

REDACTED – FOR PUBLIC INSPECTION

- Post-merger market shares of cable, DIRECTV and EchoStar in small DMAs (ranked 71-210): These fields (which are located in the “Welfare Changes: Consumer Counts in DMAs 71-210” table) display shares in small DMAs following the merger. The shares are taken from the “post_cable_share”, “post_dtv_share”, and “post_es_share” fields in *Scaled_div[/i> *_Output_MC_*.csv* output files (we calculate the average share in these DMAs, weighted by adj_hh).*
- Effects of LIL introduction on DBS shares: These are in the “Inputs for LIL Value” table, and they are taken from the regression results in Table 2 Addendum 1.
- The value of α (the logit price coefficient) is in the “General Inputs” table, and it’s taken from the regression results in Table 1 in Addendum 1.

b. Output Fields (listed in Output Tables)

The output fields are calculations based on data in the input fields. Key output fields include the following:

- Value of LIL service: The methodology for calculating the dollar value of LIL service is described in Section 3. The historical value of LIL (*i.e.*, the value of LIL based on the share lift experienced by the parties over the past few years) is calculated in the “Outputs for Historical LIL Value” table and the “Auxiliary LIL Calculations: Historical Estimates” table. Note that the value of LIL is allowed to differ across various categories of households. For example, the historical value of LIL to an existing EchoStar subscriber is \$[]. This value is listed in cell J31. In contrast, the value of EchoStar having LIL to someone who moves from cable to EchoStar in response to EchoStar’s introduction of LIL is listed as “Cable to ES” and the corresponding dollar amount (\$[]) is in cell J41. Similarly, the value of EchoStar LIL to those who move from antenna to EchoStar is listed as “Antenna to ES” and the corresponding dollar amount (\$[]) is in cell J43. Note that in each case, examining the formula used in each cell allows the user to trace the exact method for calculating these LIL values.
- Share changes in response to LIL: These calculations are in the “Auxiliary LIL Calculations”, and “Outputs for Historical LIL Value” tables. *E.g.* cells J35 and J37 show that—net of the cable price reaction to LIL introduction—EchoStar gains [] percentage points share from cable and [] percentage

REDACTED – FOR PUBLIC INSPECTION

points from antenna in response to both EchoStar and DIRECTV introducing LIL post merger.⁴⁵

c. *Welfare Changes fields (in the Welfare Changes table)*

These fields use data from the output and input fields to calculate the welfare effects of the merger. To illustrate how this table works, we delineate the analysis for the case where marginal costs decline following the merger include SAC savings (*i.e.*, EchoStar marginal costs declines by \$[] and DIRECTV by \$[]). The welfare calculation begins with the consumer welfare change before LIL Introduction *i.e.*, welfare changes *only* due to the price effects of the merger. Since prices decline because of the marginal cost decreases, the initial effect of the merger is to increase consumer welfare by \$[] a month, which is seen in cell G24. Next, we include the direct effects of LIL introduction on existing DBS subscribers (post-merger) of \$[] a month in cell G41 and the indirect effect of LIL introduction (*i.e.*, the cable price reduction) on existing cable subscribers of \$[] in cell G42. Finally, the welfare effect of households switching to DBS in response to LIL introduction post merger is \$[] per month, which is in cell G43; the welfare effect of households switching from antenna to cable in response to the cable price drop is \$[] is in cell G44. The sum of these four (which equals the monthly consumer welfare change due to LIL introduction) equals \$[], which is in cell G45. Annualizing this figure and adding to the initial (pre-LIL expansion) change in welfare generates a total welfare increase of \$[] (which is in cell G53).

Key results are summarized in the “Summary of Results” table.

⁴⁵ Note that the predicted percentage point increase in EchoStar share due to LIL introduction post-merger is [] rather than the [] in the LIL regressions in Table 3 in Addendum 1. The reason is that the [] percentage point increase is the effect of EchoStar introducing LIL without conditioning on DIRECTV's LIL service status, whereas the [] percentage point increase is the share lift to EchoStar when it introduce LIL simultaneously with DIRECTV.

REDACTED – FOR PUBLIC INSPECTION

ADDENDUM 5: DESCRIPTION OF RESCALING PROCESS TO ALIGN SAMPLE CABLE FRANCHISE AREA SHARES WITH NATIONAL JANUARY 2002 MARKET SHARES

The final dataset that we use to simulate the merger includes cable franchise areas that account for 76 million households, and the national average shares of DBS products in our data are very close to their national shares in 2002. However, the share of cable is less than its national share.

To test the robustness of our simulation results, we re-weighted our data such that weighted average shares equal their national shares. The weight of each area is the number of households in that area multiplied by a rescaling factor. In other words, the goal of the rescaling exercise was to create a rescaled household count for each area such that (a) the weighted average shares of MVPD products equal their known national shares, and (b) the total number of households in the nation equals 106.3 million⁴⁶. To minimize distorting the original data, the rescaling procedure did not alter any characteristics of a cable franchise area, such as the cable franchise area share of each product. Instead, the rescaling procedure simply rescaled the household count in cable franchise areas. Also, to minimize the distortion of the original data, the change in the observed households count in an area was constrained (*i.e.*, we try to minimize (numerically) the range of rescaling factors).

In short, the goal of the rescaling is to create a vector of household numbers, $\hat{N} = \{\hat{N}_1, \dots, \hat{N}_H\}$, such that:

$$\sum_i \hat{N}_i = \hat{N},$$

where \hat{N}_i denotes the rescaled number of households in cable franchise area i , and \hat{N} is the actual national household count, and

⁴⁶ Census Bureau data show 106.3 million households in the U.S. in 2001.

$$\frac{1}{\hat{N}} \sum_i \hat{N}_i s_i^j = \bar{s}^j,$$

where s_i^j represents the share of product j in cable franchise area i , and \bar{s}^j is the true national share of product j .

Each cable franchise area was assigned a *scalar* proportional to the difference between cable franchise area-specific and true national cable franchise area shares. The scalar for cable franchise area i is given by:

$$Scalar_i = \sum_{j \in \{DTV, ECHO, CABLE\}} \omega^j (s_i^j - \bar{s}^j)$$

Where ω^j is a product-specific weight defined below. The rescaling was accomplished through an exponential rescaling function, r , given by:

$$r_i = \alpha \frac{\exp\{scalar_i\}}{1 + \exp\{scalar_i\}} + \beta$$

Where α and β bound the minimum and maximum scaling for a cable franchise area; they are scaling parameters which control how much cable franchise areas are allowed to “grow” and “shrink.” Given the scaling function, r :

$$\hat{N}_i = \frac{r_i N_i \hat{N}}{\sum_i r_i N_i}$$

Where N_i is the original number of households in cable franchise area i .

Hence, the above rescaling procedure requires assigning values for two scaling parameters, α and β , and three product-specific parameters, ω^j , $j \in \{DTV, ECHO, CABLE\}$. In order to find appropriate values for these parameters, a loss function was constructed:

$$L = .05 \left(\frac{1}{\hat{N}} \sum_i \hat{N}_i s_i^{ECHO} - \bar{s}^{ECHO} \right)^2 + 10 \left(\frac{1}{\hat{N}} \sum_i \hat{N}_i s_i^{DTV} - \bar{s}^{DTV} \right)^2 + 20 \left(\frac{1}{\hat{N}} \sum_i \hat{N}_i s_i^{CABLE} - \bar{s}^{CABLE} \right)^2$$

The loss function represents the difference between the rescaled cable franchise area shares and the actual cable franchise area shares. The parameters in the above loss function as well as the values of α and β were obtained by trial and error. Since the vector \hat{N} depends on the values of the ω^j s, we solve for ω^j by solving the following optimization problem:

$$\min_{\omega^j, j \in \{DTV, ECHO, CABLE\}} L$$

The minimization problem is solved using a numerical reduced gradient algorithm. This is done using the “Excel Solver” add-in functionality. The trial-and-error pattern for α and β is aimed to minimize the differences between original and rescaled numbers of households within each cable franchise area. It was found that some cable franchise areas need to be scaled at least [] fold (or by []) to allow the rescaling procedure to match national cable franchise area shares. That is, a “tighter” bound on the ratio of unscaled to scaled cable franchise area sizes did not allow for rescaling that met the objective of equating the data to observed national cable franchise area shares. The obtained parameter values were:

ω^{DTV}	
ω^{ECHO}	
ω^{CABLE}	
A	
B	

REDACTED – FOR PUBLIC INSPECTION

The resulting rescaling factor, \hat{N}_i / N_i , ranged from [] to [], the resulting national household count matched the actual household count, and rescaled product shares matched actual national shares to within four decimal places.

The spreadsheet used to do the rescaling is in the file *DataScaling_Spreadsheet.xls*. This file includes a worksheet that has instructions on how to use the workbook.

While the averages of the estimated asymptotic variances are never identical—in fact, that of the e.g.f. estimator is always smaller—they converge as T increases.

These results illustrate that the c.g.f. is a valid and indeed a sensible alternative to the e.c.f. estimator, especially in cases where the c.f. is given as the exponential of some function of the parameters.

NOTE

1. Although one of the referees commented that intuitively one would expect asymptotic equivalence between e.c.f. and c.g.f., this theorem proves this equivalence.

REFERENCES

- Feuerverger, A. (1990) An efficiency result for the empirical characteristic function in stationary time series models. *Canadian Journal of Statistics* 18, 155–161.
- Feuerverger, A. & P. McDunnough (1981a) On some Fourier methods for inference. *Journal of the American Statistical Association* 76, 379–387.
- Feuerverger, A. & P. McDunnough (1981b) On the efficiency of the empirical characteristic function procedures. *Journal of the Royal Statistical Society, Series B* 43, 20–27.
- Feuerverger, A. & R. Mureika (1977) The empirical characteristic function and its applications. *Annals of Statistics* 5, 88–97.
- Jorion, P. (1988) On jump processes in the foreign exchange and stock markets. *Review of Financial Studies* 1, 427–445.
- Knight, J.L. & S.E. Satchell (1993) GARCH Processes—Via Compound Poisson Processes. Mimeo, University of Western Ontario.
- Knight, J.L. & S.E. Satchell (1994). Estimation of stationary stochastic processes via the empirical characteristic function. Mimeo, University of Western Ontario, London, Ontario.
- Quandt, R. & J. Ramsey (1978) Estimating mixtures of Normal distributions and switching regressions. *Journal of the American Statistical Association* 73, 730–738.
- Schmidt, P. (1982) An improved version of the Quandt-Ramsey mgf estimator for mixtures of Normal distributions and switching regressions. *Econometrica* 50, 501–516.

VARIANCE COMPONENTS STRUCTURES FOR THE EXTREME-VALUE AND LOGISTIC DISTRIBUTIONS WITH APPLICATION TO MODELS OF HETEROGENEITY

N. SCOTT CARDELL
Salford Systems

Two new classes of probability distributions are introduced that radically simplify the process of developing variance components structures for extreme-value and logistic distributions. When one of these new variates is added to an extreme-value (logistic) variate, the resulting distribution is also extreme value (logistic). Thus, quite complicated variance structures can be generated by recursively adding components having this new distribution, and the result will retain a marginal extreme-value (logistic) distribution. It is demonstrated that the computational simplicity of extreme-value error structures extends to the introduction of heterogeneity in duration, selection bias, limited-dependent- and qualitative-variable models. The usefulness of these new classes of distributions is illustrated with the examples of nested logit, multivariate risk, and competing risk models, where important generalizations to conventional stochastic structures are developed. The new models are shown to be computationally simpler and far more tractable than alternatives such as estimation by simulated moments. These results will be of considerable use to applied microeconomic researchers who have been hampered by computational difficulties in constructing more sophisticated estimators.

1. INTRODUCTION

Econometricians and statisticians have long been concerned with the problems of unobserved heterogeneity in cross-sectional studies, where systematic differences among economic agents cannot be captured with demographic

This paper is based in part on my unpublished Ph.D. dissertation (Harvard University, 1989). I thank Zvi Griliches and Gary Chamberlain, who advised and commented on that earlier work; Ernst Stromsdorfer, Ron Mittelhammer, Dan Steinberg, and John Trimble, who commented on this paper; and particularly Jerry Hausman and Audrey Cardell, who advised and made many useful suggestions on both my dissertation and this paper. Finally, I thank the anonymous referees, each of whom contributed valuable comments, and Peter Phillips, Editor, who supplied the key comments for the final revision of this paper. Address correspondence to: N. Scott Cardell, 1730 Kenny Drive, Pullman, WA 99163, USA.

and other available data. At a minimum, failing to account for such heterogeneity will result in inefficient estimation. It may also result in biased and inconsistent model estimates. There have been major impediments to accommodating heterogeneity in applied empirical work, however. First, without sufficient data in the form of repeated measures of some type, models may not be identified. Second, even with adequate data, tractable estimators based on reasonable rather than computationally convenient assumptions are rare. With the advent of new microeconomic databases consisting of lengthy panels in which endogenous variables are repeatedly measured, and cross-sections containing a large number of related endogenous variables, data sets capable of supporting quite complex heterogeneity modeling are now routinely available. It is therefore of considerable importance to find methods that are practicably usable by applied researchers and theoretically general enough to accommodate a wide range of heterogeneity structures. This paper derives two new classes of conjugate distributions and utilizes them to develop a variety of broadly applicable variance components structures.

Heterogeneity has been discussed widely in the hazard function literature and in the literature on cross-equation correlation, including correlated linear and discrete equations. However, heterogeneity has rarely been considered when modeling multiple discrete decisions. Furthermore, empirical investigations have often ignored important aspects of heterogeneity. The focus of this paper is on three situations in which heterogeneity is important: (1) multiple hazard processes, (2) multiple discrete decisions, and (3) combinations of discrete decisions and continuous stochastic processes. In each situation, heterogeneity will affect multiple decisions and thus can induce correlations among the decisions, resulting in inconsistent conventional estimators. For both hazard function and discrete choice models, heterogeneity can be parsimoniously represented by variance components.

Since the 1920's, logit models have been the preferred method for dealing with binomial and multinomial discrete choice for reasons of computational simplicity, parsimony, and robustness. The logit specification is based on an indicator equation that includes a logistically distributed additive stochastic term. If the indicator is positive, one choice is made; if not, the other choice is made. Because the difference of two independent Type I extreme-value variates is logistically distributed, binomial logit is equivalent to a utility maximization problem, with each utility having an independent Type I extreme-value additive stochastic term. In the multinomial logit model, the utility for each of the multiple alternatives has an independent Type I extreme-value additive stochastic term. Often, multinomial logit is the only computationally feasible multinomial choice model.¹

The chief practical advantage of the extreme-value and logistic variance components structures introduced here is that they combine computational simplicity and flexibility, a combination for which there is a very real need.

Applied modeling is often restricted by the computational difficulty of evaluating the cumulative multivariate distribution functions involved. As a result, a large and growing literature on simulation-based methods has developed.² The variance components structures in this paper avoid such computational problems altogether by providing closed-form expressions for the multivariate cumulative probabilities.

An important advantage of the methods introduced here is that they can be used in combination with nonparametric approaches. For example, in duration models they can be combined with the nonparametric baseline hazard of Han and Hausman (1990). In large samples, heterogeneity distributions could be modeled as a convolution of a finite points-of-support distribution and a parametric distribution of the type introduced in this paper. The resulting estimator would combine consistent estimation of general heterogeneity distributions with the practical advantage of computational simplicity. Such approaches avoid the perils of a "have distribution will fit" mentality that can occur when modeling becomes too complex to allow appropriate specification searches.

The remainder of the paper is organized as follows. Section 2 presents the main theoretical results, deriving the two new classes of distribution functions, denoted $C(\lambda)$ and $GL(\lambda)$, and showing how they can be used to develop variance components structures for Type I extreme-value and logistic random variables. The remaining sections are devoted to applications. Section 3 utilizes variance components to derive the nested logit model, Section 4 develops cumulative distribution functions for vectors with variance components structures based on $C(\lambda)$ and $GL(\lambda)$, and Section 5 utilizes these cumulative distributions to apply the variance components models to some of the leading econometric problems in the literature. These include (1) correlated, multiple discrete decision models, (2) hazard models with heterogeneity, and (3) simultaneous equations models with discrete endogenous equations. Concluding remarks are contained in the last section. All proofs appear in the Appendix.

2. VARIANCE COMPONENTS STRUCTURES

This section derives the classes of distribution functions $C(\lambda)$ and $GL(\lambda)$ and their properties necessary for constructing extreme-value variance components structures. In general, a linear structure for the random p -vector η has the form $\eta = \mu + A\nu$, where μ is a vector of constants, ν is a q -vector of non-degenerate independent random variables, and A is a constant $p \times q$ matrix. Kagan, Linnik, and Rao (1973) extensively discuss the use of linear structures. When η has a finite variance, the linear structure is referred to as a variance components structure.

2.1. Overview

If a $C(\lambda)$ variate is added to extreme-value variates, the resulting sum is an extreme-value random variate. The extreme-value distribution is itself a member of the $C(\lambda)$ class. Therefore, one can derive extreme-value variables recursively, which are themselves sums of $C(\lambda)$ random variables. Then, in a factor components framework in which various $C(\lambda)$ random variables appear in the construction of several extreme-value variables, quite complicated variance components structures may be developed while retaining the assumption that the marginal distributions are extreme value. Thus, the selection of the shape of the marginal distributions is kept separate from the statistical determination of the covariance between them. Corresponding results hold for the $GL(\lambda)$ class, allowing one to build sophisticated logistic variance components structures. These structures are particularly appropriate for modeling heterogeneity.

Cardell (1975) conjectured a class of distributions that could form the basis for useful variance components structures for the Type I extreme-value distribution. For all λ , $0 \leq \lambda \leq 1$, let $C(\lambda)$ denote a member of that class. The $C(\lambda)$ distribution is defined to be that unique distribution for which ν and ϵ independent, $\nu \sim C(\lambda)$, and $\epsilon \sim$ Type I extreme value, implies that $\nu + \lambda\epsilon \sim$ Type I extreme value. In other words, the $C(\lambda)$ distributions are conjugate to the Type I extreme-value distribution. This paper demonstrates that the $C(\lambda)$ distribution exists and that the nested logit can be based on $C(\lambda)$ variates. Parallel to the $C(\lambda)$ class of distributions is the $GL(\lambda)$ class. The $GL(\lambda)$ distribution is defined to be that unique distribution for which ν distributed as $GL(\lambda)$ and ϵ independent and logistically distributed implies that $\nu + \lambda\epsilon$ is distributed as logistic. Thus, the $GL(\lambda)$ distributions are conjugate to the logistic distribution. Sums of appropriately scaled $C(\lambda)$ (or $GL(\lambda)$) random variates can produce flexible variance components structures for vectors of Type I extreme-value random variables (or logistic random variables). As shown later, the binomial variance components structure for logistic random variables allows the correlation to vary freely from -1 to 1 . Researchers have searched for a freely correlated bivariate logistic distribution at least since Gumbel (1961; see also Johnson and Kotz, 1972). A similar variance components structure can be used to develop more plausible statistical properties for logit-type discrete choice models with very large samples (Cardell, 1989).

For the $C(\lambda)$ and $GL(\lambda)$ distributions, the single parameter λ determines the mean, the shape, and the scale of the distribution. It is sometimes useful to generalize the $C(\lambda)$ and $GL(\lambda)$ distributions to include a scale parameter. Accordingly, if ϵ is distributed as $C(\lambda)$ and δ is a fixed scalar, $\delta\epsilon$ is said to be distributed as $C(\lambda, \delta)$. Similarly, if ϵ is distributed as $GL(\lambda)$ and δ is a fixed scalar, $\delta\epsilon$ is said to be distributed as $GL(\lambda, |\delta|)$.³

2.2. The Existence of $C(\lambda)$ and $GL(\lambda)$

This investigation began by hypothesizing that a distribution $C(\lambda)$ exists for all λ , $0 \leq \lambda \leq 1$, such that $\nu \sim C(\lambda)$ and $\epsilon \sim$ Type I extreme value implies that $\nu + \lambda\epsilon \sim$ Type I extreme value. Obviously, $C(0)$ is the Type I extreme-value distribution. Theorem 2.1 shows that $C(\lambda)$ exists for $0 < \lambda < 1$. The motivation here can be understood by analogy to the normal distribution. The normal distribution is self-conjugate because, for ν and ϵ independently distributed, $\nu + \epsilon$ is normal iff ν and ϵ are normally distributed (Cramer, 1937). Therefore, given $\epsilon \sim N(0, \sigma_1^2)$, ν and ϵ independent, $\nu + \epsilon \sim N(0, \sigma_1^2 + \sigma_2^2)$ iff $\nu \sim N(0, \sigma_2^2)$. Similarly, Theorem 2.1 shows that for ϵ extreme value with scale parameter δ (i.e., $\epsilon \sim C(0, \delta)$, $\delta > 0$), for ν independently distributed, $\nu + \epsilon$ is extreme value with scale parameter γ ($\gamma > \delta$) iff $\nu \sim C((\delta/\gamma), \delta)$. Thus, $C(\lambda, \delta)$ is the class of conjugate distributions to the extreme value.

THEOREM 2.1. *For $0 < \lambda < 1$ and ϵ , a random variable distributed as Type I extreme value, there exists a unique distribution, denoted $C(\lambda)$, such that for ν , a random variable, ν and ϵ independent, then $\nu + \lambda\epsilon$ is a random variable distributed as Type I extreme value, iff ν is distributed as $C(\lambda)$ where the probability density function (p.d.f.) of $C(\lambda)$ is $f_\lambda(\nu) = (1/\lambda) \sum_{n=0}^{\infty} [((-1)^n e^{-n\nu}) / (n! \Gamma(-\lambda n))]$. The cumulative distribution function (c.d.f.) of the $C(\lambda)$ does not have a closed-form representation.*

It is interesting to note that as a consequence of Theorem 2.1 $C(\lambda)$ are infinitely divisible distributions. Thus, other decompositions exist for Type I extreme-value random variates and logistic random variates, for example, in terms of sums of independent and identically distributed (i.i.d.) random variables. Furthermore, the self-decomposability of the Type I extreme-value distribution is independently interesting. The stable distributions (stable under convolution) are a subclass of the self-decomposable distributions, and the exponential distribution is well known to be self-decomposable. However, the Type I extreme-value and exponential distributions are stable under maximization but not stable under convolution. Thus, Theorem 2.1 demonstrates an interesting relationship between the stable and the extreme-value or max-stable distributions.

While many useful results can be computed in closed form using the $C(\lambda)$ distribution, the c.d.f. for $C(\lambda)$ does not have a general closed-form representation and the p.d.f. appears not to as well. By contrast, both the c.d.f. and p.d.f. for the generalized logistic distribution have simple closed-form expressions. Clearly, $C(1)$ represents the degenerate case $\nu = 0$ with probability 1. Also, $\text{Var}(\epsilon) = \text{Var}(\nu + \lambda\epsilon) = \text{Var}(\nu) + \lambda^2 \text{Var}(\epsilon)$, and $C(\lambda)$ cannot be defined for $|\lambda| > 1$. Furthermore, $C(\lambda)$ cannot be defined for $\lambda < 0$ (Cardell, 1989); thus, $C(\lambda)$ is defined iff $0 \leq \lambda \leq 1$.

LEMMA 2.1. If $\eta_1, \eta_2 \stackrel{i.i.d.}{\sim}$ Type I extreme value, then $\eta_1 - \eta_2 \sim$ logistic. Furthermore, the $GL(\lambda)$ distribution exists for $0 < \lambda < 1$, and, for $\eta_1, \eta_2 \stackrel{i.i.d.}{\sim} C(\lambda)$, $\eta_1 - \eta_2 \sim GL(\lambda)$.

Just as Theorem 2.1 provides a unique class of conjugate distributions to the extreme-value distribution, so Lemma 2.1 extends Theorem 2.1 to provide a unique class of conjugate distributions to the logistic distribution. $C(\lambda, \delta)$ and $GL(\lambda, \delta)$ can now be simply defined: $C(\lambda, \delta)$ is the distribution of δv for δ , a fixed scalar, and $v \sim C(\lambda)$ and $GL(\lambda, \delta)$ is the distribution of δv for δ , a positive fixed scalar, and $v \sim GL(\lambda)$.

2.3. Nesting the Logistic Distribution

THEOREM 2.2. The p.d.f. of generalized logistic distribution is $f_\lambda(Z) = [(\sin \pi \lambda) / (\pi \lambda)] / (e^Z + 2 \cos \pi \lambda + e^{-Z})$ and the c.d.f. is $F_\lambda(Z) = (1/\pi \lambda) \tan^{-1}((\sin \pi \lambda) / (e^{-Z} + \cos \pi \lambda))$, where the range of \tan^{-1} is taken to be $(0, \pi)$.

An interesting consequence of Theorem 2.2 is⁴

$$\frac{1}{1 + e^{-z}} = \int_{-\infty}^{\infty} \frac{\frac{\sin \pi \lambda}{\pi \lambda}}{e^x + 2 \cos \pi \lambda + e^{-x}} \frac{1}{1 + e^{-(z-x)/\lambda}} dx. \quad (1)$$

To nest the logistic distribution in a broader class, it is convenient to define a class of distributions that includes $GL(\lambda)$ as a subset. Define the extended generalized logistic distribution $GL^*(a)$ by its p.d.f. and c.d.f. as follows:

$$f_a(x) = \frac{\sqrt{a^2 - 1}}{\log(a + \sqrt{a^2 - 1})} \frac{1}{e^x + 2a + e^{-x}}$$

and

$$F_a(x) = 1 - \frac{1}{2} \left(\frac{\log \left(\frac{e^x + a + \sqrt{a^2 - 1}}{e^x + a - \sqrt{a^2 - 1}} \right)}{\log(a + \sqrt{a^2 - 1})} \right).$$

Letting logs and square roots be defined to take their principal values, for $-1 \leq a \leq 1$, $GL^*(a) = GL((1/\pi) \cos^{-1}(a))$. Therefore, $GL^*(a)$ is defined for $a > -1$, and the logistic distribution corresponds to an interior point in the allowed range of a . This family of distributions thus provides a way to nest the logistic distribution and to test it statistically (i.e., $H_0: a = 1$). For situations where the mean and variance are free parameters, the generalized logistic distribution with mean μ and variance σ^2 has a p.d.f. of

$$f_a(x; \mu, \sigma^2) = \frac{c(a) \frac{g(a)}{\sigma}}{e^{(x-\mu)g(a)/\sigma} + 2a + e^{-(x-\mu)g(a)/\sigma}},$$

where

$$g(a) = \begin{cases} \sqrt{\frac{\pi^2 + (\ln(a + \sqrt{a^2 - 1}))^2}{3}} & \text{for } a \geq 1, \\ \sqrt{\frac{\pi^2 - (\cos^{-1}(a))^2}{3}} & \text{for } -1 \leq a \leq 1, \end{cases}$$

$$c(a) = \begin{cases} \frac{\sqrt{a^2 - 1}}{\ln(a + \sqrt{a^2 - 1})} & \text{for } a \geq 1, \\ \frac{\sqrt{1 - a^2}}{\cos^{-1}(a)} & \text{for } -1 \leq a \leq 1. \end{cases}$$

2.4. Building Variance Components Structures Recursively

LEMMA 2.2. For $\nu_1 \sim C(\lambda_1)$, $\nu_2 \sim C(\lambda_2)$, ν_1 and ν_2 independent, $\nu_1 + \lambda_1 \nu_2 \sim C(\lambda_1 \cdot \lambda_2)$.

Repeatedly applying Theorem 2.1 allows one to create recursively a variance components structure for the Type I extreme-value distribution that has any number of terms. For instance, if ν_1, ν_2, ν_3 , and ϵ are independent, $\nu_1 \sim C(\lambda_1)$, $\nu_2 \sim C(\lambda_2)$, $\nu_3 \sim C(\lambda_3)$, and ϵ is extreme value. Then, $(\nu_3 + \lambda_3 \epsilon)$ is extreme value. Thus, $\nu_2 + \lambda_2(\nu_3 + \lambda_3 \epsilon) = \nu_2 + \lambda_2 \nu_3 + \lambda_2 \lambda_3 \epsilon$ is extreme value, and $\nu_1 + \lambda_1(\nu_2 + \lambda_2 \nu_3 + \lambda_2 \lambda_3 \epsilon) = \nu_1 + \lambda_1 \nu_2 + \lambda_1 \lambda_2 \nu_3 + \lambda_1 \lambda_2 \lambda_3 \epsilon$ is extreme value.

THEOREM 2.3. Let Q be the number of terms in the variance components structure. Given $0 \leq \lambda_k < 1$ for $1 \leq k \leq Q$, $\lambda_0 = 1$, $\xi_k = \prod_{i=0}^k \lambda_i$. It follows that

- A. If ν_k is independently distributed as $C(\lambda_k)$ for $k = 1, \dots, Q$, $\sum_{k=1}^Q \xi_{k-1} \nu_k \sim C(\xi_Q)$.
- B. Alternatively, for η_k independently distributed as $C(\lambda_k, \xi_{k-1})$, $\sum_{k=1}^Q \eta_k \sim C(\xi_Q)$.

Using λ_k, ξ_k, ν_k , and η_k as in Theorem 2.3 with $\lambda_Q = 0, 0 < \lambda_k < 1, 1 \leq k < Q$, $\sum_{k=1}^Q \xi_{k-1} \nu_k$ or $\sum_{k=1}^Q \eta_k$ is a Q term variance components structure for a single Type I extreme-value random variable. Now consider a vector of J Type I extreme-value random variables denoted ζ_j for $1 \leq j \leq J$. Let Q_j

denote the number of terms in the variance components structure for ζ_j . Let $\lambda_{k,j}$ be fixed constants, $0 \leq \lambda_{k,j} < 1$ for $1 \leq j \leq J$ and $1 \leq k \leq Q_j$. Similarly to Theorem 2.3, let $\nu_{k,j}$ be distributed as $C(\lambda_{k,j})$, set $\lambda_{0,j} \equiv 1$, and define $\xi_{k,j} = \prod_{l=0}^k \lambda_{l,j}$. For all j , the set $\{\nu_{l,j} | 1 \leq l \leq Q_j\}$ is assumed to consist of mutually independent random variables. Thus, for all j , $1 \leq j \leq J$, $k = 1, \dots, Q_j$, $\sum_{k=1}^k \xi_{k-1,j} \nu_{k,j} \sim C(\xi_{k,j})$. In particular, if $\lambda_{Q_j,j} = 0$, then $\zeta_j = \sum_{k=1}^{Q_j} \xi_{k-1,j} \nu_{k,j} \sim C(0)$, the Type I extreme-value distribution. Clearly, if all $\nu_{k,j}$'s are mutually independent, then $\zeta_j \stackrel{i.i.d.}{\sim} C(0)$. However, by thinking of $\zeta_j = \sum_{k=1}^{Q_j} \xi_{k-1,j} \nu_{k,j}$ as a simultaneous equations system (or as the stochastic structure of the error terms in a simultaneous equations system) and applying cross-equation restrictions to this system, one can develop a wide variety of complex variance components structures. The terms in the sum are analogous to autoregressive terms in simple time-series models, whereas the ξ 's are factor loadings. Each variance components structure is a representation of a multivariate Type I extreme-value distribution. Each restriction takes the form $\lambda_{l,m} = \lambda_{k,j}$ alone or $\lambda_{l,m} = \lambda_{k,j}$ and $\nu_{l,m} \equiv \nu_{k,j}$. The set of distinct ν 's is assumed to be mutually independent; thus, $\nu_{l,m}, \nu_{k,j}$ are either independent or $\nu_{l,m} \equiv \nu_{k,j}$.⁵

For example, if one sets $Q_1 = Q_2 = Q_3 = 2$, $Q_4 = Q_5 = Q_6 = 3$, $\lambda_{2,1} = \lambda_{2,2} = \lambda_{2,3} = \lambda_{3,4} = \lambda_{3,5} = \lambda_{3,6} = 0$, $\lambda_{1,1} = \lambda_{1,2} = \lambda_{1,4} = \lambda_{2,5}$, $\nu_{1,1} \equiv \nu_{1,2} \equiv \nu_{1,4} \equiv \nu_{2,5}$, $\lambda_{1,3} = \lambda_{2,4} = \lambda_{1,5}$, $\nu_{1,3} \equiv \nu_{2,4} \equiv \nu_{1,5}$, $\nu_{2,1} \equiv \nu_{3,6} = \epsilon_1$, $\nu_{2,2} = \epsilon_2$, $\nu_{2,3} \equiv \nu_{3,4} = \epsilon_3$, and $\nu_{3,5} = \epsilon_4$, then the following variance components structure results:

$$\zeta_1 = \nu_{1,1} + \lambda_{1,1} \nu_{2,1} = \nu_{1,1} + \lambda_{1,1} \epsilon_1,$$

$$\zeta_2 = \nu_{1,2} + \lambda_{1,2} \nu_{2,2} = \nu_{1,1} + \lambda_{1,1} \epsilon_2,$$

$$\zeta_3 = \nu_{1,3} + \lambda_{1,3} \nu_{2,3} = \nu_{1,3} + \lambda_{1,3} \epsilon_3,$$

$$\zeta_4 = \nu_{1,4} + \lambda_{1,4} \nu_{2,4} + \lambda_{1,4} \lambda_{2,4} \nu_{3,4} = \nu_{1,1} + \lambda_{1,1} \nu_{1,3} + \lambda_{1,1} \lambda_{1,3} \epsilon_3,$$

$$\zeta_5 = \nu_{1,5} + \lambda_{1,5} \nu_{2,5} + \lambda_{1,5} \lambda_{2,5} \nu_{3,5} = \nu_{1,3} + \lambda_{1,3} \nu_{1,1} + \lambda_{1,3} \lambda_{1,1} \epsilon_4,$$

$$\zeta_6 = \nu_{1,6} + \lambda_{1,6} \nu_{2,6} + \lambda_{1,6} \lambda_{2,6} \nu_{3,6} = \nu_{1,6} + \lambda_{1,6} \nu_{2,6} + \lambda_{1,6} \lambda_{2,6} \epsilon_1.$$

Multivariate Type I extreme-value distributions may also be developed by using the maximum operation to combine distinct Type I extreme-value random variables. Such distributions occur naturally in competing hazard models of duration data and switching regression models. Using both multiple variance components structures and a maximum operation to combine them yields a generalization of the GEV model of McFadden (1981, 1984; see also Tawn, 1990). The distributions of the maxima are also needed to develop the probability formulae for discrete choice models such as nested logit. To demonstrate that these structures produce Type I extreme-value variates, the following theorem is needed.

THEOREM 2.4. *If ϵ, ν are i.i.d. extreme value ($C(0)$), then for any constants a and b ,*

$$t \equiv \max(a + \epsilon, b + \nu) - \log(e^a + e^b) \sim C(0),$$

the probability that $a + \epsilon > b + \nu$ is $e^a / (e^a + e^b)$ and $t | (a + \epsilon > b + \nu) \sim C(0) - t | (b + \nu > a + \epsilon)$.

Theorem 2.4 is a slight generalization of the fundamental theorem that leads to multinomial logit (McFadden, 1974).

COROLLARY 2.1. *If ϵ_j i.i.d. extreme value, $j = 1, \dots, J$, then, for any constants a_j , we have the following:*

1. $t_j = (\max_{1 \leq l \leq j} (a_l + \epsilon_l) - \log(\sum_{l=1}^j e^{a_l}))$ is extreme value.
2. Probability $(a_j + \epsilon_j \geq a_l + \epsilon_l, \text{ all } l, 1 \leq l \leq j) = (e^{a_j}) / (\sum_{l=1}^j e^{a_l})$.
3. $(t_j | a_j + \epsilon_j \geq a_l + \epsilon_l, \text{ all } l, 1 \leq l \leq j)$ is extreme value.

Note that part 2 of Corollary 2.1 is the multinomial logit model, whereas parts 1 and 3 can be combined with Theorem 2.3 to generalize from the multinomial logit model to the nested logit model. Applying parts 1 and 3 of Corollary 2.1 recursively yields the ranked logit model (Beggs, Cardell, and Hausman, 1981).

3. THE NESTED LOGIT MODEL

In both multinomial logit and nested logit models, the total stochastic term for each alternative is assumed to obey a Type I extreme-value distribution. In multinomial logit, the stochastic terms are independent, whereas in the nested logit model the alternative specific stochastic terms are correlated via a special case of the variance components structure described in Section 2 (Cardell, 1975, 1989). Disjoint subsets of the set of alternatives can each have a common variance component, and each of these subsets can have one or more disjoint subsets, each with a further common variance component. This process can be repeated indefinitely.

Assume that the utility to choosing the j th alternative (U_j) has a systematic component that is linear in the independent variables. (To allow for a nonlinear functional form, one need only replace $X_j \beta$ with $f(X_j, \beta)$ throughout.) In this section, the j th random variable in the variance components structure will be an additive stochastic term in the utility to the j th alternative. Using $\lambda_{k,j}$ and $\nu_{k,j}$ as in Section 2, let $\xi_{l,j} = \prod_{k=0}^l \lambda_{k,j}$. The utility to choosing the j th alternative is

$$U_j = X_j \beta + \sum_{l=1}^{Q_j} \xi_{l-1,j} \nu_{l,j}. \quad (2)$$

The only legal restrictions for the nested logit model involved $\lambda_{k,j}$ and $\lambda_{k,m}$ and possibly $\nu_{k,j}$ and $\nu_{k,m}$. No restrictions are allowed between $\lambda_{k,j}$ and $\lambda_{l,m}$ for $k \neq l$. Furthermore, if $\nu_{k,j} \equiv \nu_{k,m}$, then, for $1 \leq l \leq k$, $\nu_{l,j} \equiv \nu_{l,m}$. Similarly, if $\lambda_{k,j} = \lambda_{k,m}$ is a restriction, then so is $\lambda_{l,j} = \lambda_{l,m}$ for $1 \leq l \leq k$. Of course, the nested logit model is based on Type I extreme-value additive stochastic terms; thus, $\lambda_{Q_j,j} \equiv 0$.

Using the notation developed above, the general form of the nested logit model is

$$U_j = X_j\beta + \epsilon_j, \tag{3}$$

$$\epsilon_j = \sum_{l=1}^{Q_j} \xi_{l-1,j} \nu_{l,j},$$

where

1. $\nu_{l,j}$ is distributed $C(\lambda_{l,j})$,
2. U_j is the utility to choosing the j th alternative (and the j th alternative is chosen iff $U_j \geq U_l$, for all l , $1 \leq l \leq J$),
3. X_j is the vector of independent variables associated with alternative j , and
4. β is the vector of utility function coefficients.

Depending on the situation, any of U_j , X_j and β may be considered to be also indexed by a suppressed individual index. The nested logit model could be combined with the hedonic demand model (Cardell and Dunbar, 1980; Cardell, 1989), in which case β would be a stochastic vector with the parameters of its distribution the estimable parameters of the model.

The distribution of ϵ_j is Type I extreme value, the same distribution as for multinomial logit. Thus, from (3) one can see that the preceding assumptions define a variance components structured multinomial logit model. The name "nested logit" has sometimes led to an inappropriate presumption that the group of alternatives is chosen first and the alternative within the group is chosen last. Such hierarchical models are, in fact, quite different (McFadden, 1981; Tversky, 1972). The original name for the nested logit model was "non-independent logit" (Cardell, 1975). Because each variance component can be associated with a node in a tree, variance components structure (3) is a tree-type structure. McFadden (1981) introduced the name "tree extreme value" for nested logit models. The appropriateness of this name can be made more evident by rewriting (3) as

$$\epsilon_j = \nu_{1,j} + \lambda_{1,j}(\nu_{2,j} + \lambda_{2,j}(\dots + \lambda_{Q_j-1,j}\nu_{Q_j,j})\dots). \tag{4}$$

For each variance component, $\nu_{l,j}$, a set $S_{l,j}$ can be defined consisting of all alternatives that contain $\nu_{l,j}$ in their variance components structure. In other words, this is the set of alternatives grouped with j at the l th level. For example, if $\nu_{2,1} \equiv \nu_{2,3}$ is a constraint and there is no other constraint

involving $\nu_{2,1}$ (or $\nu_{2,3}$), then $S_{2,1} = S_{2,3} = \{1,3\}$. To locate the nodes below l,j , create $G_{l,j}$ from $S_{l,j}$ by removing from $S_{l,j}$ all m such that $\nu_{l+1,m} \equiv \nu_{l+1,j}$ and then selecting a new index j^* , $j^* \in S_{l,j}$. Using the new index, repeat the removal operation, continuing to select new indices, until, for all $m,n \in G_{l,j}$, $m \neq n$, and $\nu_{l+1,m}$ is not equivalent to $\nu_{l+1,n}$. For example, if $j = 1, 2, 3, 4$, and 5 represent drive alone to work, carpool to work, take a bus to work, take a subway to work, and walk to work, respectively, then one logical approach would be to select $Q_1 = Q_2 = Q_3 = Q_4 = 3$, $Q_5 = 1$ and the constraints $\nu_{1,1} \equiv \nu_{1,2} \equiv \nu_{1,3} \equiv \nu_{1,4}$, $\lambda_{1,1} \equiv \lambda_{1,2} \equiv \lambda_{1,3} \equiv \lambda_{1,4}$, $\nu_{2,1} \equiv \nu_{2,2}$, $\nu_{2,3} \equiv \nu_{2,4}$, $\lambda_{2,1} \equiv \lambda_{2,2}$, and $\lambda_{2,3} \equiv \lambda_{2,4}$.⁶ Then, $S_{1,1} = \{1,2,3,4\}$, but $G_{1,1} = \{1,3\}$. In this example, $\nu_{1,1}$ is a common component for commuting in a vehicle, $\nu_{2,1}$ is a common component for commuting by car, and $\nu_{2,3}$ is a common component for commuting by public transit. The tree graph of this example is shown in Figure 1.

Let $K = \max(Q_j)$ and define $Z_{l,j}$ by

$$Z_{l,j} = \begin{cases} e^{X_j\beta/\xi_{l-1,j}} & \text{for } Q_j \leq l \leq K, \\ \sum_{m \in G_{l,j}} Z_{l+1,m}^{\lambda_{l+1,m}} & \text{for } Q_j > l \geq 0. \end{cases} \tag{5}$$

For all $0 \leq l < Q_j$, define $U_{l,j,k} = (X_k\beta + \sum_{j=l+1}^{Q_j} \xi_{m-1,k} \nu_{m,k})/\xi_{l,k}$, $0 \leq l < Q_j$, and $U_{l,j} = \max_{k \in S_{l,j}}(U_{l,j,k})$ for $0 \leq l < Q_j$. Note that $U_j = U_{0,j,j}$. This sets the stage for the following theorem, which is the fundamental theorem of nested logit.

THEOREM 3.1. $U_{l,j} - \log Z_{l,j}$ is an extreme-value random variable. The probability that $U_{l,j,j} = U_{l,j}$, $0 \leq l \leq Q_j$ is $P_{l,j,j} = (\prod_{k=l+1}^{Q_j} Z_{k,j}^{\lambda_{k,j}}) / (\prod_{k=j}^{Q_j-1} Z_{k,j})$. The probability that $U_j > U_l$, all $l \neq j$, is $P_j = (\prod_{k=1}^{Q_j} Z_{k,j}^{\lambda_{k,j}}) / (\prod_{k=0}^{Q_j-1} Z_{k,j})$.

Because they are relatively easy to compute, limited information maximum likelihood (LIML) estimates of nested logit models are quite common.⁷ However, the parameters estimated at a given stage are based only on the identifying variation between alternatives that are grouped together at that level of the tree. When that variation is a small part of the total, the LIML estimates will have high variances and will not be robust to specification errors, a well-known phenomenon in the linear regression literature. The solution in linear regression is to specify an explicit variance components structure and use all the variation to compute an asymptotically efficient estimator (see, e.g., Swamy, 1974; Mundlak, 1978). Similarly, with nested logit, the full information maximum likelihood (FIML) estimates use all the variation between all the alternatives and are asymptotically efficient. Furthermore, FIML estimates do not require that $\lambda_{l,j} \equiv \lambda_{l,k}$. These questions are further addressed in Cardell (1989). Cardell (1989) and Cardell and Steinberg (1992) discuss one practical method for computing FIML estimates and a

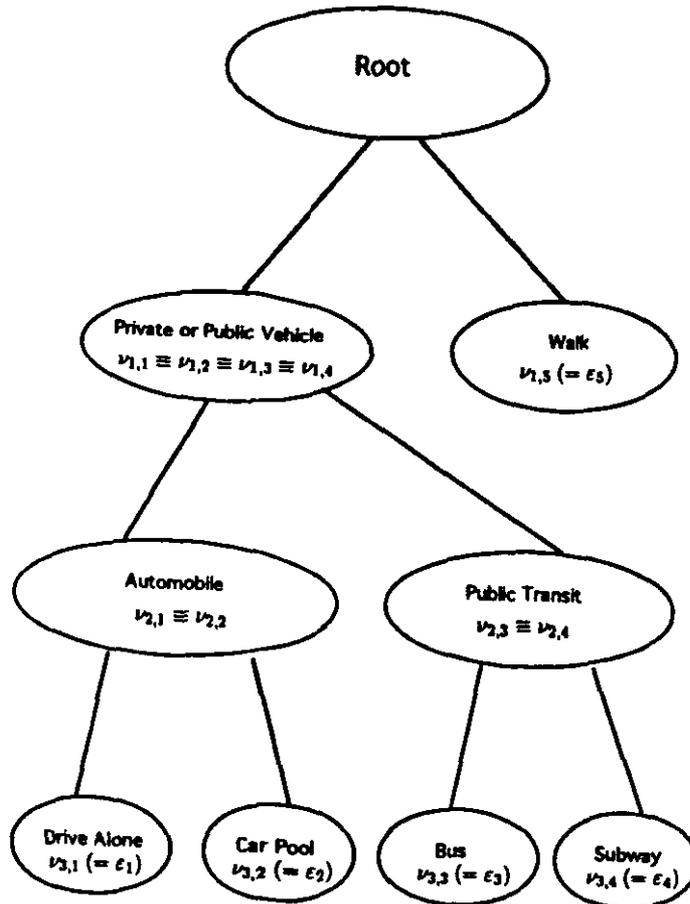


FIGURE 1. Tree diagram for a nested logit model of travel mode choice.

simple method for computing “one-step” estimators that are asymptotically equivalent to FIML.

4. THE MULTINOMIAL c.d.f.’s

4.1. Structures with a Single Common Variance Component

Researchers are often interested in only a single common variance component. Most often only a single heterogeneity factor is used when modeling

duration data due to the inherent complexity. In general, a single common variance component is a statistically useful alternative to a null hypothesis of independent stochastic terms.

THEOREM 4.1. For $0 < \lambda < 1$, $\theta \sim C(\lambda)$, v_j distributed extreme value; θ, v_1, \dots, v_j independent, $\epsilon_j = \theta + \lambda v_j$. The c.d.f. of $\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \end{pmatrix}$ is $F_{\lambda,j}(\epsilon_1, \dots, \epsilon_j) = e^{-\left(\sum_{j=1}^j e^{-\epsilon_j/\lambda}\right)^\lambda}$.

Johnson and Kotz (1972) list three possible forms, including $F_{\lambda,2}$ for the limiting joint distribution of an appropriate linear transformation of $X_{\text{Max}} = \max(X_i)$ and $Y_{\text{Max}} = \max(Y_i)$ for (X_i, Y_i) i.i.d. pairs of random variables with some continuous joint c.d.f. In their typology, $F_{\lambda,2}$ is the Type B bivariate extreme-value distribution. It follows that this distribution is max-stable. Thus, for ordered pairs $(\epsilon_{j1}, \epsilon_{j2})$ distributed with c.d.f. $F_{\lambda,2}$, with $(\epsilon_{j1}, \epsilon_{j2})$ independent of the set of all $(\epsilon_{j1}, \epsilon_{j2})$'s, $j \neq l$, then for any positive integer N $(\max(\epsilon_{j1}), \max(\epsilon_{j2}))$ is distributed as $F_{\lambda,2}$ plus a constant vector (equal to $(\lambda \log N, \lambda \log N)$). By comparison, for $(\epsilon_{j1}, \epsilon_{j2})$, distributed $N(\mu, \Sigma)$ with any nonsingular Σ , the limiting joint distribution of $\text{Max}(\epsilon_{j1}), \text{Max}(\epsilon_{j2})$ has $\text{Max}(\epsilon_{j1})$ independent of $\text{Max}(\epsilon_{j2})$. These results generalize directly from the bivariate to the multinomial case.

THEOREM 4.2. For $0 < \lambda < 1$, $\theta \sim GL(\lambda)$, $v_j \sim \text{logistic}$; θ, v_1, \dots, v_j independent, $\epsilon_j = \theta + \lambda v_j$. Then, the c.d.f. of the random vector $\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \end{pmatrix}$ is

$$F(\epsilon_1, \dots, \epsilon_j) = \sum_{l=1}^L \sum_{n=1}^{m_l} C_{j,n} \left(\frac{1}{1 + e^{-\epsilon_l}} \right)^n,$$

where $\epsilon_{j1}, \dots, \epsilon_{jL}$ are the distinct ϵ_j 's, m_l is the multiplicity of ϵ_{jl} , and the $C_{j,n}$'s are easily computed recursively.

Note that for ϵ_j all distinct, Theorem 4.2 reduces to

$$F(\epsilon_1, \dots, \epsilon_j) \equiv \sum_{j=1}^j \left(\prod_{l \neq j} \left(\frac{e^{-\epsilon_j/\lambda}}{e^{-\epsilon_j/\lambda} - e^{-\epsilon_l/\lambda}} \right) \frac{1}{1 + e^{-\epsilon_j}} \right).$$

4.2. Tree-Type Variance Components Structures

Theorem 4.1 can be used to develop the c.d.f. for variance components structure (3).

LEMMA 4.1. For $\epsilon_{i,j}$ defined by variance components structure (3),

$$F(\epsilon_1, \dots, \epsilon_j) = e^{-\left(\sum_{i \in G_{h,1}} \left(\sum_{i \in G_{i,1}} \left(\dots \left(\sum_j e^{-\epsilon_j/\lambda_{i,j}}\right)^{\lambda_{i,j-1}} \dots\right)^{\lambda_{i,2}}\right)^{\lambda_{i,3}}\right)^{\lambda_{i,4}}}$$

In the nested logit model, $\lambda_{0,j} \equiv 1$ and $\nu_{0,j} \equiv 0$, because a variance component common to all alternatives will not affect a discrete choice model. However, such components are important in other circumstances, as, for example, in hazard function models. In such cases, (3) can be replaced by

$$\epsilon_j = \sum_{i=0}^{Q_j} \xi_{i-1,j} \nu_{i,j} = \nu_{0,j} + \sum_{i=1}^{Q_j} \xi_{i-1,j} \nu_{i,j}, \tag{6}$$

with $\xi_{-1,j} \equiv 1$, $0 < \lambda_{0,j} < 1$. In equations containing cardinal information, ϵ_j may also have a free scale parameter.

Tree-type variance components structures are also potentially useful for logistic random variables. The c.d.f.'s for logistic random variables with such tree-type variance components structures are easily derived using the same method used to prove Theorem 4.2.⁸ While such tree-type structures have only positive correlations, structures with negative correlations are easily derived from tree-type logistic structures. The logistic and generalized logistic distributions are symmetric. Thus, let d_j be a fixed scalar equal to ± 1 , depending on the variance components structure to be developed, let ϵ_j be as in (3), and let $\eta_j = d_j \epsilon_j$. Then, $\text{Var}(\eta)$ can include negative correlations and the c.d.f. of (η_1, \dots, η_j) is easily calculated from the c.d.f.'s of $(\epsilon_1, \dots, \epsilon_j)$ and subsets of the ϵ 's. Huang and Cardell (1996) use this approach to investigate a heterogeneity factor that affects durations on and off welfare in opposite ways.

Tree-type variance components structures are a natural and parsimonious way to allow for stochastic correlation with a small number of parameters. Often, the variance components themselves can be identified with unmeasured factors that, from economic theory, should be important. Furthermore, in hazard functions with heterogeneity, variance components structures seem to be the only feasible way to include unmeasured factors. In most duration data, different individuals begin in different states and go through different sequences of states (i.e., spell types). Truly free correlations would require a different matrix for every possible sequence of states, frequently involving more free parameters than could possibly be identified. For example, in a data set in which 4% of the individuals have 20 or more spells, estimation of even a single 21×21 covariance matrix would require estimating 210 correlation parameters. By contrast, heterogeneity has in practice usually been represented by only a single variance component (see, e.g., Heckman and Singer, 1986).

In some situations with a limited number of stochastic components and a large number of observations, one might wish to be able to approximate a free correlation matrix. While I find it implausible that this generalization would make the best use of finite data (as compared to, e.g., generalizing $X\beta$ to $f(X, \beta)$), others may consider the capability to approximate free correlation matrices important. Therefore, I generated all 4×4 correlation matrices

that follow from $\Gamma' \Gamma = R$ with Γ upper triangular and the off-diagonal elements of Γ a multiple of .2. For each one, the variance components structure from (6) that could most closely approximate R was found. For this case, the variance components structure has three or fewer unknown continuous parameters, whereas R has six correlations to be approximated. The overall mean squared error of \hat{R}_{ij} , $i \neq j$, was .01. (The largest errors occur when two of the stochastic terms are each highly correlated with a third term but uncorrelated or only slightly correlated with each other.) Thus, in this simulation experiment, the variance components structures of this paper provided a reasonable approximation to a free correlation matrix.

5. OTHER APPLICATIONS AND EXTENSIONS

5.1. Correlated Discrete Decisions

When the same individual makes multiple discrete decisions, these choices may be correlated. Although the nested logit approach is not a useful framework for handling this situation, the same type of variance components structure can be used, but with a logistic rather than a Type I extreme-value stochastic distribution. Let Y_k be an indicator variable and d_k be a discrete response variable, where $0 \leq \lambda \leq 1$, $v \sim GL(\lambda)$, and η_k - logistic.

$$Y_k = X_k \beta_k + v + \lambda \eta_k.$$

$$d_k = \begin{cases} 1 & \text{if } Y_k \geq 0, \\ 0 & \text{if } Y_k < 0. \end{cases}$$

X_k is a row vector of characters and β_k is a column vector of unknown parameters related to the k th binary decision. This variance components structure is a natural way to include a common interdecisions stochastic correlation in the case of multiple binary decisions. Obviously, the marginal probability of $d_k = 1$ is $P_k = 1 / (1 + e^{-X_k \beta_k})$. Therefore, a single model can incorporate data on individuals who face multiple decisions and also on other individuals who face only a single decision, allowing coefficients to be directly compared between a study of correlated decisions and a simple binomial logit study.

Applying Theorem 4.2, the joint probability of $d_k = 1$ and $d_l = 1$ for $l \neq k$ is $P_{kl} = F(X_k \beta_k, X_l \beta_l)$, where

$$P_{kl} = \begin{cases} \frac{e^{X_k \beta_k / \lambda}}{e^{X_k \beta_k / \lambda} + e^{X_l \beta_l / \lambda}} \times \frac{1}{1 + e^{-X_l \beta_l}} - \frac{e^{X_l \beta_l / \lambda}}{e^{X_k \beta_k / \lambda} + e^{X_l \beta_l / \lambda}} \times \frac{1}{1 + e^{-X_k \beta_k}} & \text{if } X_k \beta_k \neq X_l \beta_l, \\ (1 - \lambda) \frac{1}{1 + e^{-X \beta}} + \lambda \left(\frac{1}{1 + e^{-X \beta}} \right)^2 & \text{if } X_k \beta_k = X_l \beta_l = X \beta. \end{cases}$$

The formulae for the joint probability of multiple binary decisions can be determined from Theorem 4.2 or the extensions to it already discussed. See Cardell, Huang, and Brown (1995) for an empirical study of correlated binary decisions. Some correlated multinomial cases can be solved in the same way. See Cardell (1989) for a multinomial case with X fixed across individuals.

5.2. Hazard Functions with Heterogeneity

Econometric modeling of duration data is based on both hazard and survival functions. The hazard function $h(t, X)$ is defined as the probability of a spell ending per unit time, conditional on it not having ended previously. The survival function $S(t, X)$ is defined as the probability that a spell will last at least a time of t . The definitions imply that $h(t, X) = -[d \log(S(t, X))] / dt$ or that $S(t, X) = e^{-\int_0^t h(\tau, X) d\tau}$.

For a proportional hazard function, the hazard $h(t, X) = g(t)e^{-X\beta}$, where g is a function that may or may not be parametrically specified, β is a column vector of unknown parameters, and X is a row vector of time-invariant independent variables. The corresponding survival function to $h(t, X)$ is $S(t, X) = e^{-\int_0^t X\beta G(\tau) d\tau}$, where $G(t) = \int_0^t g(\tau) d\tau$. Thus, $-\log(-\log(S(t, X))) = X\beta - \log G(t)$ is distributed extreme value for $t =$ time to transition, or $\log G(t) = X\beta - \eta$, where η is an extreme value stochastic term.

5.3. Proportional Hazards with $C(\lambda)$ Heterogeneity

Let $\eta_i = v + \varepsilon_i$, where $v \sim C(\lambda)$, $\varepsilon_i \sim C(0, \lambda)$, and v, ε_i independent. Then, the conditional survivor function is $S(t_i, X, v) = e^{-e^{-X\beta/\lambda} G(t_i)^{1/\lambda} e^{v/\lambda}}$, which is also the proportional hazard form with v heterogeneity. From Theorem 4.1, the multiple survival function for repeated proportional hazards with $C(\lambda)$ heterogeneity is

$$S(t_1, X_1, t_2, \dots, t_N, X_N) = e^{-\left(\sum_{i=1}^N e^{-X_i\beta/\lambda} G_i(t_i)^{1/\lambda}\right)^\lambda} \quad (7)$$

where, in general, β_i and G_i may or may not be distinct for distinct i . These models can be used in such situations as heterogeneity combined with repeated hazards, sample selection modeling, and switching regression.

The survival function in (7) has been used in the biometric literature, though generally restricted to the special case $G_i(t_i) = a_i t_i^{b_i}$, the Weibull hazard function (Hougaard, 1986, equation 3.1; Oakes and Manatunga, 1992, equation 1; Tawn, 1990, equations 2.3, 3.2, and 3.8; Tawn, 1988, equation 5.2). Hougaard (1986) also discusses more general $G_i(t_i)$. Other related distributions in the biometric literature correspond to variance components structures with two heterogeneity factors affecting a given duration (Hou-

gaard, 1986, Sect. 7; Tawn, 1990, equations 3.3 and 3.10). Other bivariate distributions in Johnson and Kotz (1972) are discussed in Tawn (1988).⁹ The general multifactor case given in Lemma 4.1 is discussed in Cardell (1989).

5.4. Hazards with $GL(\lambda)$ Heterogeneity

Logit-based variance components structures allow for hazards that may be either positively or negatively correlated between distinct spell types. The log-logistic hazard function model is widely available in standard packages for the log-logistic model $S(t, X) = 1/(1 + t^\alpha e^{-X\beta})$ (e.g., SAS[®] and LIM-DEP[®]). Although not a type of proportional hazard model, the logistic model is similar and closely related. Consider the following model of durations based on the survival function $S(t, X; \beta, \delta)$:

$$\frac{d}{dt} \left(\frac{S^{-\delta} - 1}{\delta} \right) = g(t)e^{-X\beta} \quad \text{or} \quad \frac{S^{-\delta} - 1}{\delta} = G(t)e^{-X\beta}, \quad (8)$$

where $S(t, X; \beta, 0) = \lim_{\delta \rightarrow \infty} S(t, X; \beta, \delta)$ is a proportional hazard model. Setting $\delta = 1$ and $G(t) = t^\alpha$ yields the standard log-logistic model. In general, $-\log((S^{-\delta} - 1)/\delta)$ has a Verhulst (1845, as cited in Ahuja and Nash, 1967) distribution. Thus, let $\varepsilon = -\log((S^{-\delta} - 1)/\delta)$ and $\log G(t) = X\beta - \varepsilon$, with ε Verhulst.

Therefore, both the proportional hazard model and the log-logistic model are special cases of (8). The log-logistic model can be generalized by allowing $G(t)$ to be nonparametric, as in Han and Hausman (1990). Then, $\log(S/(1 - S)) = X\beta - \log G(t) \sim GL(0)$. Thus, $\log G(t) = X\beta + \eta$ with η logistic. Note that a test of $H_0: \delta = 0$ in (8) is one way to test the proportional hazards assumption. Similarly, a test of $H_0: \delta = 1$ in (8) is a test of the log-logistic duration model.

Given multiple spells (indexed by i) in a generalized log-logistic model, the η 's can be given a variance components structure. The joint survival function is then

$$S(t_1, X_1, \dots, t_N, X_N) = F(X_1\beta - \log G(t_1), \dots, X_N\beta - \log G(t_N)),$$

where F is the joint c.d.f. of the η 's discussed in the previous section. If N completed spells of length t_i are observed, the likelihood of the observation is

$$(-1)^N \frac{\partial^N S(t_1, X_1, \dots, t_N, X_N)}{\partial t_1 \cdots \partial t_N}.$$

If the last spell is right truncated at t_N , then the likelihood of the observation is

$$(-1)^{N-1} \frac{\partial^{N-1} S(t_1, X_1, \dots, t_N, X_N)}{\partial t_1 \cdots \partial t_{N-1}}.$$

Thus, the explicit joint survival functions derived here can be used to estimate the structural coefficient vector β , the variance components parameters, and $G(t)$ by FIML. Note that a completed spell of length t_1 and a spell truncated at time t_2 is the combination of a continuous duration result of t_1 and a discrete event of spell 2 "surviving" to time t_2 ; that is,

$$\log(G(t_2)) < X_2\beta_2 - \eta_2. \tag{9}$$

However, one can replace (9) with an indicator equation for any other binary event. In particular, selection biases can be dealt with in exactly the same framework.

5.5. Competing Hazards

In the competing hazards situation, the likelihood that a spell ends at time t_j by cause j of N competing causes is

$$\frac{\partial S(t_1, X_1, \dots, t_N, X_N)}{\partial t_j} \Big|_{t_1=t_2=\dots=t_N}$$

An important special case is the competing proportional hazards model, with $C(\lambda)$ heterogeneity, $G_i \equiv G$, and $\beta_i \equiv \beta$. From (7) the probability that the spell will end due to cause j is $P_j = (e^{-X_j\beta/\lambda}) / (\sum_{i=1}^N e^{-X_i\beta/\lambda})$, a multinomial logit probability. Furthermore, P_j is independent of the length of the spell. Thus, one can estimate λ , β , and G by a LIML method. In the first stage, β/λ can be consistently estimated using a conventional multinomial logit program. Then, λ and G can be consistently estimated using $\log(\sum_{i=1}^N e^{-X_i(\beta/\lambda)})$ as an independent variable in a conventional proportional hazards program. This procedure is obviously very similar to LIML estimation of nested logit, and, as with nested logit, coefficients unidentified at the first stage can be estimated at the second. Repeated and competing hazards can be combined in the obvious way.

Consider the practical competing hazards situation where the hazards apply to spells of working and not working. Let the working/not working dichotomy define the spells in a repeated hazards framework. Assume that some, but not all, factors that lead to longer (shorter) working spells also lead to shorter (longer) nonworking spells. Let s index spell type, with 1 = working and 2 = nonworking, and let i index the number of spells of a given type. Then,

$$\log(G(t_{is})) = X_{is}\beta_s + (-1)^{s-1}\nu_0 + \lambda_0\nu_s + \lambda_0\lambda_s\varepsilon,$$

where $\nu_n \sim GL(\lambda_n)$, and $\varepsilon \sim$ logistic represents the desired variance components structure. The joint survival function can be calculated using the results in Section 4.

5.6. Models with Discrete Endogenous Equations

Let $Y = X\beta + \alpha\varepsilon$, $Z = X\gamma + \nu + \lambda\varepsilon$, and $d = 1$ if $Z > 0$, where $\nu \sim GL(\lambda)$, ε logistic, and ν , ε independent. $0 < \lambda < 1$ is an unknown scalar, β and γ are unknown parameter vectors, and α is an unknown scale factor. From Theorem 2.2, the likelihood of an observation d_i, Y_i, X_i is then

$$\frac{\left(\tan^{-1}\left(\frac{\sin \pi \lambda}{e^{-X_i\gamma - \lambda(Y_i - X_i\beta)/\alpha} + \cos \pi \lambda}\right)\right)^{d_i} \left(\tan^{-1}\left(\frac{\sin \pi \lambda}{e^{+X_i\gamma + \lambda(Y_i - X_i\beta)/\alpha} + \cos \pi \lambda}\right)\right)^{1-d_i}}{\alpha(e^{(Y_i - X_i\beta)/\alpha} + 2 + e^{-(Y_i - X_i\beta)/\alpha})\pi \lambda} \tag{10}$$

This formula can form the basis for a FIML estimation of a logit-based version of Heckman's (1976, 1978) simultaneous equation system with dummy endogenous variables. If Y_i is not observed for $d_i = 0$, the endogenous discrete equation becomes a selection equation and the likelihood function for $d_i = 0$ is $1/(1 + e^{X_i\delta})$. Combining this function with (10) for $d_i = 1$ allows FIML estimation of a logit-based selection bias model. The previous subsection demonstrated how to account for selection bias in modeling duration data. Clearly, the same approach can be followed for any combination of duration equations, linear equations, endogenous discrete equations, and selection equations.

6. CONCLUSIONS

The variance components structures developed in this paper allow parsimonious modeling of heterogeneity in a wide variety of contexts. In general, one may have $Y_j = X_j\beta_j + \sum_{i=0}^2 \xi_{i-1,j}\nu_{i,j}$, for $1 \leq j \leq J$, where each Y may be an ordinary continuous variable, a latent variable that determines a binary outcome, a (possibly nonparametric) transformation of a (possibly latent) spell duration, or even the utility to a discrete choice alternative. Furthermore, different Y 's may be of different types. Thus, one may address heterogeneity of many types in a unified framework. For instance, a single model could include competing repeated hazard functions with heterogeneity, selection bias, and endogenous discrete equations.

Because the variance components structures developed in this paper have relatively simple and differentiable c.d.f.'s, the models generated from them can be easily estimated by FIML. While the $C(\lambda)$ and $GL(\lambda)$ distributions allow more general variance components structures than those discussed here, the question of whether any other structures, aside from a few simple cases, have closed-form joint c.d.f.'s is still open.

The variance components structures in this paper are for stochastic terms that are Type I extreme value or logistic. These distributions have been found to be very useful in modeling discrete decisions and duration data. The nested

Other models based on these structures will likely become common in the future.

NOTES

1. Cardell (1989) estimated models with more than 100 alternatives.

2. See *Review of Economics and Statistics* 76, November 1994, for a special section on simulation methods in econometrics. See also Cardell (1989, Ch. II).

3. The logistic distribution is symmetric. Thus, the generalized logistic distribution is symmetric and the distributions of bc and $-bc$ are identical.

4. Elementary techniques do not appear to be able to generate this result. I would welcome a more direct proof of (1).

5. If $v_i^m = v_{k,j}$, then (i,m) and (k,j) become in effect the same index. Thus, if \cdot is a mapping from $i \times l$ to j defined such that $i \cdot m = k \cdot j$ iff $v_i^m = v_{k,j}$, then $k \cdot j$ can be used as a single index. While elegant and parsimonious, this approach has proven to confuse readers. For computer implementations based on these variance component structures, it is convenient to define an array $f(k,j) = k \cdot j$ as earlier. A version of this paper that uses the $k \cdot j$ notation is available from the author on request.

6. Users of limited information maximum likelihood, or LIML, would need to add $\lambda_{2,2} = \lambda_{2,3}$.

7. Details of the LIML method using the notation in this paper are available from the author on request.

8. Further details are available from the author on request.

9. Tawn (1988) is interested in random variables with marginal exponential distributions. Johnson and Kotz (1972) base their bivariate distributions of random variables on a Type I extreme-value distribution. The bivariate distributions in Tawn (1988) can be generated from (8) in Johnson and Kotz (1972) and conditions (9) by setting $x = e^{-x_1}$ and $y = e^{-x_2}$. Thus, the random variables in Johnson and Kotz (1972) correspond to minus the log of those in Tawn (1988). For example, on page 401 of Tawn (1988), $P(X \geq x, Y \geq y) = e^{-\frac{1}{2}(x+y) + ((\theta y)/(x+y))} = P\{X \leq x^*, Y^* \leq y^*\} = F_{X^*, Y^*}(x^*, y^*) = e^{-e^{-x^*} - e^{-y^*} + \theta e^{-x^*} e^{-y^*}}^{-1}$ (for $X^* = -\log(X)$, $x^* = -\log(x)$, $Y^* = -\log(Y)$, and $y^* = -\log(y)$), which is (11), on page 252, of Johnson and Kotz (1972), to which I have added *s to distinguish the variables in Johnson and Kotz (1972) from those in Tawn (1988).

REFERENCES

Ahuja, J. C. & S. W. Nash (1967) The generalized Gompertz-Verhulst family of distributions. *Sankhya: The Indian Journal of Statistics: Series A* 29, 141-156.
 Beggs, S. D., N. S. Cardell, & J. Hausman (1981) Assessing the potential demand for electric cars. *Journal of Econometrics* 17, 1-19.
 Cardell, N. S. (1975) A Statistical Test of the Independence of Irrelevant Alternatives Property of Multinomial Logit. Working paper, Charles River Associates, Massachusetts.
 Cardell, N. S. (1989) *Extensions of Multinomial Logit: The Hedonic Demand Model, the Non-independent Logit Model, and the Ranked Logit Model*. Unpublished Ph.D. Dissertation, Harvard University, Cambridge, Massachusetts.
 Cardell, N. S. & F. C. Dunbar (1980) Measuring the societal impacts of automobile downsizing. *Transportation Research* 14A, 389-404.
 Cardell, N. S., S. Huang, & S. Brown (1995) Decisionmaking in the motor carrier industry. *Transportation Research* 29A, 401-419.

Cardell, N. S. & D. Sechnberg (1992) A Gauss-Newton Method for Maximum Likelihood Estimation, with an Application to Nested Logit. Working paper, San Diego State University, San Diego.
 Cramer, H. (1937) *Random Variables and Probabilities Distributions*. Cambridge Tracts in Mathematics and Mathematical Physics 36. Cambridge: Cambridge University Press.
 Erdelyi, A. (ed.) (1953) *Higher Transcendental Functions*. Bateman Manuscript Project, vol. 1. New York: McGraw-Hill.
 Feller, W. (1971) *Introduction to Probability Theory and its Applications*, vol. 2, 2nd ed. New York: Wiley & Sons.
 Gradshteyn, I. S. & I. M. Ryzhik (1965) *Table of Integrals, Series, and Products*, 4th ed. New York: Academic Press.
 Gumbel, E. J. (1961) Bivariate logistic distributions. *American Statistical Association Journal* 56, 335-349.
 Han, A. & J. A. Hausman (1990) Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics* 5, 1-28.
 Harvard University (1994) Special section on simulation methods. *Review of Economics and Statistics* 76, 591-702.
 Heckman, J. J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 4, 475-492.
 Heckman, J. J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica* 46, 931-959.
 Heckman, J. J. & B. Singer (1986) Econometric analysis of longitudinal data. In Z. Griliches & M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 3, pp. 1689-1763. New York: North-Holland.
 Hougaard, P. (1986) A class of multivariate failure time distributions. *Biometrika* 73, 671-678.
 Huang, S. & N. S. Cardell (1996) Duration Analysis on Multiple Welfare Spells: A Repeated Hazard Model with Heterogeneity. Working paper, Washington State University, Pullman.
 Johnson, N. L. & S. Kotz (1972) *Distributions in Statistics: Continuous and Multivariate Distributions*. New York: Wiley & Sons.
 Kagan, A. M., Y. V. Linnik, & C. R. Rao (1973) *Characterization Problems in Mathematical Statistics*. New York: Wiley & Sons.
 Lukacs, E. (1970) *Characteristic Functions*, 2nd ed. London: Griffin.
 McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*, pp. 105-142. New York: Academic Press.
 McFadden, D. (1981) Econometric models of probabilistic choice. In C. F. Manski & D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 198-272. Cambridge, Massachusetts: MIT Press.
 McFadden, D. (1984) Econometric analysis of qualitative response models. In Z. Griliches & M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, pp. 1395-1457. New York: North-Holland.
 Mundlak, Y. (1978) On the pooling of time series and cross section data. *Econometrica* 46, 69-85.
 Oakes, D. & A. K. Manjunaga (1992) Fisher information for a bivariate extreme value distribution. *Biometrika* 79, 827-832.
 Swamy, P. A. V. B. (1974) Linear models with random coefficients. In P. Zarembka (ed.), *Frontiers in Econometrics*, pp. 143-168. New York: Academic Press.
 Tawn, J. A. (1988) Bivariate extreme value theory: Models and estimation. *Biometrika* 75, 397-413.
 Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika* 77, 245-253.
 Tversky, A. (1972) Elimination-by-aspects: A theory of choice. *Psychological Review* 79, 281-299.
 Verhulst, P. F. (1845) Recherches mathématiques sur la loi d'accroissement de la population. *Nouvelles Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles* 18, 38 pp.

APPENDIX

Proof of Theorem 2.1. The characteristic function of the Type I extreme-value distribution is

$$\psi_\epsilon(t) = \int_{-\infty}^{\infty} e^{it\epsilon} e^{-\epsilon} e^{-e^{-\epsilon}} d\epsilon = \Gamma(1 - it).$$

$$\log(\Gamma(1 - it)) = \int_0^{\infty} (e^{itu} - 1 - it(1 - e^{-u})) \frac{du}{u(e^u - 1)}$$

(see Gradshteyn and Ryzhik, 1965, p. 934).

Let $a = \int_0^{\infty} \{u/(1 + u^2)\} - (1 - e^{-u}) [du/(u(e^u - 1))]$. Obviously, $-\infty < a < \infty$. Thus, $\log \Gamma(1 - it) = ita + \int_0^{\infty} (e^{itu} - 1 - \{itu/(1 + u^2)\}) [du/(u(e^u - 1))]$. Now, let $N(u) = -\int_u^{\infty} [dx/(x(e^x - 1))]$ and $M(u) \equiv 0$. Thus,

$$\log \psi_\epsilon(t) = ita + \int_0^{\infty} \left(e^{itu} - 1 - \frac{itu}{1 + u^2} \right) dN(u)$$

is the Lévy canonical representation of $\psi_\epsilon(t)$. Furthermore, $M(u)$ and $N(u)$ are nondecreasing over $(-\infty, 0)$ and $(0, \infty)$, respectively, and $M(-\infty) = N(\infty) = 0$. Finally, $0 < \int_0^{\infty} u^2 dN(u) < \int_0^{\infty} 1 du = \epsilon$. Thus, $\int_{-\infty}^0 u^2 dM(u)$ and $\int_0^{\infty} u^2 dN(u)$ are finite for every $\epsilon > 0$. By Theorem 5.5.2, on page 118, in Lukacs's (1970) text, $\psi_\epsilon(t)$ is an infinitely divisible characteristic function. Furthermore, $M'(u) = 0$ and $N'(u) = 1/[u(e^u - 1)]$, and $uM'(u)$ and $uN'(u)$ are nonincreasing for $u < 0$ and $u > 0$, respectively. Therefore, $\psi_\epsilon(t)$ is self-decomposable (Lukacs, 1970, Theorem 5.11.2, p. 164). Thus, from the definition of a self-decomposable characteristic function, $\psi_\nu(t) = [\psi_\epsilon(t)]/[\psi_\epsilon(\lambda t)] = [\Gamma(1 + it)]/[\Gamma(1 - i\lambda t)]$ is a characteristic function. Furthermore, for ν, ϵ independent, $0 < \lambda < 1$, $\epsilon \sim$ Type I extreme value, then $\nu + \lambda\epsilon \sim$ Type I extreme value iff the characteristic function of ν is $[\Gamma(1 - it)]/[\Gamma(1 - i\lambda t)]$. Let $f_\lambda(\nu) = (1/2\pi) \int_{-\infty}^{\infty} e^{-it\nu} [\Gamma(1 - it)]/[\Gamma(1 - i\lambda t)] dt$ and $F_\lambda(\nu) = \int_{-\infty}^{\nu} f_\lambda(y) dy$.

The functions e^z and $1/\Gamma(Z)$ are entire and $\Gamma(Z)$ is analytic everywhere except for simple poles at the nonpositive integers, which have residues $[(-1)^{Z+1}]/[(-Z)!]$ (Erdelyi, 1953). Therefore, consider the contour C_m as defined by the four segments: $\text{Im}(t) = 0, -m \leq \text{Re}(t) \leq m; \text{Re}(t) = -m, 0 \geq \text{Im}(t) \geq -i(m + \frac{1}{2}); \text{Im}(t) = -i(m + \frac{1}{2}), -m \leq \text{Re}(t) \leq m; \text{and } \text{Re}(t) = m, 0 \geq \text{Im}(t) \geq -i(m + \frac{1}{2})$, where m is a positive integer. From the Cauchy Theorem,

$$\oint_{C_m} e^{-it\nu} \frac{\Gamma(1 - it)}{\Gamma(1 - i\lambda t)} dt = \sum_{n=1}^m \frac{(-1)^{n+1} e^{-n\nu}}{(n-1)! \Gamma(1 - \lambda n)} \tag{A.1}$$

Taking the limit of (A.1) as $m \rightarrow \infty$ yields

$$f_\lambda(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-it\nu} \frac{\Gamma(1 - it)}{\Gamma(1 - i\lambda t)} dt = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} e^{-n\nu}}{(n-1)! \Gamma(1 - \lambda n)}$$

$$\equiv \frac{1}{\lambda} \sum_{n=1}^{\infty} \frac{(-1)^n e^{-n\nu}}{n! \Gamma(-\lambda n)} \equiv e^{-\nu} \sum_{n=0}^{\infty} \frac{(-1)^n e^{-n\nu}}{n! \Gamma(1 - \lambda(n+1))} \tag{A.2}$$

The first infinite series can be integrated from ν to ∞ to yield

$$F_\lambda(\nu) = 1 + \sum_{n=1}^{\infty} \frac{(-1)^n e^{-n\nu}}{n! \Gamma(1 - \lambda n)} = \sum_{n=0}^{\infty} \frac{(-1)^n e^{-n\nu}}{n! \Gamma(1 - \lambda n)}.$$

Let $C(\lambda)$ denote the distribution of a random variable that has p.d.f. f_λ and c.d.f. F_λ . Distinct probability distributions have distinct characteristic functions (Feller, 1971, p. 508). Thus, for ν, ϵ independent and $\epsilon \sim$ Type I extreme value, $\nu + \lambda\epsilon \sim$ Type I extreme value iff $\nu \sim C(\lambda)$. For $\lambda = \frac{1}{2}$, (A.2) is easily simplified to $f_{1/2}(\nu) = (e^{-\nu} e^{-\frac{1}{2}\nu^2})/\sqrt{\pi}$. It follows that $F_{1/2}(\nu) = 2(1 - F_N(e^{-\nu}/\sqrt{2}))$, where F_N is the c.d.f. of the standard normal distribution. Therefore, if $x \sim N(0, 2)$, $\nu = -\log |x| \sim C(\frac{1}{2})$. Because $F_{1/2}$ does not have a closed-form representation, F_λ cannot in general have a closed-form representation. ■

Proof of Lemma 2.1. Let $\zeta = \eta_1 - \eta_2$, and let $F(\zeta)$ represent the c.d.f. of ζ . Then,

$$F(\zeta) = \int_{-\infty}^{\infty} e^{-\eta} e^{-e^{-\eta}} e^{-e^{-(\zeta+\eta)}} d\eta = \frac{1}{1 + e^{-\zeta}}. \tag{A.3}$$

From Theorem 2.1, the preceding η_1 and η_2 admit the linear structures: $\eta_j = \nu_j + \lambda\epsilon_j$, $j = 1, 2$, where ν_1, ν_2, ϵ_1 , and ϵ_2 are all mutually independent and $\nu_j \sim C(\lambda)$, $\epsilon_j \sim$ Type I extreme value $j = 1, 2$. Thus, $\eta_1 - \eta_2 = \nu_1 - \nu_2 + \lambda(\epsilon_1 - \epsilon_2)$. From (A.3), $\epsilon_1 - \epsilon_2 \sim$ logistic, and, in addition, $\nu_1 - \nu_2$ and $\epsilon_1 - \epsilon_2$ are independent. Therefore, from the definition of $GL(\lambda)$, $\nu_1 - \nu_2 \sim GL(\lambda)$. Furthermore, the uniqueness of $GL(\lambda)$ follows directly from the uniqueness of $C(\lambda)$. ■

Proof of Lemma 2.2. Let ϵ be an extreme-value random variable independent of ν_1 and ν_2 . Then, $(\nu_2 + \lambda_2\epsilon)$ is an extreme-value random variable and independent of ν_1 . Hence, $\nu_1 + \lambda_1(\nu_2 + \lambda_2\epsilon)$ is extreme value, and $\nu_1 + \lambda_1(\nu_2 + \lambda_2\epsilon) \equiv (\nu_1 + \lambda_1\nu_2) + (\lambda_1\lambda_2)\epsilon$. By Theorem 2.1, $(\nu_1 + \lambda_1\nu_2) \sim C(\lambda_1\lambda_2)$. ■

Proof of Theorem 2.2. The distribution of $z \sim GL(\lambda)$ can be determined by inverting $\Psi_{GL(\lambda)}(t)$, where $\Psi_{GL(\lambda)}(t) = \Psi_{C(\lambda)}(t)\Psi_{C(\lambda)}(-t) = [(\Gamma(1 - it))/(\Gamma(1 - i\lambda t))][(\Gamma(1 + it))/(\Gamma(1 + i\lambda t))]$. Using $\Gamma(1 - Z)\Gamma(1 + Z) \equiv (\pi Z)/(\sin \pi Z)$ (Erdelyi, 1953),

$$\Psi_{GL(\lambda)}(t) = \begin{cases} \left(\frac{1}{\lambda}\right) \frac{e^{\pi\lambda t} - e^{-\pi\lambda t}}{e^{\pi t} - e^{-\pi t}}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

Obviously, $\lim_{t \rightarrow 0} \Psi_{GL(\lambda)}(t) = 1 = \Psi_{GL(\lambda)}(0)$, and $\Psi_{GL(\lambda)}$ is analytic everywhere except for simple poles at $t = ni$ for some integers n . Thus, the p.d.f. of $GL(\lambda)$ is

$$f_{GL(\lambda)}(z) = \frac{1}{2\pi\lambda} \int_{-\infty}^{\infty} e^{-itz} \frac{e^{\pi\lambda t} - e^{-\pi\lambda t}}{e^{\pi t} - e^{-\pi t}} dt. \tag{A.4}$$

Obviously, the right-hand side of (A.4) is a symmetric function of z for z real. Thus, consider $z < 0$, z real. The integrand of (A.4) goes to zero exponentially as

$\text{Im}(t) \rightarrow \infty$, or $|\text{real}(t)| \rightarrow \infty$ for $\text{Im}(t)$ positive. Therefore, the integral in (A.4) is equal to $2\pi i$ times the sum of the residues in the half plane with positive $\text{Im}(t)$. Thus,

$$\begin{aligned} f_{GL(\lambda)}(z) &= (2\pi i) \frac{1}{2\pi\lambda} \sum_{n=1}^{\infty} e^{nz} (-1)^n \frac{1}{2\pi} (e^{i\lambda\pi n} - e^{-i\lambda\pi n}) \\ &= \frac{i}{2\pi\lambda} \sum_{n=1}^{\infty} ((-e^{z+i\lambda\pi})^n - (-e^{z-i\lambda\pi})^n) = \frac{i}{2\pi\lambda} \left(\frac{-e^{z+i\lambda\pi}}{1+e^{z+i\lambda\pi}} + \frac{e^{z-i\lambda\pi}}{1+e^{z-i\lambda\pi}} \right) \\ &= \frac{i}{2\pi\lambda} \left(\frac{-e^z(e^{i\lambda\pi} - e^{-i\lambda\pi})}{e^{2z} + e^z(e^{i\lambda\pi} + e^{-i\lambda\pi}) + 1} \right) = \frac{-i}{2\pi\lambda} \frac{2i \sin \pi\lambda}{e^z + 2 \cos \pi\lambda + e^{-z}} \\ &= \frac{\sin \pi\lambda}{\pi\lambda} \frac{1}{e^z + 2 \cos \pi\lambda + e^{-z}} \end{aligned} \tag{A.5}$$

As already observed, $f_{GL(\lambda)}(z)$ is symmetric; therefore, the preceding equation applies to $z \geq 0$ as well. Using the substitution $Y = e^z + \cos \pi\lambda$, (A.5) can be integrated directly. The resulting c.d.f. of z is

$$F(z) = 1 - \frac{1}{\pi\lambda} \tan^{-1} \left(\frac{\sin \pi\lambda}{e^z + \cos \pi\lambda} \right) = \frac{1}{\pi\lambda} \tan^{-1} \left(\frac{\sin \pi\lambda}{e^{-z} + \cos \pi\lambda} \right). \quad \blacksquare$$

Proof of Theorem 2.3. The result is obvious for $K = 1$. Thus, as an induction hypothesis, assume that (A) holds for all $K \leq Q^*$, some $Q^* \geq 1$.

$$\sum_{k=1}^{Q^*+1} \xi_{k-1} \nu_k = \sum_{k=1}^{Q^*} \xi_{k-1} \nu_k + \xi_{Q^*} \nu_{Q^*+1}.$$

By induction, $\sum_{k=1}^{Q^*} \xi_{k-1} \nu_k = C(\xi_{Q^*})$. Hence, by Lemma 2.2, $\sum_{k=1}^{Q^*+1} \xi_{k-1} \nu_k = C(\xi_{Q^*+1})$, and (A) is proven by induction. From the definition of $C(\lambda_k, \xi_{k-1})$, (B) follows from (A). \blacksquare

Proof of Theorem 2.4. The c.d.f. of ϵ is $F(\epsilon) = e^{-e^{-\epsilon}}$ and hence the marginal distribution function of ϵ is $f(\epsilon) = e^{-\epsilon} \cdot e^{-e^{-\epsilon}}$ (similarly for ν).

$$t = \begin{cases} a + \epsilon - \log(e^a + e^b) & \text{if } b + \nu \leq a + \epsilon \ (\nu \leq a - b + \epsilon), \\ b + \nu - \log(e^a + e^b) & \text{if } b + \nu \geq a + \epsilon \ (\epsilon \leq -a + b + \nu). \end{cases}$$

Using the fact that probability $(b + \nu \geq a + \epsilon) + \text{probability}(b + \nu \leq a + \epsilon) = 1$,

$$\begin{aligned} f(t) &= F(a - b + \epsilon)f(\epsilon) \Big|_{t=a+\epsilon-\log(e^a+e^b)} + F(-a + b + \nu)f(\nu) \Big|_{t=b+\nu-\log(e^a+e^b)} \\ &= e^{-e^{-(t-b+\log(e^a+e^b))}} \cdot e^{-e^{-(t-a+\log(e^a+e^b))}} \cdot e^{-e^{-(t-a+\log(e^a+e^b))}} \\ &\quad + e^{-e^{-(t-a+\log(e^a+e^b))}} \cdot e^{-e^{-(t-b+\log(e^a+e^b))}} \cdot e^{-e^{-(t-b+\log(e^a+e^b))}} \\ &= \frac{e^a}{e^a + e^b} \cdot e^{-t} \cdot e^{-e^{-t}(e^a + e^b) + e^a/(e^a + e^b)}} \\ &\quad + \frac{e^b}{e^a + e^b} \cdot e^{-t} \cdot e^{-e^{-t}(e^a + e^b) + e^b/(e^a + e^b)}} \\ &= e^{-t} \cdot e^{-e^{-t}}. \end{aligned}$$

The probability $P(a + \epsilon \geq b + \nu) = \int_{-\infty}^{\infty} F(a - b + \epsilon)f(\epsilon) d\epsilon$

$$= \int_{-\infty}^{\infty} e^{-e^{-a+b-\epsilon}} e^{-\epsilon} e^{-e^{-\epsilon}} d\epsilon = \int_{-\infty}^{\infty} e^{-\epsilon} e^{-e^{-\epsilon}(1+(e^b/e^a))} d\epsilon = \frac{e^a}{e^a + e^b},$$

and

$$f(t | a + \epsilon \geq b + \nu) = \frac{F(a - b + \epsilon)f(\epsilon) \Big|_{\epsilon=t+\log(e^a+e^b)-a}}{P(a + \epsilon \geq b + \nu)}.$$

Using the same arguments as earlier, $f(t | b + \nu \leq a + \epsilon) = [(e^a/(e^a + e^b))e^{-t}e^{-e^{-t}}] / (e^a/e^a + e^b) = e^{-t}e^{-e^{-t}}$, and, by symmetry, $F(t | b + \nu \geq a + \epsilon) = e^{-t}e^{-e^{-t}}$. \blacksquare

Proof of Corollary 2.1. Corollary 2.1 is proved by induction. Note that parts 1-3 hold trivially for $j = 1$. Hence, as the induction hypothesis, assume that parts 1-3 hold for some j , $1 \leq j \leq J$.

To prove parts 1-3 for $j + 1$,

$$\begin{aligned} t_{j+1} &= \left(\max_{1 \leq l \leq j+1} (a_l + \epsilon_j) \right) - \log \left(\sum_{i=1}^{j+1} e^{a_i} \right) \\ &= \max \left(\max_{1 \leq l \leq j} (a_l + \epsilon_l), a_{j+1} + \epsilon_{j+1} \right) - \log \left(e^{\log \left(\sum_{i=1}^j e^{a_i} \right)} + e^{a_{j+1}} \right) \\ &= \max \left(\log \left(\sum_{i=1}^j e^{a_i} \right) + t_j, a_{j+1} + \epsilon_{j+1} \right) - \log \left(e^{\log \left(\sum_{i=1}^{j+1} e^{a_i} \right)} \right). \end{aligned}$$

Hence, by Theorem 2.1, t_{j+1} is extreme value, proving (1). Note that

$$a_{j+1} + \epsilon_{j+1} > a_l + \epsilon_l \text{ all } l; 1 \leq l < j + 1$$

$$\Leftrightarrow a_{j+1} + \epsilon_{j+1} > \max_{1 \leq l \leq j} (a_l + \epsilon_l) = \log \left(\sum_{i=1}^j e^{a_i} \right) + t_j.$$

Hence, (2) and (3) follow from Theorem 2.3. \blacksquare

Proof of Theorem 3.1. Note that this is obvious for $Q_j - 1 \leq l \leq K$. To prove by induction for all $l \geq 0$, assume the theorem holds for all j and some $l > 0$ and prove it therefore holds for $l - 1$.

$$\begin{aligned} U_{l-1,j} &= \max_{k \in S_{l-1,j}} (U_{l-1,j,k}) = \max_{k \in S_{l-1,j}} \left(U_{l,j,k} \frac{\xi_{l,k}}{\xi_{l-1,k}} + \nu_{l,k} \right) \\ &= \max_{k \in G_{l-1,j}} (\lambda_{l,k} \max_{m \in S_{l,k}} (U_{l,k,m}) + \nu_{l,k}) = \max_{k \in G_{l-1,j}} (\lambda_{l,k} U_{l,k} + \nu_{l,k}). \end{aligned}$$

By the induction hypothesis, $U_{l,k} - \log Z_{l,k}$ is extreme value. For $k, m \in G_{l-1,j}$, $k \neq m$, $U_{l,k}$ and $U_{l,m}$ are independently distributed. Hence, $(\lambda_{l,k}(U_{l,k} - \log Z_{l,k}) + \nu_{l,k})$ are i.i.d. extreme value for $k \in G_{l-1,j}$. Also, $\lambda_{l,k} U_{l,k} + \nu_{l,k} = \lambda_{l,k} \log Z_{l,k} + \lambda_{l,k}(U_{l,k} - \log Z_{l,k}) + \nu_{l,k}$. Thus,

$$U_{l-1,j} - \log Z_{l-1,j} = U_{l-1,j} - \log \left(\sum_{k \in G_{l-1,j}} Z_{l,k}^{\lambda_{l,k}} \right) \\ = U_{l-1,j} - \log \left(\sum_{k \in G_{l-1,j}} e^{\lambda_{l,k} \log Z_{l,k}} \right).$$

Applying Theorem 2.4, $U_{l-1,j} - \log Z_{l-1,j}$ is extreme value.

The probability of the j th alternative being chosen is $P_j = P_{0,j,j}$. In general, $P_{l,j,k}$ is the probability that U_k is maximal in the set $S_{l,j}$.

Note that for all $k \in S_{l,j}$,

$$U_{l,j,k} = U_{l,k,k} = \frac{U_k - \sum_{m=1}^l \xi_{m-1,j} \nu_{m,j}}{\xi_{l,j}}.$$

Because only U_k varies with k , and $\xi_{l,j} > 0$,

$$U_{l,j,j} = U_{l,j} [\equiv \max_{k \in S_{l,j}} (U_{l,j,k})] \quad \text{iff} \quad U_j = \max_{k \in S_{l,j}} (U_k).$$

From Theorem 3.3, the distribution of $U_{l,j} = \max_{k \in G_{l,j}} (U_{l+1,k})$ is independent of the k for which $U_{l,j} = U_{l,j,k}$. Hence, by Theorem 2.2 (recall $U_{0,j} = U_j$),

$$P_{l,j,j} = \prod_{k=l+1}^{Q_j} P_{k-1,j,k,j} = \frac{\prod_{k=l+1}^{Q_j} Z_{k,j}^{\lambda_{k,j}}}{\prod_{k=l}^{Q_j-1} Z_{k,j}}. \quad \blacksquare$$

Proof of Theorem 4.1. For ν_j extreme value and $\theta \sim C(\lambda)$, $0 < \lambda < 1$, the c.d.f. of ν_j is then $F(\nu_j) = e^{-e^{-\nu_j}}$. Thus,

$$F(\varepsilon_j) = e^{-e^{-\varepsilon_j}} = \int_{-\infty}^{\infty} f(\theta) e^{-e^{-(\varepsilon_j-\theta)/\lambda}} d\theta. \quad (\text{A.6})$$

This must hold for any ε_j . Let $A = e^{-\varepsilon_j/\lambda}$ and rewrite (A.6) as

$$\int_{-\infty}^{\infty} f(\theta) e^{-Ae^{\theta/\lambda}} d\theta = e^{-A^\lambda} \quad (\text{A.7})$$

for $A \geq 0$. Thus, $F(\varepsilon_1, \dots, \varepsilon_j) = \int_{-\infty}^{\infty} f(\theta) \prod_{j=1}^j e^{-e^{-(\varepsilon_j-\theta)/\lambda}} d\theta$

$$= \int_{-\infty}^{\infty} f(\theta) e^{-\left(\sum_{j=1}^j e^{-\varepsilon_j/\lambda}\right) e^{\theta/\lambda}} d\theta = e^{-\left(\sum_{j=1}^j e^{-\varepsilon_j/\lambda}\right)}.$$

Proof of Theorem 4.2. The c.d.f. of ν_j is $F(\nu_j) = 1/(1 + e^{-\nu_j})$. Therefore,

$$F(\varepsilon_j) = \frac{1}{1 + e^{-\varepsilon_j}} = \int_{-\infty}^{\infty} F(\varepsilon_j | \theta) f(\theta) d\theta = \int_{-\infty}^{\infty} f(\theta) \frac{1}{1 + e^{-(\varepsilon_j-\theta)/\lambda}} d\theta. \quad (\text{A.8})$$

Again, this must hold for all ε_j . Thus, let $A = e^{-\varepsilon_j/\lambda}$. Rewriting (A.8) yields

$$\int_{-\infty}^{\infty} f(\theta) \frac{1}{1 + Ae^{\theta/\lambda}} d\theta = \frac{1}{1 + A^\lambda} \quad (\text{A.9})$$

for all $A > 0$. Thus, for $\varepsilon_j \neq \varepsilon_l$, all $l \neq j$,

$$F(\varepsilon_1, \dots, \varepsilon_j) = \int_{-\infty}^{\infty} f(\theta) \prod_{j=1}^j \left(\frac{1}{1 + e^{-(\varepsilon_j-\theta)/\lambda}} \right) d\theta = \int_{-\infty}^{\infty} f(\theta) F(\varepsilon_1, \dots, \varepsilon_j | \theta) d\theta \\ = \int_{-\infty}^{\infty} f(\theta) \sum_{j=1}^j \left(\prod_{l \neq j} \left(\frac{e^{-\varepsilon_l/\lambda}}{e^{-\varepsilon_j/\lambda} - e^{-\varepsilon_l/\lambda}} \right) \frac{1}{1 + e^{-\varepsilon_j/\lambda} e^{\theta/\lambda}} \right) d\theta \quad (\text{A.10}) \\ = \sum_{j=1}^j \left(\prod_{l \neq j} \left(\frac{e^{\varepsilon_l/\lambda}}{e^{-\varepsilon_l/\lambda} - e^{\varepsilon_j/\lambda}} \right) \right) \int_{-\infty}^{\infty} F(\varepsilon_j | \theta) f(\theta) d\theta \\ = \sum_{j=1}^j \left(\prod_{l \neq j} \left(\frac{e^{-\varepsilon_j/\lambda}}{e^{-\varepsilon_j/\lambda} - e^{-\varepsilon_l/\lambda}} \right) \frac{1}{1 + e^{-\varepsilon_j}} \right). \quad (\text{A.11})$$

Equation (A.10) follows from the well-known partial fractions expansion

$$\prod_{j=1}^j \frac{1}{1 + A_j X} = \sum_{j=1}^j \left(\prod_{l \neq j} \left(\frac{A_j}{A_j - A_l} \right) \frac{1}{1 + A_j X} \right),$$

with $A_j = e^{-\varepsilon_j/\lambda}$ and $X = e^{\theta/\lambda}$, and (A.11) follows from (A.8).

L'Hopital's rule can be used to find $F(\varepsilon_1, \dots, \varepsilon_j)$ when $\varepsilon_j = \varepsilon_l$, some $j, l, j \neq l$. However, the following method is simpler and easily adapted to a computer algorithm. Obviously, the ε_j 's are exchangeable random variables; therefore, without loss of generality, assume that $\varepsilon_1, \dots, \varepsilon_L$ are the distinct values of the ε 's with $1 \leq L \leq J$ and let m_j be the multiplicity of ε_j . Thus, $\sum_{j=1}^L m_j = J$ and $F(\varepsilon_1, \dots, \varepsilon_j | \theta) = \prod_{j=1}^L (1/(1 + e^{-(\varepsilon_j-\theta)/\lambda}))^{m_j}$, which has a partial fractions expansion of

$$F(\varepsilon_1, \dots, \varepsilon_j | \theta) = \sum_{j=1}^L \sum_{l=1}^{m_j} B_{jl} \left(\frac{1}{1 + e^{-(\varepsilon_j-\theta)/\lambda}} \right)^l = \sum_{j=1}^L \sum_{l=1}^{m_j} B_{jl} F(\varepsilon_j, \dots, \varepsilon_j | \theta),$$

where the B_{jl} 's depend on $\lambda, \varepsilon_1, \dots, \varepsilon_j$, but not θ . Therefore,

$$F(\varepsilon_1, \dots, \varepsilon_j) = \int_{-\infty}^{\infty} f(\theta) F(\varepsilon_1, \dots, \varepsilon_j | \theta) d\theta \\ = \sum_{j=1}^L \sum_{l=1}^{m_j} B_{jl} \int_{-\infty}^{\infty} f(\theta) F(\varepsilon_j, \dots, \varepsilon_j | \theta) d\theta = \sum_{j=1}^L \sum_{l=1}^{m_j} B_{jl} F(\varepsilon_j, \dots, \varepsilon_j). \quad (\text{A.12})$$

Note that the B_{jl} 's are easily computed recursively as follows. Define B_{jl}^k by

$$F(\varepsilon_1, \dots, \varepsilon_k | \theta) = \sum_{j=1}^L \sum_{l=1}^{m_j} B_{jl}^k \left(\frac{1}{1 + e^{-(\varepsilon_j-\theta)/\lambda}} \right)^l.$$

(Thus, $B_{ji} \equiv B_{ji}^j$.) Then, from (A.11), $B_{ji}^k = \prod_{l=ij} (e^{-\epsilon_j/\lambda}) / (e^{-\epsilon_j/\lambda} - e^{-\epsilon_l/\lambda})$ and $B_{ji}^k = 0$, $l > i$. Then, for $k = L, \dots, J-1$ and $1 \leq j \leq L$, if $\epsilon_j \neq \epsilon_{k+1}$,

$$B_{ji}^{k+1} = \frac{e^{-\epsilon_j/\lambda}}{e^{-\epsilon_j/\lambda} - e^{-\epsilon_{k+1}}} \sum_{n=1}^{m_j} B_{jn}^k \left(\frac{e^{-\epsilon_{k+1}/\lambda}}{e^{-\epsilon_{k+1}/\lambda} - e^{-\epsilon_j/\lambda}} \right)^{n-1};$$

if $\epsilon_j = \epsilon_{k+1}$ and $l > i$, then $B_{ji}^{k+1} = B_{ji}^{k+1}$; and, if $\epsilon_j = \epsilon_{k+1}$ and $l = i$, then

$$B_{ji}^{k+1} = \sum_{l=ij} \sum_{n=1}^{m_j} B_{jn}^k \left(\frac{e^{-\epsilon_j/\lambda}}{e^{-\epsilon_j/\lambda} - e^{-\epsilon_l/\lambda}} \right)^n.$$

The final step is to show that

$$F(\underbrace{\epsilon_1, \dots, \epsilon_j}_{i\epsilon_j}) = \sum_{n=1}^{m_j} R_{in}(\lambda) \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n.$$

Obviously, $R_{i1}(\lambda) \equiv 1$, and $R_{in} = 0$, for $n \neq i$. Thus, R_{in} is developed recursively. To derive R_{in} from R_{i-1n} , take

$$F(\underbrace{\epsilon_1, \dots, \epsilon_j}_{i-1\epsilon_j}) = \sum_{n=1}^{i-1} R_{i-1n} \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n = \int_{-\infty}^{\infty} f(\theta) \left(\frac{1}{1 + e^{-(\epsilon_j - \theta)/\lambda}} \right)^{i-1} d\theta \quad (\text{A.13})$$

and differentiate with respect to ϵ_j to obtain

$$\begin{aligned} \sum_{n=1}^{i-1} n R_{i-1n} \left(\left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n - \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^{n+1} \right) \\ = \frac{i-1}{\lambda} \int_{-\infty}^{\infty} f(\theta) \left(\left(\frac{1}{1 + e^{-(\epsilon_j - \theta)/\lambda}} \right)^{i-1} - \left(\frac{1}{1 + e^{-(\epsilon_j - \theta)/\lambda}} \right)^i \right) d\theta. \end{aligned} \quad (\text{A.14})$$

Rearranging and substituting (A.13) into (A.14) yields

$$F(\underbrace{\epsilon_1, \dots, \epsilon_j}_{i\epsilon_j}) \equiv \int_{-\infty}^{\infty} \left(\frac{1}{1 + e^{-(\epsilon_j - \theta)/\lambda}} \right)^i f(\theta) d\theta = \sum_{n=1}^{i-1} R_{in} \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n, \quad (\text{A.15})$$

where $R_{in} = R_{i-1n} + [\lambda/(i-1)]((n-1)R_{i-1n-1} - nR_{i-1n})$. Thus, $R_{2i} = 1 - \lambda$, $R_{3i} = \lambda$, $R_{31} = 1 - \frac{3}{2}\lambda + \frac{1}{2}\lambda^2$, and so forth. Note that R_{in} depends only on λ , not ϵ_j . Combining (A.15) and (A.12),

$$F(\epsilon_1, \dots, \epsilon_j) = \sum_{j=1}^L \sum_{i=1}^{m_j} B_{ji} \sum_{n=1}^i R_{in} \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n = \sum_{j=1}^L \sum_{n=1}^{m_j} \left(\sum_{i=n}^{m_j} B_{ji} R_{in} \right) \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n.$$

Thus,

$$F(\epsilon_1, \dots, \epsilon_j) = \sum_{j=1}^L \sum_{n=1}^{m_j} C_{jn} \left(\frac{1}{1 + e^{-\epsilon_j}} \right)^n \quad \text{for } C_{jn} = \sum_{i=n}^{m_j} B_{ji} R_{in}. \quad \blacksquare$$

Proof of Lemma 4.1. Let $\epsilon_{i,j}$ be the vector of all ϵ_i 's, such as $i \in S_{i,j}$. Integrate out $v_{Q_{i-1,j}}$ following the method of Theorem 4.1. Thus,

$$\begin{aligned} F(\epsilon_{Q_{j-1,j}} | v_{Q_{j-1,j}}, v_{1,j}, \dots, v_{Q_{j-2,j}}) \\ = F_{\lambda} \left(\left(\epsilon_{Q_{j-1,j}} - \left(\sum_{i=1}^{Q_{j-2}} \xi_{i-1,j} v_{1,j} \right) \right) / \xi_{Q_{j-2,j}} \right) \\ = e^{-\sum_{k \in Q_{Q_{j-1,j}}} \left(\left(\epsilon_k - \sum_{l=1}^{Q_{j-3}} \xi_{l-1,j} v_{1,j} \right) / \xi_{Q_{j-3,j}} - v_{Q_{j-2,j}} \right) / \lambda_{Q_{j-1,j}} \lambda_{Q_{j-2,j}} } \\ = e^{-\sum_{k \in Q_{Q_{j-1,j}}} \left(\left(\epsilon_k - \sum_{l=1}^{Q_{j-3}} \xi_{l-1,j} v_{1,j} \right) / \xi_{Q_{j-3,j}} - v_{Q_{j-2,j}} \right) / \lambda_{Q_{j-1,j}} \lambda_{Q_{j-2,j}} } \end{aligned}$$

Next, integrate out $v_{Q_{j-2,j}}$. Successively integrating out each variance component, starting at the deepest level, yields

$$F(\epsilon_1, \dots, \epsilon_j) = e^{-\left(\sum_{k \in Q_{Q_{1,1}}} \left(\sum_{l \in Q_{1,1,k}} \left(\dots \left(\sum_{m \in Q_{1,1,l}} e^{-\epsilon_j / \lambda_{Q_{j-1,j}}} \right)^{\lambda_{2,j}} \right)^{\lambda_{1,k}} \right)^{\lambda_{1,j}} \right) / \lambda_{Q_{j-1,j}}}$$

REDACTED -- FOR PUBLIC INSPECTION

**HIGHLY CONFIDENTIAL ATTACHMENT --
SUBJECT TO SECOND PROTECTIVE
ORDER IN CS DOCKET NO. 01-348**