}$$

where now $\theta = 1+(\rho_1+\rho_2)/2$. The values of $\rho_1$ and $\rho_2$ cannot be calculated from a value of $\theta$. The original model is therefore not normalized

for scale and level, and the parameters $\rho_1$ and $\rho_2$ are not identified. This fact is somewhat surprising, since only two parameters enter the original covariance matrix $\Omega$. It would seem, unless the researcher explicitly tested in the manner we have just done, that restricting the covariance matrix to consist of only two elements would be sufficient to normalize the model. In this case, however, it is not.

In the normalized model, only the average of the $\rho$'s appears: $(\rho_1 + \rho_2)/2$. It is possible to calculate the average $\rho$ from $\theta$, simply as $\theta - 1$. This means that the average $\rho$ is identified, but not the individual values. When $\rho_1 = \rho_2$, as in the previous example, the model is normalized because each $\rho$ is equal to the average $\rho$. However, as we now see, any model with the same average $\rho$'s is equivalent, after normalizing for scale and level. Hence, assuming that $\rho_1 = \rho_2$ is no different than assuming that $\rho_1 = 3\rho_2$, or any other relation. All that matters for behavior is the average of these parameters, not their values relative to each other. This fact is fairly surprising and would be hard to realize without using our procedure for normalization.

Now that we know how to assure that a probit model is normalized for level and scale, and hence contains only economically meaningful information, we can examine how the probit model is used to represent various types of choice situations. We look at the three issues for which logit models are limited and show how the limitation is overcome with probit. These issues are: taste variation, substitution patterns, and repeated choices over time.

## 5.3 Taste variation

Probit is particularly well-suited for incorporating random coefficients, provided that the coefficients are normally distributed. Hausman and Wise (1978) were the first, to my knowledge, to give this derivation. Haaijer, Wedel, Vriens and Wansbeek (1998) provide a compelling application. Assume that representative utility is linear in parameters and that the coefficients vary randomly over decision-makers instead of being fixed as we have assumed so far in this book. Utility is: $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$ where $\beta_n$ is the vector of coefficients for decision-maker $n$ representing that person's tastes. Suppose the $\beta_n$ is normally distributed in the population with mean $b$ and covariance $W$: $\beta_n \sim N(b, W)$. The goal of the research is to estimate parameters $b$ and $W$.

Utility can be rewritten with $\beta_n$ decomposed into its mean and deviations from its mean: $U_{nj} = b'x_{nj} + \tilde{\beta}_n' x_{nj} + \varepsilon_{nj}$, where $\tilde{\beta}_n = b - \beta_n$. The last two terms in utility are random; denote their sum as $\eta_{nj}$ to obtain $U_{nj} = b'x_{nj} + \eta_{nj}$. The covariance of the $\eta_{nj}$'s depends on $W$ as well as the $x_{nj}$'s, such that the covariance differs over decision-makers.

The covariance of the $\eta_{nj}$'s can be described easily for a two-alternative model with one explanatory variable. In this case, utility is

$$\begin{aligned} U_{n1} &= \beta_n x_{n1} + \varepsilon_{n1} \\ U_{n2} &= \beta_n x_{n2} + \varepsilon_{n2}. \end{aligned}$$

Assume that $\beta_n$ is normally distributed with mean $b$ and variance $\sigma_\beta$. Assume that $\varepsilon_{n1}$ and $\varepsilon_{n2}$ are identically normally distributed with variance $\sigma_\varepsilon$. The assumption of independence is for this example and is not needed in general. Utility is then rewritten as

$$\begin{aligned} U_{n1} &= b x_{n1} + \eta_{n1} \\ U_{n2} &= b x_{n2} + \eta_{n2}. \end{aligned}$$

where $\eta_{n1}$ and $\eta_{n2}$ are jointly normally distributed. Each has zero mean: $E(\eta_{nj}) = E(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = 0$. The covariance is determined as follows. The variance of each is $V(\eta_{nj}) = V(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = x_{nj}^2 \sigma_\beta + \sigma_\varepsilon$. Their covariance is

$$\begin{aligned} Cov(\eta_{n1}, \eta_{n2}) &= E[(\tilde{\beta}_n x_{n1} + \varepsilon_{n1})(\tilde{\beta}_n x_{n2} + \varepsilon_{n2})] \\ &= E(\tilde{\beta}_n^2 x_{n1} x_{n2} + \varepsilon_{n1}\varepsilon_{n2} + \varepsilon_{n1}\tilde{\beta}_b x_{n2} + \varepsilon_{n2}\tilde{\beta}_n x_{n1}) \\ &= x_{n1} x_{n2} \sigma_\beta. \end{aligned}$$

The covariance matrix is

$$\begin{aligned} \Omega &= \begin{pmatrix} x_{n1}^2\sigma_\beta + \sigma_\varepsilon & x_{n1}x_{n2}\sigma_\beta \\ x_{n1}x_{n2}\sigma_\beta & x_{n2}^2\sigma_\beta + \sigma_\varepsilon \end{pmatrix} \\ &= \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1}x_{n2} \\ x_{n1}x_{n2} & x_{n2}^2 \end{pmatrix} + \sigma_\varepsilon \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

One last step is required for estimation. Recall that behavior is not affected by a multiplicative transformation of utility. We therefore need to set the scale of utility. A convenient normalization for this case is $\sigma_\varepsilon = 1$. Under this normalization,

$$\Omega = \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1}x_{n2} \\ x_{n1}x_{n2} & x_{n2}^2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The values of $x_{n1}$ and $x_{n2}$ are observed by the researcher, and the parameters $b$ and $\sigma_\beta$ are estimated. Thus, the researcher learns both the mean and variance of the random coefficient in the population. Generalization to more than one explanatory variable and more than two alternatives is straightforward.

## 5.4 Substitution patterns/non-IIA

Probit can represent any substitution pattern. The probit probabilities do not exhibit the IIA property that gives rise to the proportional substitution of logit. Different covariance matrices $\Omega$ provide different substitution patterns, and by estimating the covariance matrix, the researcher determines the substitution pattern that is most appropriate for the data.

A full covariance matrix can be estimated, or the researcher can place structure on the covariance matrix to represent particular sources of non-independence. This structure usually reduces the number of parameters and facilitates interpretation of the parameters. We consider first the situation where the researcher estimates a full covariance matrix, and then turn to a situation where the researcher places structure on the covariance matrix.

### Full covariance: unrestricted substitution patterns

For notational simplicity, consider a probit model with four alternatives. A full covariance matrix for the unobserved components of utility takes the form of $\Omega$ in (5.5). When normalized for scale and level, the covariance matrix becomes $\tilde{\Omega}_1^*$ in (5.6). The elements of $\tilde{\Omega}_1^*$ are estimated. The estimated values can represent any substitution pattern; importantly, the normalization for scale and level does not restrict the substitution patterns. The normalization only eliminates aspects of $\Omega$ that are irrelevant to behavior.

Note, however, that the estimated values of the $\theta^*$'s provide essentially no interpretable information in themselves (Horowitz, 1991). For example, suppose $\theta_{33}^*$ is estimated to be larger than $\theta_{44}^*$. It might be tempting to interpret this result as indicating that the variance in unobserved utility of the third alternative is greater than that for the fourth alternative; that is, that $\sigma_{33} > \sigma_{44}$. However, this interpretation is incorrect. It is quite possible that $\theta_{33}^* > \theta_{44}^*$ and yet $\sigma_{44} > \sigma_{33}$, if

covariance $\sigma_{13}$ is sufficiently greater than $\sigma_{14}$. Similarly, suppose that $\theta_{23}$ is estimated to be negative. This does not mean that unobserved utility for the second alternative is negatively correlated with unobserved utility for the third alternative (that is, $\sigma_{23} < 0$). It is possible that $\sigma_{23}$ is positive and yet $\sigma_{12}$ and $\sigma_{13}$ are sufficiently large to make $\theta^*_{23}$ negative. The point here is: estimating a full covariance matrix allows the model to represent any substitution pattern but renders the estimated parameters essentially uninterpretable.

**Structured covariance: restricted substitution patterns**

By placing structure on the covariance matrix, the estimated parameters usually become more interpretable. The structure is a restriction on the covariance matrix and, as such, reduces the ability of the model to represent various substitution patterns. However, if the structure is correct (that is, actually represents the behavior of the decision-makers), then the true substitution pattern will be able to be represented by the restricted covariance matrix.

Structure is necessarily situation-dependent: an appropriate structure for a covariance matrix depends on the specifics of the situation being modeled. Several studies using different kinds of structure were described above in section 5.2. As an example of how structure can be placed on the covariance matrix and hence substitution patterns, consider a homebuyer's choice among purchase-money mortgages. Suppose four mortgages are available to the homebuyer from four different institutions: one with a fixed rate, and three with variable rates. Suppose the unobserved portion of utility consists of two parts: the homebuyer's concern about the risk of rising interest rates, labeled $r_n$, which is common to all the variable rate loans, and all other unobserved factors, labeled collectively $\eta_{nj}$. The unobserved component of utility is then

$$\varepsilon_{nj} = -r_n d_j + \eta_{nj}$$

where $d_j = 1$ for the variable rate loans and zero for the fixed rate loan, and the negative sign indicates that utility decreases as concern about risk rises. Assume that $r_n$ is normally distributed over homebuyers with variance $\sigma$, and that $\eta_{nj}\ \forall\, j$ is iid normal with zero mean and

variance $\omega$. Then the covariance matrix for $\varepsilon_n = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$ is

$$\Omega = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \cdot & \sigma & \sigma & \sigma \\ \cdot & \cdot & \sigma & \sigma \\ \cdot & \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 1 & 0 & 0 & 0 \\ \cdot & 1 & 0 & 0 \\ \cdot & \cdot & 1 & 0 \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}$$

The model needs to normalized for scale but, as we will see, is already normalized for level. The covariance of error differences is

$$\tilde{\Omega}_1 = \begin{pmatrix} \sigma & \sigma & \sigma \\ \cdot & \sigma & \sigma \\ \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 2 & 1 & 1 \\ \cdot & 2 & 1 \\ \cdot & \cdot & 2 \end{pmatrix}.$$

This matrix has no fewer parameters than $\Omega$. In this sense, the model was already normalized for level. To normalize for scale, set $\sigma+2\omega = 1$. Then the covariance matrix becomes:

$$\tilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta & \theta \\ \cdot & 1 & \theta \\ \cdot & \cdot & 1 \end{pmatrix}$$

where $\theta = (\sigma+\omega)/(\sigma+2\omega)$. The values of $\sigma$ and $\omega$ cannot be calculated from $\theta$. However, the parameter $\theta$ provides information about the variance in utility due to concern about risk relative to that due to all other unobserved factors. For example, suppose $\theta$ is estimated to be 0.75. This estimate can be intrepreted as indicating that the variance in utility attributable to concern about risk is twice as large as the variance in utility attributable to all other factors:

$$\begin{aligned} \theta &= 0.75 \\ \frac{\sigma+\omega}{\sigma+2\omega} &= 0.75 \\ \sigma+\omega &= 0.75\sigma + 1.5\omega \\ .25\sigma &= .5\omega \\ \sigma &= 2\omega. \end{aligned}$$

Stated equivalently, $\hat{\theta} = 0.75$ means that concern about risk accounts for two-thirds of the variance in the unobserved component of utility.

Since the original model was already normalized for level, the model could be estimated without reexpressing the covariance matrix in terms of error differences. The normalization for scale could be accomplished

simply by setting $\omega = 1$ in the original $\Omega$. Under this procedure, the parameter $\sigma$ is estimated directly. Its value relative to 1 indicates the variance due to concern about risk relative to the variance due to perceptions about ease of dealing with each institution. An estimate $\hat{\theta} = 0.75$ corresponds to an estimate $\hat{\sigma} = 2$.

## 5.5 Panel data

Probit with repeated choices is similar to probit on one choice per decision-maker. The only difference is that the dimension of the covariance matrix of the errors is expanded. Consider a decision-maker who faces a choice among $J$ alternatives in each of $T$ time periods or choices situations. The alternatives can change over time, and $J$ and $T$ can differ for different decision-makers; however, we suppress the notation for these possibilities. The utility that decision-maker $n$ obtains from alternative $j$ in period $t$ is $U_{njt} = V_{njt} + \varepsilon_{njt}$. In general, one would expect $\varepsilon_{njt}$ to be correlated over time as well as over alternatives, since factors that are not observed by the researcher can persist over time. Denote the vector of errors for all alternatives in all times periods as $\varepsilon_n = \langle \varepsilon_{n11}, \ldots, \varepsilon_{nJ1}, \varepsilon_{n12}, \ldots, \varepsilon_{nJ2}, \ldots, \varepsilon_{n1T}, \ldots, \varepsilon_{nJT} \rangle$. The covariance matrix of this vector is denoted $\Omega$, which has dimension $JT \times JT$.

Consider a sequence of alternatives, one for each time period, $\mathbf{i} = \{i_1, \ldots, i_T\}$. The probability that the decision-maker makes this sequence of choices is

$$\begin{aligned} P_{n\mathbf{i}} &= Prob(U_{ni_t t} > U_{njt} \; \forall j \neq i_t, \; \forall t) \\ &= Prob(V_{ni_t t} + \varepsilon_{ni_t t} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \; \forall t) \\ &= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n. \end{aligned}$$

where $B_n = \{\varepsilon_n \text{ s.t. } V_{ni_t t} + \varepsilon_{ni_t t} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \; \forall t\}$ and $\phi(\varepsilon_n)$ is the joint normal density with zero mean and covariance $\Omega$. Compared to the probit probability for one choice situation, the integral is simply expanded to be over $JT$ dimensions rather than $J$.

It is often more convenient to work in utility differences. The probability of sequence $\mathbf{i}$ is the probability that the utility differences are negative for each alternative in each time period, when the differences in each time period are taken against the alternative identified by $\mathbf{i}$ for

that time period:

$$\begin{aligned} P_{n\mathbf{i}} &= Prob(\tilde{U}_{nji_t t} < 0 \; \forall j \neq i_t, \; \forall t) \\ &= \int_{\tilde{\varepsilon}_n \in \tilde{B}_n} \phi(\tilde{\varepsilon}_n) d\tilde{\varepsilon}_n. \end{aligned}$$

where $\tilde{U}_{nji_t t} = U_{njt} - U_{ni_t t}$; $\tilde{\varepsilon}'_n = \langle (\varepsilon_{n11} - \varepsilon_{ni_1 1}), \ldots, (\varepsilon_{nJ1} - \varepsilon_{ni_1 1}), \ldots, (\varepsilon_{n1T} - \varepsilon_{ni_T T}), \ldots, (\varepsilon_{nJT} - \varepsilon_{ni_T T}) \rangle$ with each ... being over all alternatives except $i_t$; and the matrix $\tilde{B}_n$ is the set of $\tilde{\varepsilon}_n$'s for which $\tilde{U}_{nji_t t} < 0 \; \forall j \neq i_t, \; \forall t$. This is a $(J-1)T$ dimensional integral. The density $\phi(\tilde{\varepsilon}_n)$ is joint normal with covariance matrix derived from $\Omega$. Simulation of the choice probability is the same as for situations with one choice per decision-maker, which we describe in section (5.6), but with a larger dimension for the covariance matrix and integral. Borsch-Supan *et al.* (1991) provide an example of a multinomial probit on panel data that allows covariance over time and over alternatives.

For binary choices, such as whether a person buys a particular product in each time period or works at a paid job each month, the probit model simplifies considerably (Gourieroux and Monfort, 1993). The net utility of taking the action (e.g., working) in period $t$ is $U_{nt} = V_{nt} + \varepsilon_{nt}$, and the person takes the action if $U_{nt} > 0$. This utility is called net utility because it represents the difference between the utility of taking the action compared to not taking the action. As such, it is already expressed in difference terms. The errors are correlated over time, and the covariance matrix for $\varepsilon_{n1}, \ldots, \varepsilon_{nT}$ is $\Omega$, which is $T \times T$.

A sequence of binary choices is most easily represented by a set of $T$ dummy variables: $d_{nt} = 1$ if person $n$ took the action in period $t$ and $d_{nt} = -1$ otherwise. The probability of the sequence of choices $d_n = d_{n1}, \ldots, d_{nT}$ is

$$\begin{aligned} P_{nd_n} &= Prob(U_{nt} d_{nt} > 0 \; \forall t) \\ &= Prob(V_{nt} d_{nt} + \varepsilon_{nt} d_{nt} > 0 \; \forall t) \\ &= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n \end{aligned}$$

where $B_n$ is the set of $\varepsilon_n$'s for which $V_{nt} d_{nt} + \varepsilon_{nt} d_{nt} > 0 \; \forall t$, and $\phi(\varepsilon_n)$ is the joint normal density with covariance $\Omega$.

Structure can be placed on the covariance of the errors over time. Suppose in the binary case, for example, that the error consists of a portion that is specific to the decision-maker reflecting his proclivity

to take the action, and a part that varies over time for each decision-maker: $\varepsilon_{nt} = \eta_n + \mu_{nt}$ where $\mu_{nt}$ is iid over time and people with a standard normal density, and $\eta_n$ is iid over people with a normal density with zero mean and variance $\sigma$. The variance of the error in each period is $V(\varepsilon_{nt}) = V(\eta_n + \mu_{nt}) = \sigma + 1$. The covariance between the errors in two different periods $t$ and $s$ is $Cov(\varepsilon_{nt}, \varepsilon_{ns}) = E(\eta_n + \mu_{nt})(\eta_n + \mu_{ns}) = \sigma$. The covariance matrix therefore takes the form:

$$\Omega = \begin{pmatrix} \sigma+1 & \sigma & \dots & \dots & \sigma \\ \sigma & \sigma+1 & \sigma & \dots & \sigma \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma & \dots & \dots & \sigma & \sigma+1 \end{pmatrix}.$$

Only one parameter, $\sigma$, enters the covariance matrix. Its value indicates the variance in unobserved utility across individuals (the variance of $\eta_n$) relative to the variance across time for each individual (the variance of $\mu_{nt}$). It is often called "the cross-subject variance relative to the within-subject variance."

The choice probabilities under this structure on the errors can be easily simulated using the concepts of convenient error partitioning from section 1.2. Conditional on $\eta_n$, the probability of *not* taking the action in period $t$ is $Prob(V_{nt} + \eta_n + \mu_{nt} < 0) = Prob(\mu_{nt} < -(V_{nt} + \eta_n)) = \Phi(-(V_{nt} + \eta_n))$, where $\Phi(\cdot)$ is the cumulative standard normal function. Most software packages include routines to calculate this function. The probability of taking the action, conditional on $\eta_n$, is then $1 - \Phi(-(V_{nt} + \eta_n)) = \Phi(V_{nt} + \eta_n)$. The probability of the sequence of choices $d_n$, conditional on $\eta_n$, is therefore $\prod_t \Phi((V_{nt} + \eta_n)d_{nt})$, which we can label $H_{nd_n}(\eta_n)$. So far we have conditioned on $\eta_n$, when in fact $\eta_n$ is random. The *unconditional* probability is the integral of the conditional probability $H_{nd_n}(\eta_n)$ over all possible values of $\eta_n$:

$$P_{nd_n} = \int H_{nd_n}(\eta_n)\phi(\eta_n)d\eta_n$$

where $\phi(\eta_n)$ is the normal density with zero mean and variance $\sigma$. This probability can be simulated very simply as follows. (1) Take a draw of from a standard normal density using a random number generator. Multiply the draw by $\sqrt{\sigma}$ so that it becomes a draw of $\eta_n$ from a normal density with variance $\sigma$. (2) For this draw of $\eta_n$, calculate $H_{nd_n}(\eta_n)$. (3) Repeat steps 1-2 many times and average the results. This average is a simulated approximation to $P_{nd_n}$. This simulator is much easier

to calculate than the general probit simulators described in the next section. The ability to use this simulator arises from the structure that we placed on the model, namely that the time-dependence of the unobserved factors is captured entirely by a random component $\eta_n$ that remains constant over time for each person. Gourieroux and Monfort (1993) provide an example of the use of this simulator with a probit model of this form.

The representative utility in one time period can include exogenous variables for other time periods, the same as we discussed with respect to logit models on panel data (section 3.3.3). That is, $V_{nt}$ can include exogenous variables that relate to periods other than $t$. For example, a lagged response to price changes can be represented by including prices from previous periods in the current period's $V$. Anticipatory behavior (by which, for example, a person buys a product now because he correctly anticipates that the price will rise in the future) can be represented by including prices in future periods in the current period's $V$.

Entering a lagged dependent variable is possible but introduces two difficulties that the researcher must address. First, since the errors are correlated over time, the choice in one period is correlated with the errors in subsequent periods. As a result, inclusion of a lagged dependent variable without adjusting the estimation procedure appropriately results in inconsistent estimates. This issue is analogous to regression analysis, where the ordinary least squares estimator is inconsistent when a lagged dependent variable is included and the errors are serially correlated. To estimate a probit consistently in this situation, the researcher must determine the distribution of each $\varepsilon_{nt}$ conditional on the value of the lagged dependent variables. The choice probability is then based on this conditional distribution instead of the unconditional distribution $\phi(\cdot)$ that we use above. Second, often the researcher does not observe the decision-makers' choices from the very first choice that was available to them. For example, a researcher studying employment patterns will perhaps observe a person's employment status over a period of time (e.g., from 1998-2001), but usually will not observe the person's employment status starting with the very first time the person could have taken a job (which might precede 1998 by many years). In this case, the probability for the first period that the researcher observes depends on the choices of the person in the earlier periods that the researcher does not observe. The researcher must determine

a way to represent the first choice probability that allows for consistent estimation in the face of missing data on earlier choices. This is called the "initial conditions problem" of dynamic choice models. Both of these issues, as well as potential approaches to dealing with them, are addressed by Heckman (1981*b*, 1981*a*) and Heckman and Singer (1986). Due to their complexity, I do not describe the procedures here and refer interested and brave readers to these articles.

Papatla and Krishnamurthi (1992) avoid these issues in their probit model with lagged dependent variables by assuming that the unobserved factors are independent over time. As we discussed in relation to logit on panel data (section 3.3.3), lagged dependent variables are not correlated with the current errors when the errors are independent over time and can therefore be entered without inducing inconsistency. Of course, this procedure is only appropriate if the assumption of errors being independent over time is true in reality, rather than just by assumption.

## 5.6 Simulation of the choice probabilities

The probit probabilities do not have a closed-form expression and must be approximated numerically. Several non-simulation procedures have been used and can be effective in certain circumstances. Quadrature methods approximate the integral by a weighted function of specially chosen evaluation points. A good explanation for these procedures is provided by Geweke (1996). Examples of their use for probit include Butler and Moffitt (1982) and Guilkey and Murphy (1993). Quadrature operates effectively when the dimension of the integral is small, but not with higher dimensions. It can be used for probit if the number of alternatives (or, with panel data, the number of alternatives times the number of time periods) is no more than 4 or 5. It can also be used if the researcher has specified an error-component structure with no more than 4 or 5 terms. However, it is not effective for general probit models. And even with low-dimensional integration, simulation is often easier. Another non-simulation procedure that has been suggested is the Clark algorithm, introduced by Daganzo, Bouthelier and Sheffi (1977). This algorithm utilizes the fact, shown by Clark (1961), that the maximum of several normally distributed variables is itself approximately normally distributed. Unfortunately, the approximation can be highly inaccurate in some situations (as shown by Horowitz,

Sparmann and Daganzo (1982)), and the degree of accuracy is difficult to assess in any given setting.

Simulation has proven to be very general and useful for approximating probit probabilities. Numerous simulators have been proposed for probit models; a summary is given by Hajivassiliou, McFadden and Ruud (1996). In the section above, I described a simulator that is appropriate for a probit model that has a particularly convenient structure, namely a binary probit on panel data where the time dependence is captured by one random factor. In the current section, I describe three simulators that are applicable for probits of any form: accept-reject, smoothed accept-reject, and GHK. The GHK simulator is by far the most widely used probit simulator, for reasons that we discuss. The other two methods are valuable pedagogically. They also have relevance beyond probit and can be applied in practically any situation. They can be very useful when the researcher is developing her own models rather than using probit or any other model in this book.

### 5.6.1 Accept-reject simulator

Accept-reject (A-R) is the most straightforward of any simulator. Consider simulating $P_{ni}$. Draws of the random terms are taken from their distribution. For each draw, the researcher determines whether that value of the errors, when combined with the observed variables as faced by person $n$, would result in alternative $i$ being chosen. If so, the draw is called an "accept." If the draw would result in some other alternative being chosen, the draw is a "reject". The simulated probability is the proportion of draws that are accepts. This procedure can be applied to any choice model with any distribution for the random terms. It was originally proposed for probits (Manski and Lerman, 1981), and we give the details of the approach in terms of the probit model. Its use for other models is obvious.

We use expression (5.1) for the probit probabilities:

$$P_{ni} = \int I(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \forall\, j \neq i)\phi(\varepsilon_n)d\varepsilon_n$$

where $I(\cdot)$ is an indicator of whether the statement in parentheses holds and $\phi(\varepsilon_n)$ is the joint normal density with zero mean and covariance $\Omega$. The accept-reject simulator of this integral is calculated as follows.

1. Draw a value of the $J$-dimensional vector of errors, $\varepsilon_n$, from a normal density with zero mean and covariance $\Omega$. Label the draw $\varepsilon_n^r$ with $r = 1$ and the elements of the draw as $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.

2. Using these values of the errors, calculate the utility that each alternative obtains with these errors. That is, calculate $U_{nj}^r = V_{nj} + \varepsilon_{nj}^r \ \forall j$.

3. Determine whether the utility of alternative $i$ is greater than that for all other alternatives. That is, calculate $I^r = 1$ if $U_{ni}^r > U_{nj}^r$, indicating an "accept", and $I^r = 0$ otherwise, indicating a "reject."

4. Repeat steps 1-3 many times. Label the number of repetitions (including the first) as R, such that r takes values of 1 through R.

5. The simulated probability is the proportion of draws that are accepts: $\check{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R} I^r$.

The integral $\int I(\cdot)\phi(\varepsilon_n)d\varepsilon$ is approximated by the average $\frac{1}{R}\sum I^r(\cdot)$ for draws from $\phi(\cdot)$. Obviously, $\check{P}_{ni}$ is unbiased for $P_{ni}$: $E(\check{P}_{ni}) = \frac{1}{R}\sum E[I^r(\cdot)] = \frac{1}{R}\sum P_{ni} = P_{ni}$, where the expectation is over different sets of R draws. The variance of $\check{P}_{ni}$ over different sets of draws diminishes as the number of draws rises. The simulator is often called the "crude frequency simulator," since it is the frequency of times that draws of the errors result in the specified alternative being chosen. The word "crude" distinguishes it from the "smoothed" frequency simulator that we describe in the next section.

The first step of the A-R simulator for a probit model is to take a draw from a joint normal density. The question arises: how are such draws obtained? The most straightforward procedure is that described in section (9.2.5) which uses the Choleski factor. The covariance matrix for the errors is $\Omega$. A Choleski factor of $\Omega$ is a lower-triangular matrix $L$ such that $LL' = \Omega$. It is sometimes called the generalized square root of $\Omega$. Most statistical software packages contain routines to calculate the Choleski factor of any symmetric matrix. Now suppose that $\eta$ is a vector of $J$ iid standard normal deviates, such that $\eta \sim N(0, I)$ where I is the identity matrix. This vector can be obtained by taking J draws from a random number generator for the standard normal and stacking them into a vector. We can construct a vector $\varepsilon$ that is distributed

$N(O, \Omega)$ by using the Choleski factor to tranform $\eta$. In particular, calculate $\varepsilon = L\eta$. Since the sum of normals is normal, $\varepsilon$ is normally distributed. Since $\eta$ has zero mean, so does $\varepsilon$. The covariance of $\varepsilon$ is $Cov(\varepsilon) = E(\varepsilon\varepsilon') = E(L\eta(L\eta)') = E(L\eta\eta'L') = LE(\eta\eta')L' = LIL' = LL' = \Omega$.

Using the Choleski factor $L$ of $\Omega$, the first step of the A-R simulator becomes two substeps:

1A Draw $J$ values from a standard normal density using a random number generator. Stack these values into a vector and label the vector $\eta^r$.

1B Calculate $\varepsilon_n^r = L\eta^r$.

Then, using $\varepsilon_n^r$, calculate the utility of each alternative and see whether alternative $i$ has the highest utility.

The procedure that we have described operates on utilities and expression (5.1), which is a $J$ dimensional integral. The procedure can be applied analogously to utility differences, which reduces the dimension of the integral to $J-1$. As given in (5.3), the choice probabilities can be expressed in terms of utility differences:

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

where $\phi(\tilde{\varepsilon}_{ni})$ is the joint normal density with zero mean and covariance $\tilde{\Omega}_i = M_i\Omega M_i'$. This integral can be simulated with accept-reject methods through the following steps.

1. Draw $\tilde{\varepsilon}_{ni}^r = L_i\eta^r$ as follows:
   (a) Draw $J-1$ values from a standard normal density using a random number generator. Stack these values into a vector and label the vector $\eta^r$.
   (b) Calculate $\tilde{\varepsilon}_{ni}^r = L_i\eta^r$, where $L_i$ is the Choleski factor of $\tilde{\Omega}_i$.
2. Using these values of the errors, calculate the utility difference for each alternative, differenced against the utility of alternative $i$. That is, calculate $\tilde{U}_{nji}^r = V_{nj} - V_{ni} + \tilde{\varepsilon}_{nji}^r \ \forall j \neq i$.
3. Determine whether each utility difference is negative. That is, calculate $I^r = 1$ if $U_{nji}^r < 0 \ \forall j \neq i$, indicating an "accept", and $I^r = 0$ otherwise, indicating a "reject."

4. Repeat steps 1-3 R times.

5. The simulated probability is the number of accepts divided by the number of repetitions: $\check{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R} I^r$.

Using utility differences is slightly faster computationally than using the utilities themselves, since one dimension is eliminated. However, it is often easier conceptually to remain with utilities.

As stated above, the A-R simulator is very general. It can be applied to any model for which draws can be obtained for the random terms and the behavior that the decision-maker would exhibit with these draws can be determined. It is also very intuitive, which is an advantage from a programming perspective since debugging becomes comparatively easy. However, the A-R simulator has several disadvantages, particularly when used in the context of maximum likelihood estimation.

Recall that the log-likelihood function is $LL = \sum_n \sum_j d_{nj} log P_{nj}$ where $d_{nj} = 1$ if $n$ chose $j$ and 0 otherwise. When the probabilities cannot be calculated exactly, as in the case of probit, the simulated log-likelihood function is used instead, with the true probabilities replaced with the simulated probabilities: $SLL = \sum_n \sum_j d_{nj} log \check{P}_{nj}$. The value of the parameters that maximizes $SLL$ is called the maximum simulated likelihood estimator (MSLE). It is by far the most widely used simulation-based estimation procedure. Its properties are described in Chapter 8. Unfortunately, using the A-R simulator in $SLL$ can be problematic.

There are two issues. First, $\check{P}_{ni}$ can be zero for any finite number of draws $R$. That is, it is possible that each of the $R$ draws of the error terms result in a "reject," such that the simulated probability is zero. Zero values for $\check{P}_{ni}$ are problematic because the log of $\check{P}_{ni}$ is taken when it enters the log-likelihood function and the log of zero is undefined. $SLL$ cannot be calculated if the simulated probability is zero for any decision-maker in the sample.

The occurrence of a zero simulated probability is particularly likely when the true probability is low. Often at least one decision-maker in a sample will have made a choice that has a low probability. With numerous alternatives (such as thousands of makes and models for the choice of car) each alternative has a low probability. With repeated choices, the probability for any sequence of choices can be extremely small; for example, if the probability of choosing an alternative is 0.25

in each of 10 time periods, the probability of the sequence is $(0.25)^{10}$, which is less than 0.000001.

Furthermore, $SLL$ needs to be calculated at each step in the search for its maximum. Some of the parameters values at which $SLL$ is calculated can be far from the true values. Low probabilities can occur at these parameter values even when they do not occur at the maximizing values.

Non-zero simulated probabilities can always be obtained by taking enough draws. However, if the researcher continues taking draws until at least one "accept" is obtained for each decision-maker, then the number of draws becomes a function of the probabilities. The simulation process is then not independent of the choice process that is being modeled, and the properties of the estimator become more complex.

There is a second difficulty with the A-R simulator for MSLE. The simulated probabilities are not smooth in the parameters; that is, they are not twice differentiable. As explained in Chapter 8, the numerical procedures that are used to locate the maximum of the log-likelihood function rely on the first derivatives, and sometimes the second derivatives, of the choice probabilities. If these derivatives do not exist, or do not point toward the maximum, then the numerical procedure will not perform effectively.

The A-R simulated probability is a step function, as depicted in Figure 5.1. $\check{P}_{ni}$ is the proportion of draws for which alternative $i$ has the highest utility. An infinitesimally small change in a parameter will usually not change any draw from being a reject to an accept, or vice versa. If $U_{ni}^r$ is below $U_{nj}^r$ for some $j$ at a given level of the parameters, then it will also be below for an infinitesimally small change in any parameter. So, usually, $\check{P}_{nj}$ is constant with respect to small changes in the parameters. Its derivative with respect to the parameters is zero in this range. If the parameters change in a way that a reject becomes an accept, then $\check{P}_{nj}$ rises by a discrete amount, from $M/R$ to $(M+1)/R$ where $M$ is the number of accepts at the original parameter values. $\check{P}_{nj}$ is constant (zero slope) until an accept becomes a reject or vice versa, at which point $\check{P}_{nj}$ jumps by $1/R$. Its slope at this point is undefined. The first derivative of $\check{P}_{nj}$ with respect to the parameters is either zero or undefined.

The fact that the slope is either zero or undefined hinders the numerical procedures that are used to locate the maximum of $SLL$. As discussed in Chapter 8, the maximization procedures use the gradient

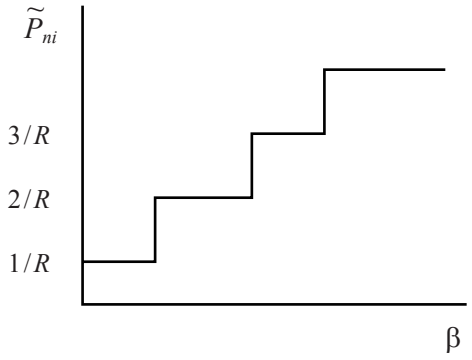


Figure 5.1: The A-R simulator is a step function in parameters.

at trial parameter values to determine the direction to move to find parameters with higher $SLL$. With the slope $\check{P}_{nj}$ for each $n$ either zero or undefined, the gradient of $SLL$ is either zero or undefined. This gradient provides no help in finding the maximum.

This issue is not actually as drastic as it seems. The gradient of $SLL$ can be approximated as the change in $SLL$ for a non-infinitesimally small change in the parameters. The parameters are changed by an amount that is large enough to switch accepts to rejects and vice versa for at least some of the observations. The approximate gradient, which can be called an arc gradient, is calculated as the amount that $SLL$ changes divided by the change in the parameters. To be precise: for parameter vector $\beta$ of length $K$, the derivate of $SLL$ with respect to the $k$-th parameter is calculated as $(SLL^1 - SLL^0)/(\beta_k^1 - \beta_k^0)$, where $SLL^0$ is calculated at the original $\beta$ with $k$-th element $\beta_k^0$ and $SLL^1$ is calculated at $\beta_k^1$ with all the other parameters remaining at their original values. The arc gradient calculated in this way is not zero or undefined, and provides information on the direction of rise. Nevertheless, the A-R simulated probability is difficult because of its inherent lack of a slope, coupled with the possibility of it being zero.

### 5.6.2 Smoothed A-R simulators

One way to mitigate the difficulties with the A-R simulator is to replace the 0-1 accept/reject indicator with a smooth, strictly positive function. The simulation starts the same as with A-R, by taking draws

of the random terms and calculating the utility of each alternative for each draw: $U_{nj}^r$. Then instead of determining whether alternative $i$ has the highest utility (that is, instead of calculating the indicator function $I^r$), the simulated utilities $U_{nj}^r \ \forall j$ are entered into a function. Any function can be used for simulating $P_{ni}$ as long as it rises when $U_{ni}^r$ rises, declines when $U_{nj}^r$ rises, is strictly positive, and has defined first and second derivatives with respect to $U_{nj}^r \ \forall j$. A function that is particularly convenient is the logit function, as suggested by McFadden (1989). Use of this function gives the "logit smoothed A-R simulator."

The simulator is implemented in the following steps, which are the same as with the A-R simulator except for step 3.

1. Draw a value of the J-dimensional vector of errors, $\varepsilon_n$, from a normal density with zero mean and covariance $\Omega$. Label the draw $\varepsilon_n^r$ with $r = 1$ and the elements of the draw as $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.

2. Using these values of the errors, calculate the utility that each alternative obtains with these errors. That is, calculate $U_{nj}^r = V_{nj} + \varepsilon_{nj}^r \ \forall j$.

3. Put these utilities into the logit formula. That is, calculate

$$S_r = \frac{e^{U_{ni}^r/\lambda}}{\sum_j e^{U_{nj}^r/\lambda}}$$

   where $\lambda > 0$ is a scale factor specified by the researcher and discussed below.

4. Repeat steps 1-3 many times. Label the number of repetitions (including the first) as R, such that r takes values of 1 through R.

5. The simulated probability is the number of accepts divided by the number of repetitions: $\check{P}_{ni} = \frac{1}{R}\sum_{r=1}^{R} S^r$.

Since $S^r > 0$ for all finite values of $U_{nj}^r$, the simulated probability is strictly positive for any draws of the errors. It rises with $U_{ni}^r$ and declines when $U_{nj}^r \ j \neq i$ rises. It is smooth (twice differentiable) since the logit formula itself is smooth.

The logit-smoothed A-R simulator can be applied to any choice model, simply by simulating the utilities under any distributional assumptions about the errors and then inserting the utilities into the

logit formula. When applied to probit, Ben-Akiva and Bolduc (1996) have called it "logit-kernel probit."

The scale factor $\lambda$ determines the degree of smoothing. As $\lambda \to 0$, $S^r$ approaches the indicator function $I^r$. Figure 5.2 illustrates the situation for a two-alternative case. For a given draw of $\varepsilon_n^r$, the utility of the two alternatives is calculated. Consider the simulated probability for alternative 1. With A-R, the 0-1 indicator function is zero if $U_{n1}^r$ is below $U_{n2}^r$ and one if $U_{n1}^r$ exceeds $U_{n2}^r$. With logit-smoothing, the step function is replaced by a smooth sigmoid curve. The factor $\lambda$ determines the proximity of the sigmoid to the 0-1 indicator. Lowering $\lambda$ increases the scale of the utilities when they enter the logit function (since the utilities are divided by $\lambda$.) Increasing the scale of utility increases the absolute difference between the two utilities. The logit formula gives probabilities that are closer to zero or one when the difference in utilities is larger. The logit-smoothed $S^r$ therefore becomes closer to the step function as $\lambda$ becomes closer to zero.

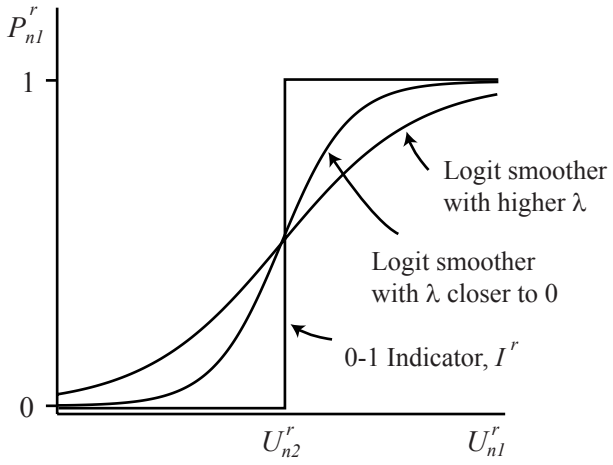


Figure 5.2: A-R smoother.

The researcher needs to set the value of $\lambda$. A lower value of $\lambda$ makes the logit-smoother a better approximation to the indicator function. However, this fact is a double-edge sword: if the logit-smoother approximates the indicator function too well, the numerical difficulties of using the unsmoothed A-R simulator will simply be reproduced in the

logit-smoothed simulator. The researcher wants to set $\lambda$ low enough to obtain a good approximation but not so low as to re-introduce numerical difficulties. There is little guidance on the appropriate level of $\lambda$. Perhaps the best approach is for the researcher to experiment with different $\lambda$'s. The same draws of $\varepsilon_n$ should be used with each different $\lambda$, so as to assure that differences in results are due to the change in the $\lambda$ rather than to differences in the draws.

McFadden (1989) describes other smoothing functions. For all of them, the researcher must specify the degree of smoothing. An advantage of the logit smoother is its simplicity. Also, we will see in Chapter 6 that the logit-smoother applied to a probit or any other model constitutes a type of mixed logit specification. That is, instead of seeing the logit smoother as providing an approximation that has no behavioral relation to the model (simply serving a numerical purpose), the logit smoother can be seen as arising from a particular type of error structure in the behavioral model itself. Under this interpretation, the logit formula applied to simulated utilities is not an approximation but actually represents the true model.

### 5.6.3 GHK simulator

The most widely used probit simulator is called GHK, after Geweke (1989, 1991), Hajivassiliou (as reported in Hajivassiliou and McFadden, 1998), and Keane (1990 1994), who developed the procedure. In a comparison of numerous probit simulators, Hajivassiliou et al. (1996) found GHK to be the most accurate in the settings that they examined. Geweke, Keane and Runkle (1994) found the GHK simulator works better than smoothed A-R. Experience has confirmed its usefulness and relative accuracy (e.g., Borsch-Supan and Hajivassiliou (1993)).

The GHK simulator operates on utility differences. The simulation of probability $P_{ni}$ starts by subtracting the utility of alternative $i$ from each other alternative's utility. Importantly, the utility of a different alternative is subtracted depending on which probability is being simulated: for $P_{ni}$, $U_{ni}$ is subtracted from the other utilities, while for $P_{nj}$, $U_{nj}$ is subtracted. This fact is critical to the implementation of the procedure.

I will explain the GHK procedure first in terms of a three-alternative case, since that situation can be depicted graphically in two dimensions for utility differences. I will then describe the procedure in general for

any number of alternatives. Bolduc (1993, 1999 provide an excellent alternative description of the procedure, along with methods to simulate the analytic derivatives of the probit probabilities. Keane (1994) provides a description of the use of GHK for transition probabilities.

**Three alternatives**

We start with a specification of the behavioral model in utilities: $U_{nj} = V_{nj} + \varepsilon_{nj}$, $j = 1, 2, 3$. The vector $\varepsilon_n' = \langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle \sim N(0, \Omega)$. We assume that the reseacher has normalized the model for scale and level such that the parameters that enter $\Omega$ are identified. Also, $\Omega$ can be a parametric function of data, as with random taste variation, though we do not show this dependence in our notation.

Suppose we want to simulate the probability of the first alternative, $P_{n1}$. We re-express the model in utility differences by subtracting the utility of alternative 1:

$$\begin{aligned} (U_{nj} - U_{n1}) &= (V_{nj} - V_{n1}) + (\varepsilon_{nj} - \varepsilon_{n1}) \\ \tilde{U}_{nj1} &= \tilde{V}_{nj1} + \tilde{\varepsilon}_{nj1} \end{aligned}$$

for $j = 2, 3$. The vector $\tilde{\varepsilon}_{n1}' = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31} \rangle$ is distributed $N(0, \tilde{\Omega}_1)$ where $\tilde{\Omega}_1$ is derived from $\Omega$.

We take one more transformation to make the model more convenient for simulation. In particular, let $L_1$ be the Choleski factor of $\tilde{\Omega}_1$. Since $\tilde{\Omega}_1$ is $2 \times 2$ in our current illustration, $L_1$ is a lower-triangular matrix that takes the form:

$$L_1 = \begin{pmatrix} c_{aa} & 0 \\ c_{ab} & c_{bb} \end{pmatrix}.$$

Using this Choleski factor, the original error differences, which are correlated, can be re-written as linear functions of *uncorrelated* standard normal deviates:

$$\begin{aligned} \tilde{\varepsilon}_{n21} &= c_{aa}\eta_1 \\ \tilde{\varepsilon}_{n31} &= c_{ab}\eta_1 + c_{bb}\eta_2 \end{aligned}$$

where $\eta_1$ and $\eta_2$ are iid $N(0, 1)$. The error differences $\tilde{\varepsilon}_{n21}$ and $\tilde{\varepsilon}_{n31}$ are correlated because both of them depend on $\eta_1$. With this way of expressing the error differences, the utility differences can be written:

$$\begin{aligned} \tilde{U}_{n21} &= \tilde{V}_{n21} + c_{aa}\eta_1 \\ \tilde{U}_{n31} &= \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2. \end{aligned}$$