On Cardinalization of Consumer Utility in Discrete Choice Models

Yuichiro Kanazawa*, Ph.D. in statistics
Professor by Special Appointment, International Christian University,
Visiting Professor, Nara Institute of Science and Technology,

1 Random utility models: A brief history

We first review how econometricians have modeled the behavioral process that leads to the person's, the firm's, or more generally the decision maker's choice since Luce (1959) [4], Marschak (1960) [6], Lancaster (1971) [3], and McFadden (1974) [8].[1)]

The set of alternatives, named the *choice set*, needs to have three characteristics:
1. The alternatives must be *mutually exclusive* from the agent's perspective;
2. The choice set must be *exhaustive*, in that all possible alternatives are included;
3. The number of alternatives must be *finite*.

For the sake of brevity, we henceforth use the term ‘agent’ for these decision makers, be it a person, a firm, or a student, or a respondent to a survey. The agents themselves exactly know all the factors that collectively determine their choices.[2)] Some of these factors are observed by the researcher, but the others are not. This dichotomy between the agents on the theater and the researcher as audience is the most striking feature of this model.

We label the agent's demografic and alternative speficic characteristics observed by all the agents as well as the researcher respectively as $\boldsymbol{d}_i$ and $\boldsymbol{x}_j$ vector, and those observed by the agents but unobserved by the researcher as ϵ_{ij}, where i indexes the agent $i = 1, \ldots, I$ and j indexes the alternative $j = 1, \ldots, J$. These factors jointly lead the agent i to choose

*This work was supported by the Japan Society for the Promotion of Science under the Grant-in-Aid for Scientific Research (C) 20K01595 and by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University.

1)Originally, the logit formula to be derived was derived by Luce (1959) [4] from assumptions about the characteristics of choice probabilities, namely the *independence from irrelevant alternatives* (IIA) property. Marschak (1960) [6] showed that these axioms implied that the model is consistent with utility maximization. The relation of the logit formula to the distribution of unobserved utility (as opposed to the characteristics of choice probabilities) was developed by Marley, as cited by Luce and Suppes (1965) [5], who showed that the Gumbel (extreme value) distribution leads to the logit formula. Lancaster (1971) [3] and McFadden (1974) [8] completed the analysis by showing the converse: that the logit formula for the choice probabilities necessarily implies that unobserved utility is distributed Gumbel (extreme value).

2)If otherwise, it is claimed that they should not be able to make such choices.

one and only one alternative j. Therefore his/her behavioral process is expressed by a function

$$h\left(\boldsymbol{d}_i, \boldsymbol{x}_j, \epsilon_{ij}\right) = \text{choosing alternative } j.$$

In this sense, this choice situation needs to be understood from a causal perspective.

There are factors that collectively determine, or cause, the agent's choice. In other words, it is deterministic in the sense that given his/her demographics $\boldsymbol{d}_i$, $\boldsymbol{x}_j$ and ϵ_{ij}, the choice of agent i is fully determined.[3)] perhaps out of convenience.[4)] The unobserved terms are assumed to have density $f(\epsilon_{ij})$. The researcher therefore must express the possible choices $j = 1, \ldots, J$ of agent $i = 1, \ldots, I$ as events with the corresponding probabilities.

Concretely, the probability that the agent i chooses a particular alternative j from the set of all possible outcomes indexed by j is the probability that the unobserved factors are such that the behavioral process results in that outcome:

$$\Pr\left\{\epsilon_i \text{ s.t. } h\left(\boldsymbol{d}_i, \boldsymbol{X}_j, \epsilon_i\right) = \text{choosing alternative } j\right\}.$$

It is researcher's responsibility to choose an appropriate distribution for ϵ_i.

I now introduce the standard derivation of choice probabilities following Train (2009) [9]. The agent i would assign a certain level of utility U_{ij} to each alternative j. The agent then chooses the alternative that provides the **highest** utility. The behavioral model is therefore: choose alternative k if and only if

$$U_{ik} > U_{ij} \text{ for } \forall j \neq k.$$

The reseacher does not observe the agent i's utility, however. Instead, the researcher only observes: **demographics of the agents** *observable by the researcher*, labeled $\boldsymbol{d}_i$; and **attributes of the alternatives** presumed to be in the mind of the agent when he/she makes decision among alternatives *observable by the reseacher*, labeled $\boldsymbol{x}_j$.

We assume that the researcher can specify a function that relates these observed factors to the agent i's utility towards alternative j. This function V_{ij} of $\boldsymbol{d}_i$ and $\boldsymbol{x}_j$ expressed as

$$V_{ij} = V\left(\boldsymbol{d}_i, \boldsymbol{x}_j\right) \tag{1}$$

is called *representative utility*. Usually, V_{ij} depends on parameters $\boldsymbol{\alpha}$ on $\boldsymbol{d}_i$ and $\boldsymbol{\beta}$ on $\boldsymbol{x}_j$ that are unknown to the researcher, for instance, in the linear form below:

$$V_{ij} = V\left(\boldsymbol{x}_j, \boldsymbol{d}_i\right) = \boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j. \tag{2}$$

These parameters must be estimated statistically.

There are some part of utility that the researcher does not or cannot observe, and this fact makes

$$U_{ij} \neq V_{ij}.$$

[3)] Here we assume that the choice agent i makes is independent of choices other agents make.

However the researcher does not observe ϵ_{ij}, and thus cannot predict the agent's behavior precisely. As a result, the researcher is forced to assume the unobserved terms ϵ_{ij} to be *random*

[4)] We will give one most standard interpretation as to what is meant by the distribution of ϵ_{ij} below.

Under such circumstances, the researcher assumes that the utility U_{ij} can be decomposed as

$$U_{ij} = V_{ij} + \epsilon_{ij}, \tag{3}$$

where ϵ_{ij}[5)] captures all the factors that jointly affect utility as perceived by agent i for alternative j, but are not included in V_{ij}.

1.1 Logit choice probabiity

This additive construction in (3) is fairly general because ϵ_{ij} is defined as simply the difference between true utility U_{ij} and the part of utility that the researcher captures in Vij. However, it also means that the random ϵ_{ij} and its characteristics such as its distribution depend critically on the researcher's specification of V_{ij} rather than a particular choice situation. In other words, it is determined by how a researcher expresses that choice situation.

The joint density of the random vector $\boldsymbol{\epsilon}_{i.} = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$ is denoted by $f_{\boldsymbol{\epsilon}_{i.}}(\cdot)$. With this density, the researcher can make probabilistic statements about the agent's choice. Specifically, the probability agent i chooses alternative j is

$$\begin{aligned} P_{ij} &= \Pr\{U_{ij} > U_{ik} \quad \forall k \neq j\} \\ &= \Pr\{V_{ij} + \epsilon_{ij} > V_{ik} + \epsilon_{ik} \quad \forall k \neq j\} \\ &= \Pr\{\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik} \quad \forall k \neq j\}, \end{aligned} \tag{4}$$

where we use the notation {statement} as the indicator function taking 1 if the statement is true and 0 otherwise as Bruno de Finetti introduced.

This probability P_{ij} is cumulative in the sense that the probability that each random term $\epsilon_{ik} - \epsilon_{ij}$, $k \neq j$, is below the observed quantity $V_{ij} - V_{ik}$, $k \neq j$. With the density $f_{\boldsymbol{\epsilon}_{i.}}(\cdot)$, this cumulative probability in (4) can be rewritten with multiple integral as

$$\begin{aligned} & P_{ij} = \Pr\{\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik} \quad \forall k \neq j\} \\ = & \int \cdots \int_{\boldsymbol{\epsilon}_{i.}} \{\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik} \quad \forall k \neq j\} \\ & \quad \times f_{\boldsymbol{\epsilon}_{i.}}(\epsilon_{i1}, \ldots \epsilon_{iJ}) d\epsilon_1 \cdots \boldsymbol{d\epsilon_{j-1} d\epsilon_{j+1}} \cdots d\epsilon_J \boldsymbol{d\epsilon_j} \\ = & \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\epsilon_{ij}+V_{ij}-V_{i1}} \cdots \int_{-\infty}^{\epsilon_{ij}+\boldsymbol{V_{ij}-V_{ij-1}}} \int_{-\infty}^{\epsilon_{ij}+\boldsymbol{V_{ij}-V_{ij+1}}} \cdots \int_{\infty}^{\epsilon_{ij}+V_{ij}-V_{iJ}} \right. \\ & \quad \left. \times f_{\boldsymbol{\epsilon}_{i.}}(\epsilon_{i1}, \ldots \epsilon_{iJ}) d\epsilon_1 \cdots \boldsymbol{d\epsilon_{j-1} d\epsilon_{j+1}} \cdots d\epsilon_J \right] \boldsymbol{d\epsilon_j}. \end{aligned} \tag{5}$$

Completing integration surrounded by the brackets [and] in (5) obtains a conditional choice probabiity $F_{\epsilon_j}(\epsilon_{ij} + V_{ij} - V_{i1}, \cdots, \epsilon_{ij} + V_{ij} - V_{ij-1}, \epsilon_{ij} + V_{ij} - V_{ij+1}, \cdots, \epsilon_{ij} + V_{ij} - V_{iJ})$ given random variables ϵ_j. If ϵ_{ij} is given, this expression is also the conditional cumulative

[5)]This quantity ϵ_{ij} is sometimes referred to as idiosyncratic preference or idiosyncratic utility in marketing and economics literature.

distribution for each ϵ_{ij} evaluated at $\epsilon_{ij} + V_{ij} - V_{nk}$. Then we integrate this distribution function of ϵ_j from negative infinity to positive infinity to obtain the choice probability.

To derive the logit model, the *researcher* uses the general notation in (5), and assume that each ϵ_{ij} is independently, identically distributed Gumbel (extreme value), or whose cumulative distribution function is

$$F(\epsilon_{i1}, \ldots, \epsilon_{iJ}) = \prod_{k=1}^{J} \exp\left\{-\exp(-\epsilon_{ik})\right\}. \tag{6}$$

with the corresponding density for each ϵ_{ij} is

$$f_{\epsilon_{ij}}(\epsilon_{ij}) = \frac{dF(\epsilon_{ij})}{d\epsilon_{ij}} = \exp(-\epsilon_{ij}) \exp\left\{-\exp(-\epsilon_{ij})\right\}. \tag{7}$$

Substituting (2), (6) and (7) into (5) obtains

$$\begin{aligned} P_{ij}(\boldsymbol{V}) &= \int_{-\infty}^{\infty} F_{\epsilon_j}(\epsilon_{ij} + V_{ij} - V_{i1}, \cdots, \epsilon_{ij} + V_{ij} - V_{ij-1}, \\ &\qquad \epsilon_{ij} + V_{ij} - V_{ij+1}, \cdots, \epsilon_{ij} + V_{ij} - V_{iJ}) f_{\epsilon_j}(\epsilon_j) d\epsilon_{ij} \\ &= \int_{-\infty}^{\infty} \prod_{k=1, k\neq j}^{J} \exp\left[-\exp\left(-\epsilon_{ij} + V_{ij} - V_{ik}\right)\right] \\ &\qquad \times \exp(-\epsilon_{ij}) \exp\left\{-\exp(-\epsilon_{ij})\right\} d\epsilon_{ij} \\ &= \frac{\exp(V_{ij})}{\sum_{k=1}^{J} \exp(V_{ik})} = \frac{\exp\left(\boldsymbol{\alpha} \boldsymbol{d}_i + \boldsymbol{\beta} \boldsymbol{x}_j\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{\alpha} \boldsymbol{d}_i + \boldsymbol{\beta} \boldsymbol{x}_j\right)}. \end{aligned} \tag{8}$$

the so-called "logit choice probability."Notice that the index i disappears in (8) because it is integrated out. [6)]

We note that, in expression (8), this probability no longer depends on the agent's taste term ϵ_{ij}. With (8), the market share of the alternative j aggregated over the agents or over $i = 1, \ldots, I$ is expressed as

$$P_j(\boldsymbol{V}) \;=\; \int_{\boldsymbol{d}_i} \frac{\exp\left(\boldsymbol{\alpha} \boldsymbol{d}_i + \boldsymbol{\beta} \boldsymbol{x}_j\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{\alpha} \boldsymbol{d}_i + \boldsymbol{\beta} \boldsymbol{x}_j\right)} P(\boldsymbol{d}_i) \tag{9}$$

where we assume that $P(\boldsymbol{d}_i)$ is the underlying joint density of demographics $\boldsymbol{d}_i$.

1.2 Interpreting idiosyncratic taste ϵ_{ij} under logit choice probability

How to interpret the distribution of ϵ_{ij}? The most widely-held interpretation of this distribution is as follows: Consider a population of agents who face the same observed repre-

6) If choice data for alternative $j = 1, \ldots, J$ are available for agents $i = 1, \ldots, I$ with each making choices independently, then estimation methods such as maximum likelihood can be employed to estimate the parameters associated with product characteristics $\boldsymbol{x}_j$ and with $\boldsymbol{d}_i$ in principle.

sentative utility V_{ij} for each alternative $j = 1, \ldots, J$ as agent i. Among these agents[7)], the researcher assumes the values of the unobserved factors ϵ_{ij} differ. That is, the researcher assumes heterogeneity of agents is manifested only[8)] in the heterogeneity of ϵ_{ij}.[9)] The unconditional density $f_{\boldsymbol{\epsilon}_{ij}}(\cdot)$ is the distribution of the *unobserved portion of utility within the population of agents who face the same observed representative portion of utility as agent i* for $i = 1, \ldots, I$. Under this interpretation, the probability P_{ij} is the share of people who choose alternative j within the population of people who face the same observed representative utility for each alternative as agent i.

Different discrete choice models are obtained from different specifications of this density. For instance, *logit* model and *nested logit* model are known to have *closed-form* choice probabilities. These models are derived under the assumption that the unobserved portion $\boldsymbol{\epsilon}_{i.} = (\epsilon_{i1}, \ldots, \epsilon_{iJ})$ of utility is distributed i.i.d *Gumbel* [10)] (*extreme value*) and a type of *generalized Gumbel* (*generalized extreme value*), respectively.

On the other hand, *probit* model is derived under the assumption that $f_{\boldsymbol{\epsilon}_{i.}}(\cdot)$ is a multivariate normal, and *mixed logit model* is based on the assumption that the unobserved portion of utility consists of a part that follows any distribution specified by the researcher plus a part that is i.i.d. Gumbel (extreme value). With probit and mixed logit, the resulting integral does *not have a closed form* and is thus evaluated numerically through simulation.

1.3 "Only differences in utility matter" and "the scale of utility is arbitrary"

Train (2009) [9] wrote and I quote that "[s]everal aspects of the behavioral decision process affect the specification and estimation of any discrete choice model. The issues can be summarized easily in two statements: "**[o]nly differences in utility matter**"and "**[t]he scale of utility is arbitrary**."In essence, what Train (2009) [9] stated is the utility as defined as in (3) is **ordinal** and **not cardinal**: (1) The absolute level of utility is irrelevant to both the decision maker's behavior and the researcher's model. If a constant is added to the utility of all alternatives, the alternative with the highest utility doesn't change; The decision maker chooses the same alternative with U_{ij} for $\forall j$ as with $U_{ij} + k$ for $\forall j$ for any constant k.[11)] (2) Just as adding a constant to the utility of all alternatives does not

---

7) It is likely that such agents share the same demographics $\boldsymbol{d}_i$ with i, but not necessarily. This is because the difference in $\boldsymbol{d}_i$ can be compensated with the (perceived) difference in $\boldsymbol{x}_j$, resulting in the same V_{ij}

8) It appears as if heterogeneity also exists because demographics $\boldsymbol{d}_i$ of agents. However, please note that demographic heterogeneity presides in the data to be analyzed and is not assumed by nor under control of the researcher.

9) This assumption is considerably extended later in Berry (1994) [1] and Berry, Levinsohn, Pakes (1995, henceforth BLP(1995)) [2] when they introduced the random coefficient in the utility. We will discuss this extension later.

10) One could argue that the choice of this distributional assumption is purely out of convenience because of its well known property: If $X \sim \text{Gumbel}(\alpha_X, \beta)$ and $Y \sim \text{Gumbel}(\alpha_Y, \beta)$ are independent, then their difference $X - Y$ is distributed as $X - Y \sim \text{Logistic}(\alpha_X - \alpha_Y, \beta)\}$.

11) A colloquial way to express this fact is, "A rising tide raises all boats."

change the decision maker's choice, neither does multiplying each alternative's utility by a constant. The alternative with the highest utility is the same no matter how utility is scaled. The model $U_{ij}^0 = V_{ij} + \epsilon_{ij}$ for $\forall j$ is equivalent to $U_{ij}^1 = \lambda V_{ij} + \lambda \epsilon_{ij}$ for $\forall j$ for any $\lambda > 0$. To take account of this fact, the researcher must normalize the scale of utility.

1.4 Logit choice probability with unobserved demand characteristic ξ_j

Berry (1994) [1] introduced to the discrete choice model of product differentiation *unobserved—by the researcher who is not a agent by definition—demand characteristics* ξ_j, while *assuming all other characteristics and all decisions are observable and observed by all agents.* Berry (1994) [1] claimed that the proposed estimation methods do *not require the statistician and econometrician to observe all relevant product characteristics.*

Now the expression of the utility in (2) can be rewritten, for instance, linearly as

$$V_{ij} = V\left(\boldsymbol{x}_j, \boldsymbol{d}_i\right) = \boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j \tag{10}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are parameters to be estimated.

From (3) and (10), the utility of agent i for alternative j is written as

$$U_{ij} = V_{ij} + \epsilon_{ij} = \boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j + \epsilon_{ij}, \tag{11}$$

with two terms ξ_j and ϵ_{ij} are random.

If we can further assume that **the conditional distribution of ϵ_{ij} given ξ_j is Gumbel**, the logit-type choice probability is derived as

$$P_{ij}(\boldsymbol{V}) = \frac{\exp(\boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j)}{\sum_{k=1}^{J} \exp(\boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_k + \xi_k)}. \tag{12}$$

1.5 Interpreting idiosyncratic taste ϵ_{ij} in the presence of unobserved demand characteristic ξ_j

Berry (1994) [1] claimed that "**[t]he term $\boldsymbol{\xi_j}$ might be thought of as the mean of consumers' valuations of an unobserved product characteristic such as product quality, while the $\boldsymbol{\epsilon_{ij}}$ represents the distribution of consumer preferences about this mean**."

We now try to extend the previous interpretation of ϵ_{ij} forwarded by Train (2009) [9] in subsection 1.2 in the presence of ξ_j. Consider a population of agents who face the same representative (observed) utility plus unobserved demand characteristic $V_{ij} = \boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j$ for alternative j as agent i. Note that ξ_j represents the average or common utility that agents obtain from the unobserved attributes of alternative j. Among these agents, we assume the values of the unobserved idiosyncratic factors ϵ_{ij} distribute about its mean ξ_j. Berry (1994) [1] assumes that the value of unobserved demand characteristic ξ_j also varies with $j = 1, \ldots, J$, and as a result, econometricians and statisticians treat ξ_j as *stochastic.*

Let $f_{\boldsymbol{\epsilon}_{i\cdot}, \boldsymbol{\xi}_{\cdot}}(\cdot)$ of $\boldsymbol{\epsilon}_{i\cdot}$ and $\boldsymbol{\xi}_{\cdot}$ be the joint distribution of the unobserved portion of utility within the population of agents who face the same observed portion $V_{ij} = \boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j$ of utility.As a result, while the integral is taken with Gambel $f_{\epsilon_{ij}}(\epsilon_{ij})$ in deriving (8), it is now replaced by its conditional Gumbel counterpart $f_{\epsilon_{ij}|\xi_j}(\epsilon_{ij}|\xi_j)$ whose conditional mean is $E_{\epsilon_{ij}|\xi_j}(\epsilon_{ij}|\xi_j) = \xi_j$ in deriving (12).

1.6 Random coefficient logit choice probability

Berry (1994) and BLP (1995) further extended (12) so that the model now includes the random coefficients model to allow for interaction between agent and alternatives. A familiar starting point is to allow each agent to have a different preference for each different observable alternative characteristic:

$$U_{ij} \;\; = \;\; \boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j + \sum_{k=1}^{K} \sigma_k x_{jk} \nu_{ik} + \epsilon_{ij}, \tag{13}$$

where $(\boldsymbol{\zeta}_i, \boldsymbol{\epsilon}_i) = (\nu_{i1}, \ldots, \nu_{iK}, \epsilon_{i1}, \ldots, \epsilon_{iJ})$ is a mean zero vector of random variables about ξ_j with (a known) distribution function. Now the contribution of x_k units of the k-th alternative characteristic to the utility of individual i is $(\beta_k + \sigma_k \nu_{ik})x_k$, which varies over agents. Under this formulation, it is common to scale ν_{ik} such that $E(\nu_{ik}) = 0$ and $E(\nu_{ik}^2) = 1$, so that the mean and variance of the marginal utilities associated with characteristic k are β_k and σ_k^2 respectively. BLP (1995) claim that "[t]**his specification is particularly tractable if ϵ_i consists of i.i.d. extreme value deviates**."

For this model, the market share function P_{ij} is obtained in two stages: First, conditional on the $(\xi_j, \boldsymbol{d}_i, \boldsymbol{\nu}_i)$, probability P_{ij} that agent i chooses alternative j is given by the logit-type formula:

$$P_{ij} \;\; = \;\; \frac{\exp(\boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j + \sum \sigma_k x_{jk}\nu_{ik})}{\sum_{j=1}^{J} \exp(\boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j + \sum \sigma_k x_{jk}\nu_{ik})}. \tag{14}$$

For this operation to be valid, or ϵ_{ij} can be integrated out conditionally on ξ_j and ν_{ij} as in (14), this conditional distribution must be Gumbel.[12)]

Second, we assume the demographic variable $\boldsymbol{d}_i$ has a joint distribution dependent on the demographic, and it is safely assume its distribution is independent of ν_{ik}.[13)] Integrating out P_{ij} in expression in (14) over the joint distribution of $(\boldsymbol{d}_i, \boldsymbol{\nu}^{ik})$ gives the market share P_j:

$$P_j \;\; = \;\; \int_{\boldsymbol{d}_i, \boldsymbol{\nu}_i} \frac{\exp(\boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j + \sum \sigma_k x_{jk}\nu_{ik})}{\sum_{j=1}^{J} \exp(\boldsymbol{\alpha}\boldsymbol{d}_i + \boldsymbol{\beta}\boldsymbol{x}_j + \xi_j + \sum \sigma_k x_{jk}\nu_{ik})} P(\boldsymbol{d}_i, d\boldsymbol{\nu}_i), \tag{15}$$

where $P(\boldsymbol{d}_i, \boldsymbol{\nu}_i)$ is the underlying joint density of $(\boldsymbol{d}_i, \nu_{i1}, \ldots, \nu_{iK_1})$.

[12)] If these two random taste terms ϵ_{ij} and ν_{ik}, both varying with respect to i, $i = 1, \ldots, I$ and j, $j = 1, \ldots, J$, are independent, the logistic choice probabilities in (14) can be justified. However, this assumption is almost impossible to verify.

[13)] Demographic distribution can be obtained from the database and thus is independent of ν_{ik}.

1.7 Interpreting unobserved idiosyncratic taste ϵ_{ij} in the presence of unobserved demand characteristic ξ_j and the random coefficient $(\beta_k + \sigma_k \nu_{ik})x_k$ under logit choice probability

Let us try to come up with interpretation of the meaning of ϵ_i in the presence of ξ_j and the random coefficient $(\beta_k + \sigma_k \nu_{ik})x_k$ as previously done in subsection 1.5 after Train (2009) [9].

Consider a population of agents who face the same observed utility $V_{ij} = \boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j$ for each alternative as agent i. Among these agents, we assume the values of the unobserved taste term ϵ_i whose mean is ξ_j[14] differ, but its random behavior is unaffected by (independent of) another taste term ν_{ik} that evaluates interaction between agent and his/her alternatives.

1.8 Problem with unobserved idiosyncratic taste term ϵ_{ij}

We have seen that the assumptions of Gumbel unobserved idiosyncratic terms ϵ_{ij} unconditionally, conditionally given ξ_j, or conditionally given ξ_j and ν_{ik} enable us to derive the logit-type choice probabiities such as the (8), (12), and (14) possible.

$$P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j) = \frac{\exp\left(\boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j\right)}, \tag{8}$$

$$P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j,) = \frac{\exp(\boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j + \xi_j)}{\sum_{k=1}^{J} \exp(\boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_k + \xi_k)}, \tag{12}$$

$$P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij}) = \frac{\exp(\boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j + \xi_j + \sum \sigma_k x_{jk} \nu_{ik})}{\sum_{j=1}^{J} \exp(\boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j + \xi_j + \sum \sigma_k x_{jk} \nu_{ik})}. \tag{14}$$

Since the terms ϵ_{ij} and ν_{ik} with the same value of i index the same agent, they are unlikely to be independent in general, especially when $j = k$. In such cases, ϵ_{ij} and ν_{ik} may even be indistinguishable. Therefore we need to critically examine the validity of assumptions that conditional distribution of ϵ_{ij} given ξ_j is Gumbel when we derive (12), or that conditional distribution of ϵ_{ij} given ξ_j and ν_{ik} is Gumbel when we derive (14), even when we can safely assume unconditional distribution of ϵ_{ij} is Gumbel when we derive (8). This is because deriving logit-type choice probabilities listed above in (8), (12), and (14) all critically dependent on the assumption that **difference between two indepedent Gumbel random variables ϵ_{ik} and** ϵ_{ij} as in (5) has logistic distribution.

We need to ask ourselves if we can remove the idiosyncratic taste random term ϵ_{ij} altogether and still able to model the choice situation?

---

[14]The term ξ_j represents the average utility that consumers obtain from the unobserved attributes of product j.

2 A statistical formulation of logit choice probability without the unobserved idiosyncratic taste term ϵ_{ij}

Suppose we have a baseline alternative,[15] and we represent such an alternative indexed by $j = 0$ as $\boldsymbol{x}_j = 0$, $\boldsymbol{d}_i = 0$, $\xi_j = 0$ with the choice probability of the baseline alternative as P_{i0}. As in page 1, we retain the following assumptions:

Assumption 1. The agent himself/herself knows exactly all the factors that collectively determine his or her choice. [16]

Assumption 2. Some of these factors are observed by the researcher, but the others are not.

This means that the reseacher cannot predict the agent's behavior precisely, but the agents know exactly what his/her demographics and the alternative characterisitcs made him/her choose a particular alternative.

As in page 1, we assume the *choice set* needs to have three characteristics:

Assumption 3. The alternatives must be *mutually exclusive* from the agent's perspective;

Assumption 4. The choice set must be *exhaustive*, in that all possible alternatives are included;

Assumption 5. The number of alternatives must be *finite*.

In addition, we assume the following for alternative $j = 0, \ldots, J$:

Assumption 6. There is no innate underlying ordering of these alternatives, and numerical labels of from 0 to J are attached to alternatives for convenience in describing the distribution.

Under **Assumptions 1** to **6**, agent i trying to choose an alternative j from the choice set indexed by $j = 1, \ldots, J$ is *ex ante* characterized by a version of **categorical distribution** with one trial, where a sample space is 1 in the chosen alternative with the highest probability and 0 otherwise. The distribution describes the possible results of a random variable that can take on one of $J + 1$ possible categories, with the probability of each category indexed by $P_{i0}, \ldots, P_{iJ}$.[17]

Concretely, for agent i, $i = 1, \ldots, I$ with his/her demographic as $\boldsymbol{d}_i$, the characteristics of alternative j as $\boldsymbol{x}_j$, $j = 0, \ldots, J$, unobserved demand characteristic of alternative j as ξ_j, interaction between agent i and alternative j as ν_{ij}, and with his/her choice probability

[15] For instance, in marketing and empirical industrial organization, the alternative of not purchasing anything or purchasing outside good is often used as the most natural baseline alternative.

[16] If otherwise, she/he should not be able to make such choices.

[17] I extended the category from $1, \ldots, J$ to $0, \ldots, J$ to include a baseline-category. For instance, in purchasing decisions, it is natural to include an alternative of not purchasing anything, or purchasing *outside good*.

P_{ij} of alternative j written as a function of those variables as $P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij})$, his or her *ex ante categorical distribution* with one opportunity to choose one alternative from the choice set **as perceived by the researcher** is

$$\begin{aligned} &\Pr\{N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}; P_{i0}, \ldots, P_{iJ}\} \\ &\quad = \frac{1!}{n_{i0}! \cdots n_{iJ}!} P_{i0}^{n_{i0}}(\boldsymbol{d}_i, \boldsymbol{x}_0, \xi_0, \nu_{i0}) \times \cdots \times P_{iJ}^{n_{iJ}}(\boldsymbol{d}_i, \boldsymbol{x}_J, \xi_J, \nu_{iJ}), \end{aligned} \tag{16}$$

where j-th element of $(n_{i0}, \ldots, n_{iJ})$ is 1 and 0 otherwise.

After the agent chooses one alternative from the choice set and the researcher observed that choice, the random variables $N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}$ is fully specified. For instance, if agent i chooses k-th alternative, then $N_{i0} = 0, \ldots, N_{i(k-1)} = 0, N_{ik} = 1, N_{i(k+1)} = 0, \ldots, N_{iJ} = 0$. *Post ante categorical distribution* with one opportunity to choose an alternative k from the choice set **as observed by the researcher** obtained from (16) is

$$\begin{aligned} &\Pr\{N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}; P_{i0}, \ldots, P_{iJ}\} \\ = \; &\frac{1!}{n_{i0}! \cdots n_{iJ}!} P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_0, \xi_0, \nu_{i0})^{n_{i0}=0} \times \cdots \times P_{i(k-1)}(\boldsymbol{d}_i, \boldsymbol{x}_{k-1}, \xi_{k^1}, \nu_{i(k-1)})^{n_{i(k-1)}=0} \\ &\times P_{ik}(\boldsymbol{d}_i, \boldsymbol{x}_k, \xi_k, \nu_{ik})^{n_{ik}=1} \\ &\times P_{i(k+1)}(\boldsymbol{d}_i, \boldsymbol{x}_{k+1}, \xi_{k+1}, \nu_{i(k+1)})^{n_{i(k+1)}=0} \times \cdots \times P_{iJ}(\boldsymbol{d}_i, \boldsymbol{x}_J, \xi_J, \nu_{iJ})^{n_{iJ}=0} \\ = \; &P_{ik}(\boldsymbol{d}_i, \boldsymbol{x}_k, \xi_k, \nu_{ik})^{n_{ik}=1} = P_{ik}(\boldsymbol{d}_i, \boldsymbol{x}_k, \xi_k, \nu_{ik}). \end{aligned} \tag{17}$$

Suppose that I agents with each agent having one opportunity to choose one alternative independently from the choice set, and that the researcher observe that alternatives $j = 0, \ldots, J$ are chosen $n_0, \ldots, n_J$ times respectively, or $N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}$. Then the *Post ante categorical distribution* **as observed by the researcher** likelihood of such an event obtained from (17) is

$$\begin{aligned} &\Pi_i^I \Pr\{N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}; P_{i0}, \ldots, P_{iJ}\} \\ = \; &\Pi_i^I \frac{1!}{n_{i0}! \cdots n_{iJ}!} P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_0, \xi_0, \nu_{i0})^{n_{i0}} \times \cdots \times P_{iJ}(\boldsymbol{d}_i, \boldsymbol{x}_J, \xi_J, \nu_{iJ})^{n_{iJ}} \\ = \; &\Pi_i^I \left[\frac{1!}{n_{i0}! \cdots n_{iJ}!}\right] P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_0, \xi_0, \nu_{i0})^{\sum_i^I n_{i0}} \times \cdots \times P_{iJ}(\boldsymbol{d}_i, \boldsymbol{x}_J, \xi_J, \nu_{iJ})^{\sum_i^I n_{iJ}}, \end{aligned} \tag{18}$$

where $\sum_i^I N_{i0} = \sum_i^I n_{i0}, \ldots, \sum_i^I N_{iJ} = \sum_i^I n_{iJ}$ are respectively the numbers of observed choices from alternative $j = 0, \ldots, J$ for this set of agent $i = 1, \ldots, I$.

The choice probabilities or parameters $P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_0, \xi_0, \nu_{i0}), \ldots, P_{iJ}(\boldsymbol{d}_i, \boldsymbol{x}_J, \xi_J, \nu_{iJ})$ specifying the probabilities of each possible outcome in (18) are constrained by the fact that each must be in the range $(0, 1)$ or $P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij}) \in (0, 1)$ for $j = 0, \ldots, 1$ for all i, and all must sum to 1, or $\sum_{j=0}^J P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij}) = 1$ for all i.

Under this setting, it is obvious that there is **redundancy in parametrization** among the $P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij})$ because $\sum_{j=0}^J P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij}) = 1$ for all i. We basically have only two choices, that is we discard one of the choice probabilities, or we focus on the contrast between the choice probability associated with alternative j to a baseline alternative, for instance, $j = 0$. Since the choice behavior of agent i is dictated by the probability of choosing alternative j $P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij})$ relative to his/her choice probability of alternative other than j or to "outside alternative" or "baseline alternative" with $j = 0$, we can assume the following without the loss of generality:

Assumption 7. We standardize observed demographics $\boldsymbol{d}_i$, observed alternative characteristic $\boldsymbol{x}_j$, unobserved alternative characteristics ξ_j, and interaction between agent and alternative ν_{ik} against a baseline alternative indexed by $j = 0$ as $\boldsymbol{d}_i = 0$, $\boldsymbol{x}_j = 0$, $\xi_j = 0$, and $\nu_{ik} = 0$.

Remark 1 In other words, **Assumption 7** impllies the representative utility V_{i0} for "outside alternative" or "baseline alternative" is zero $V_{i0} = 0$.

2.1 What standard statistics says about the issue

As seen in such standard textbooks as McCullagh and Nelder (1989) [7], each of the distributions such as the normal, Poisson, binomial, multinomial, gamma, and inverse Gaussian has a special *canonical link function* for which there exists a sufficient statistic equal in dimension to the parameter in the linear predictor.

For brevity, we suppress the arguments $\boldsymbol{d}_i$, $\boldsymbol{x}_j$, ξ_j, and ν_{ij} of P_{ij} for the moment. With this abbreviated notation, we let the probability of agent i choosing alternative j be P_{ij} in a trial with $J+1$ possible outcomes. If each N_{ij}, $i = 1, \ldots, I$ and $j = 0, \ldots, J$ taking values either 0 or 1 and the quantity $\sum_{j=0}^{J} N_{ij} = 1$ for $i = 1, \ldots, I$ denotes the number of trials for agent i resulting in choosing alternative j, $j = 0, \ldots, J$, the distribution (18) is canonically rewritten as

$$\begin{aligned}
& \Pi_i^I \Pr\{N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}; P_{i0}, \ldots, P_{iJ}\} \\
= \;& \exp\left[\sum_{i=1}^{I} n_{i1} \log\left(\frac{P_{i1}}{P_{i0}}\right) + \cdots + \sum_{i=1}^{I} n_{iJ} \log\left(\frac{P_{iJ}}{P_{i0}}\right) + I \log P_{i0}\right] \Pi_i^I \left(\frac{1}{n_{i0}! \cdots n_{iJ}!}\right) \\
= \;& \exp\left[\sum_{i=1}^{I} n_{i1} \log\left(\frac{P_{i1}}{P_{i0}}\right) + \cdots + \sum_{i=1}^{I} n_{iJ} \log\left(\frac{P_{iJ}}{P_{i0}}\right) + I \log P_{i0}\right], \qquad (19)
\end{aligned}$$

because $I = \sum_i^I \sum_{j=0}^{J} N_{ij} = \sum_i^I 1$ and $1!/(n_{i0}! \cdots n_{iJ}!) = 1$.

When the number of trials fixed, this is a $J \cdot I$-parameter exponential family with the canonicall (natural) parameter $\eta_{ij} = \log(P_{ij}/P_{i0})$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$, or

$$\boldsymbol{\eta}_i \;=\; (\eta_{i1}, \ldots, \eta_{iJ}) = \left(\log \frac{P_{i1}}{P_{i0}}, \ldots, \log \frac{P_{iJ}}{P_{i0}}\right). \qquad (20)$$

with corresponding sufficient statistics $N_{i0} = n_{i0}, \ldots, N_{iJ} = n_{iJ}$ for $i = 1, \ldots, I$ where $\sum_{j=0}^{J} n_{ij} = 1$ and n_{ij} is 1 when alternative j is chosen and the rest are 0 for any agent $i = 1, \ldots, I$.

Since $0 < P_{ij} < 1$, the natural parameter space is all $(\eta_{i1}, \ldots, \eta_{iJ})$ with $-\infty < \eta_{ij} < \infty$ in $\Re^J$ for $i = 1, \ldots, I$ and $j = 0, \ldots, J$. The ratio of agent i choosing alternative j relative to him/her choosing the baseline or outside alternative $j = 0$ is referred as his/her *odds ratio* of alternative j relative to alternative 0. Thus the canonical parameters $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iJ})$ are also *log odds ratio* of agent i choosing alternative j relative to the baseline or outside alternative.

2.2 Added insight from compositional data analysis

Alternative to *log odds ratio* characterization in (20), we could characterize the choice probabilities P_{ij} for agent i for $j = 0, \ldots, J$ as a compositional data point (or composition for short) in a simplex:

$$\mathcal{S}^{J+1_i} = \left\{ \boldsymbol{P}_i = (P_{i0}, P_{i1}, \ldots, P_{iJ}) \in \Re^{J+1} | P_{ij} > 0, j = 0, 1, \ldots, J; \sum_{j=0}^{J} P_{ij} = 1 \right\} \tag{21}$$

If we can consider the only information is given by the ratios $P_{i0}, P_{i1}, \ldots, P_{iJ}$ between components, the information of a composition is preserved under multiplication by any positive constant. Therefore, the **sample space** of compositional data $P_{i0}, P_{i1}, \ldots, P_{iJ}$ can always be assumed to be a standard simplex in (21)

Since the choice probabilities P_{ij} for agent i for $j = 0, \ldots, J$ resides in this simplex, one of the three well-characterized isomorphisms[18)] that transform from the Aitchison simplex to real space is *additive logratio transform*

$$\mathrm{alr}(\boldsymbol{P}_{i.}) = \left[\log\left(\frac{P_{i1}}{P_{i0}}\right), \log\left(\frac{P_{i2}}{P_{i0}}\right), \ldots, \log\left(\frac{P_{iJ}}{P_{i0}}\right) \right], \tag{22}$$

In other words, the utilities on the right hand side of (8), (12), and (14) are isomorphic transform of the probabilities of choosing alternative j relative to alternative 0 on the left hand side of (8), (12), and (14) from the Aithison simplex to real space respectively as follows for $i = 1, \ldots, I$ and for $j = 0, \ldots, J$:

$$\begin{aligned}
\log\left(\frac{P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j)}{P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_j)}\right) &= \boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j, \\
\log\left(\frac{P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j)}{P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j)}\right) &= \boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j + \xi_j, \\
\log\left(\frac{P_{ij}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij})}{P_{i0}(\boldsymbol{d}_i, \boldsymbol{x}_j, \xi_j, \nu_{ij})}\right) &= \boldsymbol{\alpha d}_i + \boldsymbol{\beta x}_j + \xi_j + \sum \sigma_k x_{jk} \nu_{ik}.
\end{aligned}$$

3 Conclusion and Discussion

Although the logit-type probabilities are historically derived as the difference $\epsilon_{ik} - \epsilon_{ij}$ in the idyosyncratic unobserved preference as shown in (5) by **econometricians** on the ground that this method of derivation gives a precise meaning to the probabilities, **statisticians** in generalized linear models and in compositional data analysis or Aitcheson geometry are able to show in section 2 that it is just as possible to characterize the choice situation without a single operation involving them and relate that back to the utilities in (8), (12), and (14) using the *additive logratio transform* of the utility of agent i choosing alternative

[18)] In mathematics, an isomorphism is a structure-preserving mapping between two structures of the same type that can be reversed by an inverse mapping. Two mathematical structures are isomorphic if an isomorphism exists between them.

j to the baseline or outside alternative $j = 0$. With the latter derivation, statisticians additionally gain sufficiency information as well.

Therefore statisticians are right in asking if the idyosyncratic unobserved preferences $\boldsymbol{\epsilon}$ are really needed in order to derive the logit-type choice probabilities, and hence if the argument surrounding the unobserved idiosyncratic taste ϵ_{ij} being Gumbel is needed as well.By doing so, increasingly tenuous and harder-to-justify interpretations of idiosyncratic taste ϵ_{ij} in the presence of ξjin subsection 1.5, and then in the presence of ξj and the random coefficient $(\beta_k + \sigma_k \nu_{ik})x_k$ in subsection 1.7 can be avoided. For instance, the argument "[t]he term $\boldsymbol{\xi_j}$ might be thought of as the mean of consumers' valuations of an unobserved product characteristic such as product quality, while the $\boldsymbol{\epsilon_{ij}}$ represents the distribution of consumer preferences about this mean"by Berry (1994) [1] are not necessary.

Finally, with the latter derivation, we are able to see that the statements like " [o]nly differences in utility matter " and " [t]he scale of utility is arbitrary,"are corresponding to the *redundancy in parametrization* inherent in categorical distribution of (16) and (18) in the choice probability domain. This issue is taken care of by canonical representation as in (19) using canonical parameter in (20), and by the fact that the choice probabilities for any agent must sum to unity over $j = 0, \ldots, J$.

References

[1] Berry, S. (1994), "Estimating Discrete Choice Models of Product Differentiation,"RAND Journal of Economics, Vol 25(2), 242-262.

[2] Berry, S., J. Levinsohn and A. Pakes (1995), "Automobile Prices in Market Equilibrium,"Econometrica, Vol.63, 841-890.

[3] Lancaster, K. (1971), Consumer Demand: A New Approach, Columbia University Press, New York.

[4] Luce, D. (1959), Individual Choice Behavior, John Wiley and Sons, New York.

[5] Luce, D. and P. Suppes (1965), ' Preferences, utility and subjective probability ', in R. Luce, R. Bush, and E. Galanter, eds., Handbook of Mathematical Psychology, John Wiley and Sons, New York, pp. 249–410.

[6] Marschak, J. (1960), 'Binary choice constraints on random utility indications,' in K. Arrow, ed., Stanford Symposium on Mathematical Methods in the Social Sciences, Stanford University Press, Stanford, CA, pp. 312–329.

[7] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman and Hall, London. http://dx.doi.org/10.1007/978-1-4899-3242-6.

[8] McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior,"in P. Zarembka eds. Frontiers of Econometrics, Academic Press, New York.

[9] Train, K. E. (2009), "Discrete Choice Methods with Simulation"Second Edition, Cambridge University Press.