Re-examining discrete-choice model through the lens of Aitchison geometry

Yuichiro KANAZAWA[*†]
International Christian University

1 Introduction

This article examines if statements such as "only differences in utility matter"and "the scale of utility is arbitrary"still hold water when we re-examine the discrete choice models widely employed in industrial organization and marketing science through the lens of *Aitchison geometry.* The implications of these statements are at times subtle and complex, and thus difficult. As a preparation, we review briefly three developments in this area in this section: First we review briefly how the author has been involved in the research in this area; We then briefly review the mainstream understanding of how discrete choice models can be derived from utility theory; Finally we briefly review the implications placed on the utility in order to justify the derivations of discrete choice models.

1.1 New Empirical industrial organization: A strand of history the author has been involved

Industrial organization and marketing science literature is concerned with the structure of industries in the economy and the behavior of firms and individuals in these industries. A dramatic shift in the 1980s toward what Bresnahan (1989) coined the "New Empirical Industrial Organization (NEIO),"which tries to takes advantage of the fact that individual industries are sufficiently distinct, and industry details are sufficiently important.

In NEIO, products are regarded as bundles of characteristics, and preferences are defined on those characteristics, so each consumer chooses a bundle that maximizes its utility. Consumers have different preferences for different characteristics, and hence make different choices, however. In other words, consumers are heterogeneous.

Simulation is used to obtain aggregate demand from the heterogeneous consumers' choices in the following manner: first we draw vectors of consumer characteristics from the distribution of those characteristics; second we determine the choice probability that

[*]yukanazawa@icu.ac.jp

[†]We are grateful for financial support from the Japan Society for the Promotion of Science under the Grant-in-Aid for Scientific Research (C) 20K01595 and from the Research Institute for Mathematical Sciences, a Joint Usage/Research Center located in Kyoto University.

each of the drawn consumers would make for a given value of the parameter; now we aggregate those choice probabilities into a prediction for aggregate demand conditional on the parameter vector; finally we employ a search routine that finds the value of that parameter which makes these aggregate probabilities as close as possible to the observed market shares.

The theoretical and econometric groundwork for product characteristic based demand systems dates back to the work of Lancaster (1971) [15] and McFadden (1974, 1981) [17],[18]. Applications of the Lancaster/McFadden framework, however, increased significantly after Berry, Levinsohn, and Pakes (1995, henceforth BLP(1995) [4]) showed how to circumvent a problem that had made it difficult to apply the early generation of characteristic-based models in I.O. contexts.

The early generation of models did not allow for unobserved product characteristics. Consumer goods typically are differentiated in many ways. As a result even if the econometricians measured all the relevant characteristics, they could not expect to obtain precise estimates of their impacts. One solution proposed in Berry (1994) [3] is to put in the " important " differentiating characteristics and an unobservable, say ξ, which picks up the aggregate effect of the multitude of characteristics that are being omitted.

Often the econometricians find that there is not enough information in product level demand data to estimate the entire distribution of preferences with sufficient precision. This should not be surprising given that they are trying to estimate a whole distribution of preferences from just aggregate choice probabilities. The literature has added information in two ways: One is to add an equilibrium assumption and work out its implications for the estimation of demand parameters, the other is to add more data.

It is not surprising that when the pricing system is added to the demand system, the precision of the demand parameters estimates tends to improve (see, BLP (1995) [4]). Almost all of it has assumed static or myopic profit maximization, and that one side of the transaction has the power to set prices while the other can only decide whether and what to buy conditional on those prices. However, models in marketing science started looking one period ahead (see, Che, et al. (2007) [8], Kamai and Kanazawa (2016) [13]) with interactions between manufacturers and a retailer factored in.

On adding more data, there are a number of types of micro data that might be available: Surveys that match individual characteristics to a product chosen by the individual (point-of-sales data); Surveys providing information on the proprietary consumer's second choice (Berry, Levinsohn, and Pakes (2004) [5]); Alternatively, the econometricians or market scientists may have access to summary statistics that provide information on the joint distribution of consumer and product characteristics (Petrin 2002 [25]). Petrin (2002) proposes a technique for obtaining more precise estimates of demand and supply curves when the econometricians or market scientists are constrained to market-level data. The technique allows them to augment market share data with information relating the average demographics of consumers to the observable characteristics–Myojo and Kanazawa (2012) called "discriminating attributes"–of products they purchase that determines a subset of products in the market. Petrin (2002) states that "[t]his extra information plays the same role as consumer-level data..."(p.705, [25]).

Berry, Linton, and Pakes (2004) [6] provides asymptotic theory of the estimate of the demand system objective function as the number of products increases in one (national) market. The paper shows that, provided one accounts for simulation and sampling error in the estimate of the objective function, standard approximations to the limit distribution work (see, for e.g. Pakes and Pollard, 1989 [23]). Myojo and Kanazawa (2012) [19] provides a sharper asymptotic variance- covariance of the estimate of the demand than Berry, Linton, and Pakes (2004) by adding the pricing equation and micro moment objective functions as the number of products increases in one (national) market.

For durable goods like automobiles, the number of models increased from somewhere in the 150's in 1980s to in the 260's in 2018. Asymptotic theories of Berry, Linton, and Pakes (2004) and Myojo and Kanazawa (2012) presuppose such a market. On the other hand, for many non-durable consumer good products, the number of product offering is limited because of the limited shelf space at the retail outlets. In Nevo (2001), for example, there are 1124 markets and 24 products.[1)] Therefore a different type of asymptotics, the one in the number of markets, is needed.

Freyberger (2015) [12] provides asymptotically normal *biased* parameter estimate (that depends on the number m of markets, the fixed number R of simulation draws from the distribution of consumers to calculate the observed market share of the demand system objective function) for a fixed number of products as the number of markets increases. It also provides how the leading asymptotic bias terms can be eliminated by using an *analytic bias correction* method.

Asymptotic bias is generated because, for each market, its participating households are self-selected and unique in its own way. As a result simply increasing the number of such markets will not do. One general estimation idea when you have estimates in many comparable but heterogeneous subgroups within a population is to combine the individual estimates, each unbiased, to manufacture an overall unbiased estimate, using the local variances and overall variance between subgroup means to select the best linear estimator. An approach for the current problem along this idea is to alter the number of simulation draws from the fixed R in Freyberger (2015) to market-dependent R_m to reflect the population. This approach should work in theory, but a proper choice of R_m presupposes we know the market-by-market variation of the consumer preferences, the exact knowledge we are trying to estimate. In reality, different portions of the population may still be over- or under-represented as we increase the number m of markets.

If we wish to have a data-driven method as an alternative to analytic (asymptotic) bias correction proposed by Freyberger (2015), however, there is another idea we can employ (for non-durable product markets with a limited number of product offering), however. That is, for these markets, information that encompasses many regions are available, and we can take advantage of such information to adjust the bias. In Kanazawa (2018) [14], we show that we can pursue the second idea, namely, we can achieve the data-driven asymptotic bias correction by incorporating 1) the pricing (profit maximizing) equation for *national* suppliers and 2) the *national* micro moment regarding consumers as the

[1)] We study non-durable goods such as ready-to-eat cereals because serious policy implications abound for markets of such non-durable goods: For instance, Nevo(2001) [21] claims that "Previous researchers have concluded that the ready-to-eat cereal industry is a classic example of an industry with nearly collusive pricing behavior and intense non-price competition"(p.307).

number m of markets increases.

1.2 Discrete Choice Models: A review

All the works in NEIO explained in section 1.1 assumes the so-called "discrete choice models." We now review "discrete choice models"in more detail.

A person, a household, a firm, or a decision maker faces a choice, or a series of choices over time, among a set of options. For instance, a consumer chooses one product among several competing products; a household purchases a vehicle from the many models offered by a variety of manufacturers; a firm decides which technology to invest and to use in production; a student chooses one (hopefully correct) answer on a multiple-choice test; a respondent to a survey chooses a response ranging from 1 to 7 on a Likert-scale question. Our purpose is to understand the behavioral process that leads to the person's, the household's, the firm's or the decision maker's choice.

For the sake of brevity, we henceforth use the term 'agent' for these decision makers, be it a person, a household, a firm, or a student, or a respondent to a survey. The agent himself/herself knows exactly all the factors that collectively determine his or her choice. Some of these factors are observed by the researcher but the others are not. We label the factors observed by all the agents as well as the researcher as $\boldsymbol{X}$, and those observed by the agents but unobserved by the researcher as ϵ. These factors jointly lead the agent i to choose one alternative k. Therefore agent i's behavioral process is expressed by a function

$$h(\boldsymbol{X}, \epsilon) = \text{choosing alternative } k.$$

In this sense, we look at this choice situation from a causal perspective and assume that there are factors that collectively determine, or cause, the agent's choice. In other words, it is deterministic in the sense that given $\boldsymbol{X}$ and ϵ, the choice of the agent i is fully determined.

However the researcher does not observe ϵ_{ij}, where in this article i indexes the agent $i = 1, \ldots, I$ and j indexes the alternative $j = 1, \ldots, J$, and thus cannot predict the agent's behavior precisely. As a result, the researcher is forced to assume ϵ_{ij} to be random due to the lack of evidence shown to be otherwise, and to have density $f(\epsilon_{ij})$. The probability that the agent i chooses a particular alternative k from the set of all possible outcomes indexed by j is simply the probability that the unobserved factors are such that the behavioral process results in that outcome

$$\Pr\left\{\epsilon \text{ s.t. } h(\boldsymbol{X}, \epsilon) = \text{choosing alternative } k\right\}.$$

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

1.3 Random utility models

We now rephrase what we described in section 1.2 in terms of random utility models used extensively in NEIO and marketing science. 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 some attributes of the alternatives presumed to be in the mind of the agent when he/she makes decision among alternatives and at the same time observable by the reseacher, labeled $\boldsymbol{x}_j$, and some attributes such as demographics of the agent observable by the researcher, labeled $\boldsymbol{d}_i$. 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 $\boldsymbol{x}_j$ and $\boldsymbol{d}_i$ expressed as

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

is called *representative utility.*

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

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

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

$$U_{ij} = V_{ij} + \epsilon_{ij}, \tag{1}$$

where ϵ_{ij} captures all the factors that jointly affect utility as perceived by agent i for alternative j, but are not included in V_{ij}. It is sometimes referred as the agent's *idiosyncratic utility.* it is defined according to the researcher's representation of the choice situation the agents are facing in a setting.

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 that agent i chooses alternative k is

$$\begin{aligned} P_{ik} &= \Pr\left\{U_{ik} > U_{ij} \quad \forall j \neq i\right\} \\ &= \Pr\left\{V_{ik} + \epsilon_{ik} > V_{ij} + \epsilon_{ij} \quad \forall j \neq i\right\} \\ &= \Pr\left\{\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik} \quad \forall j \neq i\right\}, \end{aligned} \tag{2}$$

where we use the notation {statement} as the indicator function taking 1 if the statement is true and 0 otherwise as Bruno de Finetti did. This probability is cumulative in the sense that the probability that each random term $\epsilon_{ik} - \epsilon_{ij}$ is below the observed quantity

$V_{ij} - V_{ik}$. Using the density $f_{\boldsymbol{\epsilon}_{i.}}(\cdot)$, this cumulative probability in (2) can be rewritten as

$$\begin{aligned} P_{ik} &= \Pr\left\{\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik} \quad \forall j \neq k\right\}. \\ &= \int_{\boldsymbol{\epsilon}_{i.}} \left\{\epsilon_{ik} - \epsilon_{ij} < V_{ij} - V_{ik} \quad \forall j \neq k\right\} f_{\boldsymbol{\epsilon}_{i.}}(\cdot) d\epsilon_{i.} \\ &= \int_{\{\epsilon_{ik}-\epsilon_{ij}<V_{ij}-V_{ik} \quad \forall j\neq k\}} f_{\boldsymbol{\epsilon}_{i.}}(\cdot) d\epsilon_{i.}. \end{aligned} \tag{3}$$

Note that this is a $J-1$-dimensional integral of the differences of random variables $\epsilon_{ik} - \epsilon_{ij}$ for all j, $j = 1, \ldots, J$ but $j \neq k$ over the regions the differences are taking less than $V_{ij} - V_{ik}$ for all j, $j = 1, \ldots, J$ but $j \neq k$.

Different discrete choice models are obtained from different specifications of this density. *Logit* model and *nested logit* model are known to have closed-form choice probabilities for this integral. 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 (extreme value) and a type of generalized Gumbel (extreme value), respectively. *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 evaluated numerically through simulation.

1.4 Model-based Market shares

Model-based market shares as aggregate outcome variables can be obtained consistently from discrete choice models either by sample enumeration or segmentation. For the current purpose, we only briefly discuss how the the probability of agent i choosing alternative j be P_{ij} as in (3) can be utilized.

In sample enumeration, the choice probabilities of each agent in a sample are averaged over the agents under consideration. Suppose that we draw a sample of N agents, from the population of $i = 1, \ldots, I$ agents for which the model-based aggregate market share is to be compared with the estimated market share. Depending on the sampling scheme, we associate some weight w_i for each agent. This weight represents the number of agents in a population similar to the agent. When the sample is randomly drawn, then w_i's are the same and if it is stratified, then w_i's are the same within a stratum. We obtain a consistent estimate of the average number of agents in the population who choose alternative i, denoted by $\hat{P}_{.j}$, by the weighted sum of the individual choice probabilities $\hat{N}_{.j} = \sum_{i=1}^{N} w_i P_{ij}$ divided by the number N of the sample. This average probability is the model-based simulated market share and written as

$$\hat{P}_{.j} = \frac{\hat{N}_{.j}}{N} = \frac{\sum_{i=1}^{N} w_i P_{ij}}{N}. \tag{4}$$

1.5 Multinomial logit model

The most widely employed discrete choice model is logit. Its popularity is due to the fact that the formula for the choice probabilities takes a closed form and is readily interpretable. It was first derived by Luce (1959) [16] with the assumption on the characteristics called the *independence from irrelevant alternatives* (IIA) of choice probabilities.

The logit model is obtained by assuming that each ϵ_{ij} in (1) is independently, identically distributed as a random variable from Gumbel or type I extreme value distribution. The key assumption is that the errors are independent of each other, meaning that the unobserved portion of utility for one alternative is unrelated to the unobserved portion of utility for another alternative. In other words, employing this assumption implies that the error for one alternative provides no information to the researcher about the error for another alternative. One can go one step further and claim employing this assumption is tantamount to the researcher is able to specify the representative utility V_{ij} so well that the remaining, unobserved portion of utility can be treated essentially as "random."Some authors think this assumption is fairly restrictive and many models have been derived to allow for correlated errors between these two unobserved portions of utilities for different alternatives.

The cumulative distribution function of Gumbel or type I extreme value is

$$F_{\epsilon_{ij}}(\epsilon_{ij}) = \exp\left(-\exp\left(-\epsilon_{ij}\right)\right). \tag{5}$$

The difference between two independent extreme value variables is known to be distributed as logistic distribution with its cumulative distribution function as

$$F_{\epsilon_{ikj}}(\epsilon_{ikj}) = \frac{\exp\left(\epsilon_{ikj}\right)}{1+\exp\left(\epsilon_{ikj}\right)}, \tag{6}$$

where $\epsilon_{ikj} = \epsilon_{ik} - \epsilon_{ij}$.

With some algebraic manipulation of the integral in (3), we arrive at the closed form expression of logit choice probabilities for agent i towards alternative k as

$$P_{ik} = \frac{\exp\left(V_{ik}\right)}{\sum_{j=1}^{J} \exp\left(V_{ij}\right)}. \tag{7}$$

Limitations of logit model is summarized on pages 42-43 in Train (2009) [28]:

> The value or importance that agent places on each attribute of the alternatives varies, in general, over agents. For example, the size of a car is probably more important to households with many members than to smaller households. Low-income households are probably more concerned about the purchase price of a good, relative to its other characteristics, than higher-income households. In choosing which neighborhood to live in, households with young children will be more concerned about the quality of schools than those without children, and so on. Decision makers' tastes also vary for reasons that are not linked to observed demographic characteristics, just because different people are different. Two people who have the same income, education, etc., will make different choices, reflecting their individual preferences and concerns.

The same author also states on page 43 of Train (2009) [28]:

> Logit models can capture taste variations, but only within limits. In particular, tastes that vary systematically with respect to observed variables can be incorporated in logit models, while tastes that vary with unobserved variables or purely randomly cannot be handled.

1.6 Implications on the utility for justifying the derivations of discrete choice models

On the aspects of behavioral decision process that affect the specification and estimation of any discrete choice model, Train (2009, p.19) [28] claims that "Only differences in utility matter"and "The scale of utility is arbitrary."He also claims that "[t]he implications of these statements are far-reaching, subtle, and, in many cases, quite complex."Specifically, he states in Train (2009, p.19) with n (instead of i in our case) indexes the agent and i and j (instead of j and k in our case) index the alternative that

> 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 $U_{nj}\ \forall j$ as with $U_{nj}+k\ \forall j$ for any constant k. A colloquial way to express this fact is, "A rising tide raises all boats."
>
> The level of utility doesn't matter from the researcher's perspective either. The choice probability is $Pni = \text{Prob}(U_{ni} > U_{nj} \forall j \neq i) = \text{Prob}(U_{ni} - U_{nj} > 0 \forall j \neq i)$, which depends only on the difference in utility, not its absolute level.

Similarly, he states in Train (2009, p.23) again with n (instead of i in our case) indexes the agent and i and j (instead of j and k in our case) index the alternative that

> Just as adding a constant to the utility of all alternatives does not 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 $U^0_{nj} = V_{nj} + \epsilon_{nj}\ \forall j$ is equivalent to $U^1_{nj} = \lambda V_{nj} + \lambda\epsilon_{nj}$ $\forall j$ for any $\lambda > 0$. To take account of this fact, the researcher must normalize the scale of utility.

We will examine if these statements still hold water when we see the discrete choice models through the lens of *Aitchison geometry.*

2 Aitchison Geometry

In the following, we mainly use the logit choice probability as a concrete example to reexamine the derivation described in sections 1.3 and 1.5. The reexamination is necessary because we feel that the derivation is heavily influenced by its unconscious choice of Euclidean metric. The nature of the argument does not alter fundamentally, however, if we employ probit, nested logit, probit, or mixed logit models instead of logit model.

2.1 Compositional data

Certain types of multivariate data were constrained like proportions or percentages and their sum must be a fixed constant such as 1 or 100, respectively. Obviously the agent i's probability of choosing alternative k in (7) is a proportion with a fixed constant sum constraint $\sum_{j=1}^{J} P_{ij} = 1$. In such case, the researcher needs to ask if *relative* rather than *absolute* information is relevant for the analysis. Here relative information refers to a representation of quantitatively described contributions on a whole. Information about the total amount itself is irrelevant. For instance, in case of household expenditure on food, housing, transportation, and communications, the researcher may not necessarily be interested in the wealth of the household expressed by the actual amounts on each of those expenditure items expressed in terms of JPY, USD, or Euro, but rather interested in the proportion of the income spent on those categories. In those cases, it is more natural to consider them as observations carrying relative information. One could argue, therefore, that all relevant information in this type of data is contained in unit-less ratios (of course, taking ratios cancel out the unit) between components using one of the categories as a baseline.

Such constrained data for which the researcher is mainly interested in unit-less ratios between components are called *compositional data.* We quote from Filzmoser et al. (2018, p.11) [11] below

> ...a compositional vector, or simply a composition, $\boldsymbol{x} = (x_1, \ldots, x_D)^{`}$ with D parts (arranged into a column vector) is by definition a positive real vector with D components, describing quantitatively the parts of some whole, which carry relative information between the parts.

Egozcue (2009) [9] states that compositional data analysis should respect the following principles:

Scale invariance: The information in a composition does not depend on the particular units in which the composition is expressed. Proportional positive vectors represent the same composition. Any sensible characteristic of a composition should be invariant under a change of scale. This principle thus corresponds to the fact that a multiplication of a compositional vector by an arbitrary positive number does not alter the ratios between compositional parts.

Permutation invariance: Permutation of parts of a composition does not alter the information conveyed by the compositional vector, similarly as in standard multivariate statistics.

Subcompositional coherence: Information conveyed by a composition of D parts should not be in contradiction with that coming from a subcomposition (i.e., a subvector of the original compositional vector) containing d parts, $d < D$. This principle can be formulated more precisely as

Subcompositional dominance: If $\triangle_p(\boldsymbol{x}, \boldsymbol{y})$ is any distance between compositions of p parts, then

$$\triangle_D(\boldsymbol{x}, \boldsymbol{y}) \geq \triangle_d(\boldsymbol{x_d}, \boldsymbol{y_d}),$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ are compositions with D parts and $\boldsymbol{x_d}$, $\boldsymbol{y_d}$ are subcompositions of the previous ones with d parts, $d < D$.

Ratio preserving: Any relevant characteristic expressed as a function of the parts of a composition is exclusively a function of the ratios of its parts. In a subcomposition, these characteristics depend only on the ratios of the selected parts and not on the discarded parts of the parent composition. Scale invariance applies to the subcomposition.

2.2 Should the choice probabilities or market shares treated as compositional data?

In the following, we examine if it is reasonable to require those three characteristics—scale invariance, permutation invariance, and subcompositional coherence—for the choice probabilities based on the discrete choice models we introduced in section 1.5. If so, we can treat the choice probabilities or market shares in NEIO and marketing science as compositional. Furthermore, we need to examine the appropriateness of Euclidean metric.

First, we recognize that the choice probabilities of agent i over all alternatives j, $j = 1, \ldots, J$ must add up to 1. Therefore P_{ij}, $j = 1, \ldots, J$ in (7) have to exist in the so-called $(J-1)$-standard simplex—a generalization of the notion of a triangle or a tetrahedron to higher dimensions—that is a subset of the J-dimensional real space $\Re^J$:

$$\left\{ \mathbf{P}_i = (P_{i1}, \ldots, P_{iJ}) \in \Re^J | P_{ij} \geq 0, \sum_{j=1}^{1} P_{ij} = 1 \right\}.$$

Similarly, we recognize the simulated estimate of the market share $\hat{P}_{.j}$ for alternative $j = 1, \ldots, J$ in (4) must also exist in the $(J-1)$-standard simplex because all the components P_{ij} from which $\hat{P}_{.j}$ is computed are in the $(J-1)$-standard simplex as well and the weights in (4) are so set that the resulting $\hat{P}_{.j}$ does not go beyond the area of this simplex. In addition, it is obvious that the corresponding observed market shares $\mathbf{S} = (s_1, \ldots, s_J)$, where s_j, $j = 1, \ldots, J$ is the market share of alternative j, must also exist in the $(J-1)$ simplex as well.

Second, on *scale invariance*. In Table 1, we present the observed sales of passenger vehicles in March 2020 in Japan in terms of units sold as well as in terms of maket share. As discussed in case of household expenditure on food, housing, transportation, and communications in section 2.1, the researcher is not likely to be interested in the actual units sold or the absolute level of the market share. Rather the researcher is interested in which models were sold among these models listed in Table 1 during March 2020 and why so, given the researcher's assumption that the agents were to buy one vehicle in March, 2020, and given their product characteristics including their prices. Or including not choosing any alternative often referred as “outside good”or “outside alternative,”the researcher may be interested why certain models were bought at all if the agent was allowed to have an option of not purchasing any vehicle during March, 2020. Either way, we need to ensure that the analyses based on the market share be the consistent with the analysis based on the sales volume.

Model Name	Manufacturer	Sales in Units	Market share
Carolla	Toyota	16,327	0.05072624
Fit	Honda	14,845	0.04612182
Yaris	Toyota	13,164	0.04089913
Raize	Toyota	12,009	0.03731067
Note	Nissan	10,999	0.03417271
Sienta	Toyota	10,456	0.03248567
Prius	Toyota	9,717	0.03018968
Roomy	Toyota	9,700	0.03013686
Freed	Honda	9,528	0.02960247
Selena	Nissan	9,130	0.02836593
Voxy	Toyota	8,963	0.02784708
Aqua	Toyota	8,488	0.02637130
Tank	Toyota	8,261	0.02566604
Alphard	Toyota	7,885	0.02449785
RAV4	Toyota	6,286	0.01952993
Solio	Suzuki	5,702	0.01771550
Noah	Toyota	5,649	0.01755084
CX-30	Mazda	5,647	0.01754462
MAZDA2	Mazda	5,616	0.01744831
Imprezza	SUBARU	5,459	0.01696053
⋮	⋮	⋮	
Total		321,865	1.0

Table 1: Japanese Automobile Sales in March 2020

Relative information between the parts as described in the block quotation from Filzmoser et al. (2018, p.11) [11] above can also be applied to the ratios between the units sold or the market shares in an example in Table table:JapaneseAutomobileSalesinMarch2020. For example, using the best selling Toyota Carolla as baseline, we obtain the ratios of the popularity for Fit relative to Carolla by 14845/16327, for Yaris relative to Carolla by 13164/16327, and so on if we employ the units sold. Obviously, computing these ratios from the corresponding market share data gives exactly the same value. Overall, there are $\binom{20}{2} = 190$ ratios, up to their reciprocals, which form this representation of relative information, if we are only interested in the Top 20 selling vehicles in Table 1. If, on the other hand, we include outside alternative, but are limiting our interest to the Top 20 selling vehicles in Table 1, this ratios will be increase to $\binom{21}{2} = 210$ ratios, again up to their reciprocals. Filzmoser et al. (2018, p.p.3-4) [11] claim:

> One can see that ratios contain much more detailed information than just percentages to the total, and they remain the same if the data are rescaled. Ratios will thus form the representation of relative information that is considered in compositional data analysis.

Third, on *permutation invariance.* In the example in Table 1, it is obvious that permutation of parts of a composition does not alter the information conveyed by the compositional vector. In other words, if those models are presented in the increasing order of units sold or market share, or in any random order, the researcher is still interested in either which models are popular, or why certain model is bought at all.

Fourth, on *subcompositional coherence.* In the example in Table 1, it is also obvious that information conveyed by a composition of 20—21 if outside alternative is included—models should not be in contradiction with that coming from a subcomposition containing d parts, $d < 20$ in the sense that the ratios are preserved. It is not clear, however, *subcompositional dominance* is guaranteed or not and we need to question applicability of Euclidean metric for compositional data.

2.3 Is Euclidean metric appropriate for choice probabilities or market shares?

Subcompositional dominance asserts that the distance computed between two compositions cannot be less than the distance between the corresponding subcompositions. However, there are many distance measure we can employ. The very insightful counterexample in Filzmoser et al. (2018, p.13) [11] shows us that, for compositional data, the Euclidean metric is not appropriate:

> Consider two compositions $\boldsymbol{x} = (0.55, 0.40, 0.05)^{‘}$ and $\boldsymbol{y} = (0.10, 0.80, 0.10)^{‘}$, expressed in proportional representation. Their Euclidean distance is $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(0.55-0.10)^2 + (0.40-0.80)^2 + (0.05-0.10)^2} = 0.604$. When computing the Euclidean distance between the vectors consisting of the first two components, $\sqrt{(0.55-0.10)^2 + (0.40-0.80)^2} = 0.602$, everything seems to work well. But the point is that such a property should be fulfilled for any representation of

these subcompositions. If the subcompositions are expressed as proportions, i.e. $0.55/(0.55+0.40)$, etc., resulting in $\boldsymbol{x_s} = (0.579, 0.421)^{`}$ and $\boldsymbol{y_s} = (0.111, 0.889)^{`}$, their Euclidean distance is 0.661, what clearly contradicts the assumption of subcompositional dominance.

2.4 What metric is appropriate for choice probabilities or market shares?

In NEIO or marketing science papers discussed in section 1.1, the researcher is rarely interested in agent i's probability P_{ik} of choosing alternative k in (7) per se, but usually interested in the P_{ik} relative to his/her choice probability of either alternative other than k or to "outside alternative."For outside alternative, which we index $j=0$, it is natural to set their representative utility as $V_{i0}=0$ because outside alternative has no attributes and the attributes of the agent facing the alternative $k=0$ is hardly observable by the researcher. The logit probability of agent i choosing outside alternative $k=0$ is thus $P_{i0} = \exp(0) / \sum_{j=1}^{J} \exp(V_{ij}) = 1/\sum_{j=1}^{J} \exp(V_{ij})$. The ratio of agent i choosing alternative k relative to him/her choosing the outside alternative $k=0$, sometimes referred as *odds ratio* in statistics, is thus

$$\frac{P_{ik}}{P_{i0}} = \frac{\exp(V_{ik})}{\exp(V_{i0})} = \exp(V_{ik} - V_{i0}) = \exp(V_{ik}). \tag{8}$$

Expression in (8) reveals that the representative utility V_{ik} of agent i choosing alternative k when first presented in section 1.3 is actually relative to the outside alternative $V_{i0}=0$. More importantly, if employ the natural logarithm to (8), we can recover the difference in the representative probabilities as

$$\log\left(\frac{P_{ik}}{P_{i0}}\right) = V_{ik} - V_{i0}. \tag{9}$$

This fact that log odds ratio applied to the choice probabilities obtains the difference in representative utilities of agent i choosing alternative k and outside alternative $k=0$ explicitly and strongly points to the logratio transformation of the following type

$$\log\left(\frac{P_{i1}}{P_{i0}}\right), \quad \log\left(\frac{P_{i2}}{P_{i0}}\right), \quad \ldots, \quad \log\left(\frac{P_{iJ}}{P_{i0}}\right), \tag{10}$$

if we employ outside alternative $j=0$. From this derivation, it becomes apparent that statement "only differences in utility matter"in Train (2009, p.19) [28] needs modified to "only differences in *representative* utility matter."We also learn from this derivation that differences in *representative* utility can be measured through this logratio transformation. Furthermore, it also becomes apparent that the statement "the overall scale of utility is irrelevant"must be questioned because the logratio transformed choice probabilities give the actual numerical value of the differences in representative utility. The expression in (10) is called *additive logratio coordinates.*

2.5 Aitchison Geometry on the Simplex

In sections 2.3 and 2.4, we learned neither the choice probabilities or the market shares does not follow the usual Eucledean geometry. We also see that the sample space of compositions is the simplex in section 2.2. Therefore an appropriate geometrical concept needs to be developed. Although the work of Aitchison (1986) [1] is pioneering, it did not include the geometrical perspective of compositional data analysis. The geometrical structure of compositions is examined and referred to as the *Aitchison geometry* in Pawlowsky-Glahn and Egozcue (2001) [24] and Egozcue et al. (2003) [10]. Their work is designed to define a vector space structure of the simplex.

They first introduced basic operations—*perturbation* and *powering*—corresponding to the addition of two vectors, i.e. the shifting operation, and multiplication of a vector by a real number in the Euclidean geometry.

For two compositions $\boldsymbol{x}$ and $\boldsymbol{y}$ from the simplex sample space $\tilde{S}^D$, the perturbation of $\boldsymbol{x}$ and $\boldsymbol{y}$ is a composition defined as

$$\boldsymbol{x} \oplus \boldsymbol{y} = (x_1 y_1, \ldots, x_J y_J)^T. \tag{11}$$

The power transformation of a composition $\boldsymbol{x} \in \tilde{S}^D$ by a constant $c \in \Re$ is defined as

$$c \odot \boldsymbol{x} = (x_1^c, \ldots, x_J^c)^T. \tag{12}$$

These two operations are sufficient to obtain a vector space, and the usual commutative, associative, distributive properties are maintained. For instance, the perturbation difference is obtained as

$$\boldsymbol{x} \ominus \boldsymbol{y} = \boldsymbol{x} \oplus [(-1) \odot \boldsymbol{y}] = (x_1/y_1, \ldots, x_J/y_J)^T, \tag{13}$$

and

$$\boldsymbol{x} \ominus \boldsymbol{x} = \boldsymbol{x} \oplus [(-1) \odot \boldsymbol{x}] = (x_1/x_1, \ldots, x_J/x_J)^T = (1, \ldots, 1) = \boldsymbol{n}, \tag{14}$$

has all pairwise logratios equal to zero and corresponds to the zero vector in the Euclidean geometry.

A Euclidean vector space structure is obtained when we define norm, inner product, and distance in the *Aitchison* sense. Let $\mathbf{x} = (x_1, \ldots, x_J)^T \in \tilde{S}^D$、 $\mathbf{y} = (y_1, \ldots, y_J) \in \tilde{S}^D$ be two compositions. Then *Aitchison inner product*, *Aitchison norm*, and *Aitchison distance* is respectively defined in the following:

$$< \boldsymbol{x}, \boldsymbol{y} >_A = \frac{1}{2d} \sum_{i=1}^{J} \sum_{j=1}^{J} \ln\left(\frac{x_i}{x_j}\right) \cdot \ln\left(\frac{y_i}{y_j}\right), \tag{15}$$

$$\|\boldsymbol{x}\|_A = \sqrt{< \boldsymbol{x}, \boldsymbol{y} >_A} = \sqrt{\frac{1}{2d} \sum_{i=1}^{J} \sum_{j=1}^{J} \left(\ln\left(\frac{x_i}{x_j}\right)\right)^2}, \tag{16}$$

$$d_A(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{2d} \sum_{i=1}^{J} \sum_{j=1}^{J} \left(\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right) \right)^2}, \tag{17}$$

which is equivalent to

$$d_A(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{2d} \sum_{i=1}^{J} \sum_{j=1}^{J} \left(\ln\left(\frac{x_i}{y_i}\right) - \ln\left(\frac{x_j}{y_j}\right) \right)^2}, \tag{18}$$

where the expression in (18) shows that a component of $\boldsymbol{x}$ is compared with the corresponding component of $\boldsymbol{y}$.

These definitions lead to a Euclidean linear vector space structure, and in the literature this is simply denoted by the *Aitchison geometry* on the simplex. As can be clearly seen in those definitions, the definitions in (15), (16), (17), and (18) are all based on logarithms of ratios (logratios) between the compositional parts. It is critically important also to note that in computing the distance between two compositions either by (17) or by (18), all the possible permutations of the logratios are employed.

3 Conclusion and Discussion

We learned that the choice probabilities—individual or aggregated—and the market shares need to be regarded as compositional because, in NEIO and marketing science, the major interest is to uncover the behavioral processes of all the agents i, $i = 1, \ldots, I$ choosing their respective alternative j, $j = 1, \ldots, J$. Therefore, when we manipulate the choice probabilities or the market shares, we need to transform it to the logratios to caluculate inner product, norm, and distance. For instance, any attempt including those papers in subsection 1.1 utilizing the generalized method of moment and simulations—BLP(1995) [4], Petrin (2002) [25], Berry, Levinsohn, and Pakes (2004) [5], Berry, Linton, and Pakes (2004) [6], and Myojo and Kanazawa (2012) [19]—to estimate the parameter associated with P_{ij} by matching the observed market share $\mathbf{S} = (s_1, \ldots, s_J)$ with the model-based simulated estimate of the market share $\hat{\mathbf{P}} = \left(\hat{P_{i1}}, \ldots, \hat{P_{iJ}}\right)$ should be "staying-in-the-simplex approach"and should be based on the logratios.

Furthermore, we learned that the logratio transformed choice probabilities at least for logit models are the difference in *representative* utilities. As such, logratio transformed choice probabilities are intuitive and highly interpretable because its direct and straightforward connection to the utility theory. In the process, we learned that we needed to modify the prevalent wisdom "only differences in utility matter"to "only differences in representative utility matter"at least for logit model because its logratio transformed choice probabilities are differences in representative utility. Similarly, we found the statement "the overall scale of utility is irrelevant"to be highly questionable because the logratio transformed choice probabilities give the actual numerical value of the differences in representative utility.

To the best of the author's knowledge, this principle of Aitchison geometry has never been applied systematically to these fields of research.

Some significant amount of work needs to be done in the areas of statistics/econometrics and microeconomics to fully understand the potential of compositional analysis if it has to become a toolbox of choice. First it is often reasonable and is widely practiced to specify the representative utility to be linear in parameters possibly with an alternative-specific constant:

$$V_{ij} = c_j + \boldsymbol{X}_{ij}\boldsymbol{\beta},$$

where the $\boldsymbol{X}_{ij}$ is a $J \times p$ matrix of variables that relate to alternative j, $j = 1, \ldots, J$ as faced by agent i, $\boldsymbol{\beta}$ is the parameter vector consisting of p coefficients corresponding to these p column vectors in $\boldsymbol{X}_{ij}$variables. In section 2.4 and especially from (9), we learned, with the representative utility V_{i0} for outside alternative set to 0, that

$$\log\left(\frac{P_{ik}}{P_{i0}}\right) = V_{ik} - V_{i0} = c_j + \boldsymbol{X}_{ij}\boldsymbol{\beta}.$$

Maximum likelihood is usually employed to estimate the parameter $\boldsymbol{\beta}$ for the model. If you look at the methods carefully, however, we find it is basically iteratively-reweighted least-square methods. We certainly need to examine the validity and applicability of these methods now that we know the appropriate distance measure is not Euclidean, but rather d_A in (17) or in (18).

Second, and obviously, we need to investigate to what extent the intuitive and highly interpretable nature of logratio transformed choice probabilities will be compromised if we employ discrete choice models other than logit such as the nested logit, probit, or mixed logit models. For this, we need to theoretically investigate what happens to the differences in representative utility when nested logit, probit, or mixed logit models are estimated on the same data.

Third, we need to propose new, intuitive, and highly interpretable methods of market segmentation based on Aitchison Geometry.

Lastly, we need to address the problem of *structural zeros* or *essential zero* in Aitchison Geometry. Aitchison and Kay (2003) [2] states the definition of *essential zero* as

> by an essential zero we mean a component which is truly zero, not something recorded as zero simply because the experimental design or the measuring instrument has not been sufficiently sensitive to detect a trace of the component.

References

[1] Aitchison, J. (1986), The Statistical Analysis of Compositional Data (Chapman & Hall, London, 1986). Reprinted in 2003 with additional material by The Blackburn Press.

[2] Aitchison, J. and Kay, J. (2003) "Possible solution of some essential zero problems in compositional data,"in *Proceedings of CoDaWork' 03, The 1st Compositional Data Analysis Workshop*, ed. by S. Thi-Henestrosa, J.A. Martn-Fernndez (University of Girona, Girona. CD-ROM.

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

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

[5] Berry, S., Levinsohn, J. and Pakes, A. (2004), “Estimating Differentiated Product Demand Systems from a Combination of Micro and Macro Data: The New Car Model,”Journal of Political Economy, vol. 112(1), 68-105.

[6] Berry, S., Linton, O. and Pakes, A. (2004), “Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems,”Review of Economic Studies, Vol.71 613-654.

[7] Bresnahan, T. F. (1989), “Empirical Studies of Industries with Market Power.”In Handbook of Industrial Organization, vol. 2, ed. Richard Schmalensee and Robert D. Willig, 101157. Amsterdam: North Holland.

[8] Che, H., Sudhir, K. and Seetharaman, P. B. (2007), “Bounded Rationality in Pricing under State-Dependent Demand: Do Firms Look Ahead, and If So, How Far?”Journal of Marketing Research, Vol. 44 (3), 434-449.

[9] Egozcue, J.J. (2009). “Reply to “ On theHarker variation diagrams; . . . ”by J.A.Cortés. Math.Geosci. 41(7), 829834.

[10] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003), “Isometric logratio transformations for compositional data analysis.”Math. Geol. 35(3), 279300.

[11] Filzmoser, P., Hron, K. and Temple, M. (2018). Applied Compositional Data Analysis With Worked Examples in R. Springer Nature, Gewerbestrasse 11, 6330 Cham, Switzerland ISSN 0172-7397 ISSN 2197-568X (electronic). ISBN 978-3-319-96420-1 ISBN 978-3-319-96422-5 (eBook).

[12] Freyberger, J. (2015), “Asymptotic theory for differentiated products demand models with many markets.”Journal of Econometrics, vol. 185, 162-181.

[13] Kamai, T. and Kanazawa, Y. (2016), “Is product with a special feature still rewarding? The case of the Japanese yogurt market.”Cogent Economics & Finance, Vol. 4(1).

[14] Kanazawa, Y. (2018), “Asymptotics for differentiated product demand/supply systems with many markets in the presence of national micro moments.”Research Institute for Mathematical Sciences Kyoto University, Kokyuroku (2091), pp.125-139. October, 2018.

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

[16] Luce, R. D. (1959). Individual Choice Behavior: A Theoretical Analysis. New York: Wiley. ISBN 978-0-486-44136-8.

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

[18] McFadden, D. (1981), “Econometric Models of Probabilistic Choice,”in C. Manski and D. McFadden, eds. Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA.

[19] Myojo, S., Kanazawa, Y. (2012), “On Asymptotic Properties of the Parameters of Differentiated Product Demand and Supply Systems When Demographically-Categorized Purchasing Pattern Data are Available,”International Economic Review, Vol.53(3), 887-938.

[20] Nakayama, K. (2014), “Simulation Studies for Asymptotic Properties of the Parameters of Differentiated Product Demand and Supply Systems When the Number of Markets Increases,”University of Tsukuba, Master Thesis.

[21] Nevo, A. (2001) “Measuring Market Power in the Ready-to-Eat Cereal Industry,”Econometrica, vol. 69 (2), 307-342.

[22] Nelder, J., and R. Mead (1965), “A simplex method for function minimization,”Computer Journal 7, 308-313.

[23] Pakes, A. and D. Pollard (1989), “Simulation and the Asymptotics of Optimization Estimators,”Econometrica, vol. 57(5), 1027-1057.

[24] Pawlowsky-Glahn, V. and Egozcue, J.J. “Geometric approach to statistical analysis on the simplex.”Stoch. Env. Res. Risk A. 15(5), 384398.

[25] Petrin, A. (2002) “Quantifying the Benefits of New Products: The Case of the Minivan,”Journal of Political Economy, Vol.110, 705-729.

[26] Suga, M. (2013), “On Asymptotic Properties of an Estimator for Demand When the Number of Markets Increases,”University of Tsukuba, Master Thesis.

[27] Takeshita, K. (2015), “CAN properties of the random-coefficient model of demand for nondurable consumer goods in the presence of national micro moments: A simulation study,”University of Tsukuba, Master Thesis.

[28] Train, K. E. (2009), Discrete Choice Methods with Simulation, Second Edition, Cambridge University Press. New York, NY 10013-2473, USA. ISBN 978-0-521-76655-5 (hardback) ISBN 978-0-521-74738-7 (paperback).