WORKING PAPERS

N° 17-804

May 2017

"Using compositional and Dirichlet models for market-share regression"

Joanna Morais, Christine Thomas-Agnan, Michel Simioni

Toulouse School of Economics

Using compositional and Dirichlet models for market-share regression

Joanna Morais[ab], **Christine Thomas-Agnan**[a], **Michel Simioni**[c]

[a] Toulouse School of Economics, University of Toulouse 1 Capitole, 21 allée de Brienne, Toulouse, France

[b] BVA, 52 rue Marcel Dassault, Boulogne-Billancourt, France

[c] INRA, UMR 1110 MOISA, 2 Place Pierre Viala, Montpellier, France

joanna.morais@live.fr

May 10, 2017

Abstract

When the aim is to model market-shares as a function of explanatory variables, the marketing literature proposes some regression models which can be qualified as attraction models. They are generally derived from an aggregated version of the multinomial logit model widely used in econometrics for discrete choice modeling. But aggregated multinomial logit models (MNL) and the so-called market-share models or generalized multiplicative competitive interaction models (GMCI) present some limitations: in their simpler version they do not specify brand-specific and cross-effect parameters. Introducing all possible cross effects is not possible in the MNL and would imply a very large number of parameters in the case of the GMCI. In this paper, we consider alternative models which are the Dirichlet covariate model (DIR) and the compositional model (CODA). DIR allows to introduce brand-specific parameters and CODA allows additionally to consider cross-effect parameters. We show that these last two models can be written in a similar fashion, called attraction form, as the MNL and the GMCI models. As market-share models are usually interpreted in terms of elasticities, we also use this notion to interpret the DIR and CODA models. We compare the main properties of the models in order to explain why CODA and DIR models can outperform traditional market-share models. The benefits of highlighting these relationships is on one hand to propose new models to the marketing literature and on the other hand to improve the interpretation of the CODA and DIR models using the elasticities of the econometrics literature. Finally, an application to the automobile market is presented where we model brands market-shares as a function of media investments, controlling for the brands average price and a scrapping incentive dummy variable. We compare the goodness-of-fit of the various models in terms of quality measures adapted to shares.

Keywords: Multinomial logit; Market-shares models; Compositional data analysis; Dirichlet regression.

1 Introduction

Share data are characterized by the following constraints: they are positive and sum up to 1. By definition shares are "compositional data": a composition is a vector of parts of some whole which carries relative information. For a composition of D parts, if $D-1$ parts are known the D^{th} part is simply 1 minus the sum of the $D-1$ other parts: D-compositions lie in a space called the simplex $\mathcal{S}^D$. Because of these constraints, classical regression models cannot be used directly.

A large number of fields are concerned by the analysis of share data. In political economy, Elff [6] studies voting behaviors and analyzes the relationship between the shares of political parties and their policy positions in different groups of voters. In geology, Solana-Acosta and Dutta [21] are interested in the lithologic composition of sandstone according to whether it is quartz, feldspar or rock fragments. For environmental planning purposes, land use models focus on what are the proportions of different types of uses (forest, agriculture, urban, etc...) on a given piece of land, see for example Chakir et al. [4].

When the aim is to model market-shares as a function of explanatory variables (marketing factors like advertising or price for example), the marketing literature proposes some regression models which can be qualified as attraction models (Cooper and Nakanishi [5]). They are generally inspired from an aggregated version of the multinomial logit models, widely used in econometrics for discrete choice modeling. But aggregated multinomial logit models (MNL) and market-share models (GMCI) present some limitations: introducing all possible cross effects is not possible in the MNL and would imply a very large number of parameters in the case of the GMCI.

In this paper, we propose to use the Dirichlet covariate model (DIR) and the compositional model (CODA) in order to model market-shares. These models consider the vector of shares as a "composition" lying in the simplex. DIR allows to estimate brand-specific parameters and CODA allows to estimate additionally cross-effect parameters. We show that these last two models can be written in a similar fashion, called attraction form, as the MNL and the GMCI models. We compare the main properties of the models in order to explain why CODA and DIR models can outperform traditional market-share models.

Finally, an application to the automobile market is presented where we model brands market-shares as a function of media investments in 6 channels (TV, press, radio, outdoor, digital, cinema), controlling for the brands average price and a scrapping incentive dummy variable. We compare the goodness-of-fit of the various models by cross-validation in terms of quality measures adapted to share data. The direct elasticity of market-shares relative to the TV investments are computed for all presented models.

The present paper is organized as follows: the models adapted to model share data are presented in Section 2, and theoretically compared in Section 3. Section 4 presents an application to an automobile market data set, along with an empirical comparison of the models in terms of cross-validated goodness-of-fit measures, and an example of elasticity interpretation. Finally, the last section concludes on the findings and on further directions to be investigated.

2 Models for explaining shares

2.1 Notations

The notations used in this paper are standardized in Table 1 depending on whether the variables are considered in volume or in share, in the left or in the right part of the regression equation, and if they are alternative and/or observation dependent.

Variable	**Volumes** (absolute values)	**Shares** (relative values)	**Coordinates** (ILR)
Dependent	N_{jt}	$\mathbf{S}_t = (S_{1t}, \dots, S_{Dt})' =$ $\mathcal{C}(N_{1t}, \dots, N_{Dt})'$	$\mathbf{S}_t^* = \mathbf{N}_t^*$
Explanatory (observation and component characteristic)	X_{jt}	$\mathbf{Z}_t = (Z_{1t}, \dots, Z_{Dt})' =$ $\mathcal{C}(X_{1t}, \dots, X_{Dt})'$	$\mathbf{Z}_t^* = \mathbf{X}_t^*$
Explanatory (observation characteristic only)	W_t		
General notations			
D	Number of components (3 in the application)		
$j, l, m = 1, \dots, D$	Index of components or coordinates (brands in the application)		
T	Number of observations (123 in the application)		
$t = 1, \dots, T$	Index of observations (time in the application)		
K, K_X, K_W	Number of explanatory variables / of type X / of type W		
$k = 1, \dots, K$	Index of explanatory variables (by default)		
$k = 1, \dots, K_X$	Index of explanatory variables of type X		
$\kappa = 1, \dots, K_W$	Index of explanatory variables of type W		
s_j	Theoretical mean share (expected value of S_j)		
Notations for the application			
C	Number of media channels (6 in the application)		
$c = 1, \dots, C$	Index of media channels		
M_{cjt}	Media investment in channel c at time t for brand j		
P_{jt}	Average price at time t of brand j		
I_t	Scrapping incentive dummy at time t		

Table 1: Notations

$\mathcal{C}()$ denotes the closure operation which transforms volumes into shares:

$$\mathcal{C}(y_1, \dots, y_D)' = \left(\frac{y_1}{\sum_{j=1}^{D} y_j}, \dots, \frac{y_D}{\sum_{j=1}^{D} y_j} \right)'$$

A composition $\mathbf{S}$ is a vector of D shares S_j potentially coming from the closure of D positive numbers N_j and belonging to the simplex $\mathcal{S}^D$:

$$\mathbf{S} = (S_1, \dots, S_D)' = \mathcal{C}(N_1, \dots, N_D)' \in \mathcal{S}^D \quad \text{with} \quad S_j > 0 \quad \text{and} \quad \sum_{j=1}^{D} S_j = 1$$

For example, in the case we use for illustration, the dependent variable is the sales of vehicles observed across time; among the explanatory variables we have media investments, price and scrapping incentive (time dependent only). The sales can be considered in volume (number of sales) or in share (market-shares). Similarly, media investments in volume correspond to the amount of euros spent, whereas in share they correspond to the so-called "shares-of-voice" in marketing.

2.2 Market-share models

Market-share models were developed in the 80's, mainly by Cooper and Nakanishi [5]. To take into account the competition between brands in a market, it is often of interest to model market-shares instead of sales volumes directly. Thus, this type of model is widely used in marketing. The aim is to model market-shares of D brands using their marketing factors (price, advertising) as explanatory variables, with aggregated data (market-level data rather than individual-level data). These models are called generalized multiplicative competitive interaction (GMCI) models. The so-called market-share models are inspired from an aggregated version of the conditional multinomial logit (MNL) models. For individual data, conditional MNL models, widely used in econometrics, model discrete choices of individuals, i.e. the probability that an individual i chooses an alternative j. If explanatory variables are alternative dependent (not individual dependent), one can aggregate the data using a group variable (time for example). In that case, the resulting data may also be modeled using a multinomial distribution.

2.2.1 GMCI attraction model

The concept of "attraction" of a brand is central in this literature, and is comparable to the "utility" concept in discrete choice models for individual data. The specification of the attraction of brand j is a function of the explanatory variables (marketing variables usually, like price and media for example) describing this brand. The market-share of brand j is defined as its relative attraction compared to competitors, i.e. as its attraction divided by the sum of attractions of all the brands of the market.

$$0 < S_{jt} = \frac{\mathcal{A}_{jt}}{\sum_{l=1}^{D} \mathcal{A}_{lt}} < 1$$

where $\mathcal{A}_{jt}$ is the attraction of firm j at observation t such that $\mathcal{A}_{jt} > 0$.

Cooper and Nakanishi [5] defined a general model for market-shares, called the generalized multiplicative competitive interaction model (GMCI). It is defined as follows:

$$\mathcal{A}_{jt} = \exp(a_j + \varepsilon_{jt}) \prod_{k=1}^{K} f_k(X_{kjt})^{b_k} \quad \text{and} \quad S_{jt} = \frac{\mathcal{A}_{jt}}{\sum_{l=1}^{D} \mathcal{A}_{lt}} \tag{1}$$

where $\exp(\varepsilon_{jt})$ is a multiplicative random component, and f_k is a monotonic transformation of X_k such that $f_k(.) > 0$. If all f_k are the identity function (resp. the exponential function), it is called the MCI specification (resp. to the MNL specification):

$$\textbf{MNL spec.:}\ S_{jt} = \frac{\exp(a_j + \sum_{k=1}^{K} b_k X_{kjt} + \varepsilon_{jt})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K} b_k X_{klt} + \varepsilon_{lt})} \tag{2}$$

$$\textbf{MCI spec.:}\ S_{jt} = \frac{\exp(a_j + \sum_{k=1}^{K} b_k \log X_{kjt} + \varepsilon_{jt})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K} b_k \log X_{klt} + \varepsilon_{lt})} \tag{3}$$

The MNL specification of the GMCI is similar to the conditional multinomial logit model (MNL), except that in the MNL model an intercept has to be fixed to zero for identifiability reason:

$$\textbf{MNL model:}\ s_{jt} = \mathbb{E}(S_{jt}|\mathbf{X}_t) = \frac{\exp(a_j + \sum_{k=1}^{K} b_k X_{kjt})}{\sum_{l=1}^{D} \exp(a_l + \sum_{k=1}^{K} b_k X_{klt})} \quad \text{with } a_D = 0 \tag{4}$$

Note however that the attraction formulation of the MNL model differs from that of the GMCI models: the GMCI attraction contains the random component ε_{jt} whereas the MNL does not since the attraction form in that case corresponds to the expected share. We will further develop this aspect in section 3.2.

2.2.2 Estimation by OLS

Contrary to the MNL model which is estimated by maximum likelihood based on the multinomial distribution, Nakanishi and Cooper [17] proposed an estimation method relying on a log linearization that they call "log-centering transformation" which is actually the log ratio between a share S_{jt} and the geometric mean of all shares at observation t, $\widetilde{\mathbf{S}}_t$, also called CLR (centered log-ratio) transformation in the CODA (Compositional Data Analysis) literature. The log-centered formulations are given by:

$$\textbf{MNL spec.:}\ \log\left(\frac{S_{jt}}{\widetilde{\mathbf{S}}_t}\right) = a_1 + \sum_{l=2}^{D}(a_j - a_1)d_l + \sum_{k=1}^{K} b_k(X_{kjt} - \overline{\mathbf{X}}_{kt}) + (\varepsilon_{jt} - \overline{\varepsilon}_t)$$

$$\textbf{MCI spec.:}\ \log\left(\frac{S_{jt}}{\widetilde{\mathbf{S}}_t}\right) = a_1 + \sum_{l=2}^{D}(a_j - a_1)d_l + \sum_{k=1}^{K} b_k \log\left(\frac{X_{kjt}}{\widetilde{\mathbf{X}}_{kt}}\right) + (\varepsilon_{jt} - \overline{\varepsilon}_t)$$

where $d_l = 1$ if $l = j$, 0 otherwise (brand dummy). $\overline{\mathbf{S}}_t$ and $\widetilde{\mathbf{S}}_t$ are the arithmetic and the geometric means of S_{jt}.

This OLS estimation would be correct if error terms $\varepsilon^*_{jt} = (\varepsilon_{jt} - \overline{\varepsilon}_t)$ had a multivariate distribution with diagonal variance covariance matrix, but indeed the ε^*_{jt} can only follow a degenerate multivariate normal distribution. It is suggested to use a generalized least squares (GLS) estimation instead of an OLS estimation due to the potential heteroscedasticity and/or correlation of error terms (if observations are time periods for example). But as stated in Cooper and Nakanishi [5], we found that the GLS procedure, which is quite heavy in terms of implementation for this kind of models, does not give empirically better results than the OLS procedure.

Implementation in R: the function lm() allows to fit the log-centered model by ordinary least squares (GMCI). The package "mclogit" developped by Martin Elff [7] allows to fit conditional logit models with count data, using the Fisher-scoring/IWLS algorithm[1] (MNL).

2.3 Dirichlet covariate models

The Dirichlet distribution is the distribution of a composition obtained as the closure of a vector of D independent gamma-distributed variables with the same scale parameter. Thus, it is another distribution adapted for variables lying in the simplex. Let $\mathbf{S} = (S_1, \ldots, S_D) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_D)$

[1] For details on IWLS algorithm, see for example Green [8].

where $S_j > 0$ and $\sum_{j=1}^{D} S_j = 1$, $\alpha_j > 0$ and $\sum_{j=1}^{D} \alpha_j = \alpha_0$. α_0 is called the precision parameter. Then, $\mathbb{E}(S_j) = \frac{\alpha_j}{\alpha_0}$. Two parametrizations exist for the Dirichlet regression model: the *"common parametrization"* and the *"alternative parametrization"*[2]. We focus here on the common specification.

2.3.1 Dirichlet model

Campbell and Mosimann [3] developed Dirichlet covariate models to explain a compositional dependent variable, supposed to be Dirichlet distributed, by classical (non-Dirichlet) covariates. As explained in Hijazi and Jernigan [10], "a different Dirichlet distribution is modeled for every value of the explanatory variables, resulting in a conditional Dirichlet distribution". The conditional distributions $\mathbf{S}_t|\mathbf{X}_t$ are mutually independent: $\mathbf{S}_t|\mathbf{X}_t \sim \mathcal{D}(\alpha_1(\mathbf{X}_t), \ldots, \alpha_D(\mathbf{X}_t))$ with unknown parameters. Under the common parametrization, the parameters of the Dirichlet distribution, the α_j's, are allowed to depend on the explanatory variables X_k in a GLM fashion with a log link.

$$\log(\alpha_j(\mathbf{X}_t)) = a_j + \sum_{k=1}^{K} b_{kj} X_{kjt} \quad \text{and} \quad \mathbb{E}(S_j) = \frac{\alpha_j(\mathbf{X}_t)}{\sum_{m=1}^{D} \alpha_m(\mathbf{X}_t)} \tag{5}$$

The components may have different explanatory variables (a different number of explanatory variables and/or explanatory variables which take different values for the different components), but for the sake of simplicity $\mathbf{X}$ denotes the vector of explanatory variables for all components.

2.3.2 Estimation by maximum likelihood

The log-likelihood to maximize is:

$$\log L(\mathbf{S}|\alpha(\mathbf{X})) = \sum_{t=1}^{T} \left[\log \Gamma \left(\sum_{j=1}^{D} \alpha_j(\mathbf{X}_t) \right) - \sum_{j=1}^{D} \log \Gamma(\alpha_j(\mathbf{X}_t)) + \sum_{j=1}^{D} (\alpha_j(\mathbf{X}_t) - 1) \log S_{jt} \right]$$

Implementation in R: the package "DirichReg" created by Maier [12] allows to fit Dirichlet model for the common or alternative parametrization, by maximum likelihood.

2.4 Compositional models

Compositional data analysis was developed in the 80's by John Aitchison [1]. The first applications were for geological data, with the objective to analyze the composition of a rock sample in terms of the relative presence of different chemical elements. More generally, CODA aims to analyze relative information between the components (parts) of a composition where the total of the components is not relevant or is not of interest.

2.4.1 The log-ratio transformation approach

The CODA approach is based on log-ratio transformations of compositions in order to obtain coordinates which can be represented in a $\mathbb{R}^{D-1}$ Euclidean space. Then, classical methods suited

[2] The alternative parametrization uses the parameters $\mu_j = \mathbb{E}(S_j)$ to account for the expected values of the shares, and $\phi = \alpha_0$ to account for the precision. See Hijazi and Jernigan [10].

for modeling data in the Euclidean space can be used on coordinates. Several transformations are developed: notably the CLR (centered log-ratio) and the ILR (isometric log-ratio) transformations.

- The CLR transformation leads to D coordinates instead of $D-1$ for others. It is defined as follows: $clr(\mathbf{S}) = \left(\log \frac{S_1}{\widetilde{\mathbf{S}}}, \ldots, \log \frac{S_D}{\widetilde{\mathbf{S}}}\right)'$ where $\widetilde{\mathbf{S}}$ is the geometric mean of the D components. Its inverse transformation is: $\mathbf{S} = clr^{-1}(clr(\mathbf{S})) = \mathcal{C}(\exp(clr(\mathbf{S})_1), \ldots, \exp(clr(\mathbf{S})_D))'$.

- The ILR transformation consists in a projection of components in an orthonormal basis of $\mathcal{S}^D$ in order to obtain $D-1$ orthonormal coordinates. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_{D-1}\}$ be an arbitrary orthonormal basis in $\mathbb{R}^{D-1}$, then $\mathbf{e}_l = clr^{-1}(\mathbf{v}_l)$, $l = 1, \ldots D-1$, represent an orthonormal basis in the simplex $\mathcal{S}^D$ equipped with its "natural geometry" (see Pawlowsky-Glahn et al. [18]). Considering the $D \times (D-1)$ matrix $\mathbf{V}$ with columns $\mathbf{v}_l = clr(\mathbf{e}_l)$, $l = 1, \ldots D-1$, ILR coordinates are defined as $ilr(\mathbf{S}) = \mathbf{S}^* = \mathbf{V}'clr(\mathbf{S}) = \mathbf{V}'\log(\mathbf{S})$. Its inverse transformation is given by: $\mathbf{S} = ilr^{-1}(\mathbf{S}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{S}^*))'$.

Example A particular ILR transformation that could be used is the following:

$$S_l^* = \sqrt{\frac{D-l}{D-l+1}} \log \frac{S_l}{(\prod_{l'=l+1}^{D} S_{l'})^{\frac{1}{D-l}}}, \quad l = 1, \ldots, D-1$$

S_1^* contains all the relative information of part S_1 to the parts $S_2, \ldots, S_D$.
If $D = 3$ for example, it leads to $S_1^* = \sqrt{\frac{2}{3}} \log \frac{S_1}{\sqrt{S_2 S_3}} = \sqrt{\frac{2}{3}} \log S_1 - \frac{1}{\sqrt{6}}(\log S_2 + \log S_3)$ and $S_2^* = \sqrt{\frac{1}{2}} \log \frac{S_2}{S_3} = \frac{1}{\sqrt{2}}(\log S_2 - \log S_3)$.
Thus, $\mathbf{V} = \begin{bmatrix} \sqrt{2/3} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}$.

For regression as well as for other statistical analysis, the results are the same after inverse transformation regardless of the chosen transformation. However, as CLR introduces collinearity between coordinates, ILR is preferred for compositional regression models.

2.4.2 CODA regression models

Compositional regression models are of different types depending on whether the response variable and/or the explanatory variables are compositional. We focus here on the case where the dependent as well as the explanatory variables are compositional and of same dimension D (for example, market-shares of D brands are explained by the corresponding media investments)[3].
CODA models can be expressed either in terms of the initial compositional observations in the simplex (equation (6)) or alternatively in terms of the corresponding transformed coordinates in the Euclidean space (equation (7)), as follows:

[3] Note that the dependent composition and the explanatory compositions could be of different dimensions.

- Linear CODA model in the simplex (in terms of compositions):

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{k=1}^{K} \mathbf{B_k} \boxdot \mathbf{Z_k}_t \oplus \boldsymbol{\varepsilon}_t \tag{6}$$

with $\mathbf{S}, \mathbf{a}, \mathbf{Z_k}, \boldsymbol{\varepsilon} \in \mathcal{S}^D$ and $\mathbf{B_k} \in \mathbb{R}_{D\times D}$ such that row and column sums are equal to zero[4], and the following operations are used in the simplex:

- $\oplus$ is the *perturbation operation*, corresponding to the addition operation in the simplex: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1y_1, \ldots, x_Dy_D)'$ with $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, and $\bigoplus_{k=1}^{K}$ corresponds to $\sum_{k=1}^{K}$.
- $\odot$ is the *power transformation*, corresponding to the multiplication operation in the simplex: $\mathbf{x} \odot \lambda = \mathcal{C}(x_1^\lambda, \ldots, x_D^\lambda)'$ with $\lambda \in \mathbb{R}, \mathbf{x} \in \mathcal{S}^D$
- $\boxdot$ is the *compositional matrix product*, corresponding to the matrix product in the simplex: $\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\big(\prod_{j=1}^{D} x_j^{b_{1j}}, \ldots, \prod_{j=1}^{D} x_j^{b_{Dj}}\big)'$ with $\mathbf{B} \in \mathbb{R}_{D\times D}, \mathbf{x} \in \mathcal{S}^D$

- Linear CODA model in the Euclidean space (in terms of ILR coordinates):

$$S_{jt}^* = a_j^* + \sum_{k=1}^{K} \sum_{m=1}^{D-1} b_{kjm}^* X_{kmt}^* + \varepsilon_{jt}^* \quad \forall\, j \in 1, \ldots, D-1 \tag{7}$$

where j is the index of $\mathbf{S}$'s ILR coordinates, m is the index of $\mathbf{X}$'s ILR coordinates and $\varepsilon_j^* \sim \mathcal{N}(0, \sigma^2)$. Equation 7 corresponds to a system of $D-1$ linear models, one for each ILR coordinate of $\mathbf{S}$. Note here that compositional explanatory variables coordinates can be equivalently calculated using $\mathbf{X}$ (volumes) or $\mathbf{Z}$ (shares).

The second presentation has the advantage to look like a classical linear model but its connection with the original data is obscured by the transformation. On the other hand, the first presentation in terms of the original share data is obscured by the simplex operations involved in the model equation.

2.4.3 Estimation by OLS

After log-ratio transformation (equation (7)), the estimation is usually done with the OLS method, separately on the $D-1$ linear models expressed in coordinates[5].

Then, the estimated model can be back transformed into the simplex using the inverse transformation which transforms $\boldsymbol{\alpha}$ into $\mathbf{a}$, $\boldsymbol{\beta}$ into $\mathbf{b}$, $ilr(\mathbf{S})$ into $\mathbf{S}$ and $ilr(\mathbf{Z})$ into $\mathbf{Z}$:

$$\begin{aligned}
\mathbf{a} &= ilr^{-1}(a_1^*, \ldots, a_{D-1}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{a}^*)) \\
\mathbf{B}_{D,D} &= \mathbf{V}\mathbf{B}_{D-1,D-1}^*\mathbf{V}' \\
\mathbf{S} &= ilr^{-1}(S_1^*, \ldots, S_{D-1}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{S}^*))
\end{aligned}$$

with $\mathbf{B}^* = \begin{bmatrix} b_{1,1}^* & \ldots & b_{1,D-1}^* \\ \ldots & b_{j,l}^* & \ldots \\ b_{D-1,1}^* & \ldots & b_{D-1,D-1}^* \end{bmatrix}$, and $\mathbf{B} = \begin{bmatrix} b_{1,1} & \ldots & b_{1,D} \\ \ldots & b_{j,l} & \ldots \\ b_{D,1} & \ldots & b_{D,D} \end{bmatrix}$ where $b_{j,l}^*$ is the parameter

[4] Under these conditions, $\mathbf{B} \boxdot \mathbf{Z}$ is an endomorphism of the simplex $\mathcal{S}^D$ (See Kynclova et al. [11]). Thus model (6) is a linear model in the simplex.

[5] The orthonormality of coordinates allows us to estimate the $D-1$ models separately.

corresponding to the impact of Z_l^* on S_j^*, and $b_{j,l}$ is the parameter corresponding to the impact of Z_l on S_j.

Implementation in R: the packages “compositions” [23] and “robCompositions” [22] allow to transform compositional data, to fit the compositional model by OLS on the coordinates and to back transform the results into compositions. Implementation of CODA using R is presented in the book of Van den Boogaart and Tolosana-Delgado [24].

3 Theoretical comparison of share models

In this section, we highlight the similarities and differences of the presented models from a theoretical perspective. Because these models are deeply linked with the type of applications they have been proposed for, the following comparison refers not only to statistical properties, but also to econometric and marketing properties. Table 2 summarizes the distributional assumptions, the estimation methods, the properties and the complexity of each model[6]. These items are discussed in detail below. Finally we highlight the fact that GMCI can be expressed in a CODA way.

3.1 Distributional assumptions

In the MNL model the dependent variable is a vector of positive numbers N_j which follow a multinomial distribution. In the other three models the dependent variable is directly the vector of shares S_j which are Dirichlet distributed in the case of DIR and Gaussian in the simplex distributed for GMCI and CODA (the coordinates are Gaussian in the transformed space). Note that the MNL model differs from the MNL specification of the GMCI model by its underlying distributional assumptions.

MNL and Dirichlet models belong to the family of GLM (Generalized linear models): see Peyhardi et al. [19] for MNL and Maier [12] for Dirichlet. GMCI and CODA models belong to the family of transformation models (TRM hereafter) in which a classical linear model is postulated in the transformed space.

3.2 Expected shares and attraction formulation

Expected value of shares Let us notice that the model formulation of the two GLM models - MNL (4) and DIR (5) - involves the expected shares $\mathbb{E}(S_{jt}|\mathbf{X}_t)$, while the two transformation models formulation - GMCI (1) and CODA (6) - involves the random shares S_{jt} and a random error term. The usual expected value cannot be analitically computed for the GMCI and the CODA models. For this reason, we turn attention to the “expected value in the simplex”, defined as follows (see Theorem 6.10 p.109 in Pawlowsky-Glahn et al. [18]):

$$\mathbb{E}^{\oplus}\mathbf{S} = \mathcal{C}(\exp(\mathbb{E}\log\mathbf{S})) = clr^{-1}(\mathbb{E}clr(\mathbf{S})) = ilr^{-1}(\mathbb{E}ilr(\mathbf{S})) = ilr^{-1}(\mathbb{E}\mathbf{S}^*)$$

This means that the expected value in the simplex of the composition of shares, $\mathbb{E}^{\oplus}\mathbf{S}$, coincides with the ILR back transformation of expected values of the random coordinates, $\mathbb{E}\mathbf{S}^*$.

[6] Here the GMCI model is presented with the MNL specification. Note that if X is replaced by $\log X$, it corresponds to the MCI specification.

Remark: If the explanatory variables only consist of intercepts, the fitted shares are not the same across the four models. In the case of the CODA and the GMCI models, they correspond to the center of the compositional data, that is the closed vector of geometric means of each component, while in the case of the MNL and the DIR models, fitted shares are the arithmetic means of components (weighted in the case of MNL). The geometric mean, which is coherent with the simplex geometry, is more adapted than the arithmetic mean to summarize shares data. This is an argument in favor of CODA and GMCI models.

Attraction formulation of share models As seen before, the attraction formulations in MNL and GMCI are different (in GMCI it includes a random error term). In order to unify the presentation, we introduce a deterministic attraction A_{jt} and a random attraction u_{jt} such that $\mathcal{A}_{jt} = A_{jt}u_{jt}$. According to equations (4), (2), (3), the deterministic attraction formulations of MNL and the two GMCI models (G_{MNL} for the MNL specification and G_{MCI} for the MCI specification) are:

$$\begin{aligned}
A_{jt}^{MNL} &= \exp(a_j + \sum_{k=1}^{K} b_k X_{kjt}) \quad \text{with } a_D = 0 && \Leftrightarrow \quad \mathbb{E}S_{jt} = \frac{A_{jt}^{MNL}}{\sum_{m=1}^{D} A_{mt}^{MNL}} \\
A_{jt}^{G_{MNL}} &= \exp(a_j + \sum_{k=1}^{K} b_k X_{kjt}) && \Leftrightarrow \quad \mathbb{E}^{\oplus}S_{jt} = \frac{A_{jt}^{G_{MNL}}}{\sum_{m=1}^{D} A_{mt}^{G_{MNL}}} \\
A_{jt}^{G_{MCI}} &= \exp(a_j + \sum_{k=1}^{K} b_k \log X_{kjt}) && \Leftrightarrow \quad \mathbb{E}^{\oplus}S_{jt} = \frac{A_{jt}^{G_{MCI}}}{\sum_{m=1}^{D} A_{mt}^{G_{MCI}}}
\end{aligned}$$

This emphasizes the fact that the type of expected shares involved in the attraction formulation are different between the MNL model and the MNL specification of the GMCI.

The Dirichlet model can also be expressed with an attraction formulation:

$$A_{jt}^{DIR} = \exp(a_j + \sum_{k=1}^{K} b_{kj} X_{kjt}) = \alpha_{jt} \qquad \Leftrightarrow \quad \mathbb{E}S_{jt} = \frac{A_{jt}^{DIR}}{\sum_{m=1}^{D} A_{mt}^{DIR}}$$

This emphasizes the fact that the parameters of the DIR model are alternative-specific (they depend on j), contrary to the GMCI and MNL models.

We now derive the attraction form of the compositional model, using equation (6). We first express the market-share of brand j in the CODA model[7] as:

$$\mathbf{S}_t = \mathbf{a}_t \bigoplus_{k=1}^{K} \mathbf{B_k} \boxdot \mathbf{Z_k}_t \oplus \boldsymbol{\varepsilon}_t = \mathcal{C}\left(a_1 \prod_{k=1}^{K}\prod_{l=1}^{D} X_{klt}^{b_{k1l}} \varepsilon_{1t}, \ldots, a_D \prod_{k=1}^{K}\prod_{l=1}^{D} X_{klt}^{b_{kDl}} \varepsilon_{Dt}\right)$$

[7] The market-share S_{jt} is here expressed as a function of X_{klt} directly and not as a function of Z_{klt} because S_{jt} is obtained by a closure operation (dividing by the denominator), thus it can be shown that the explanatory variables can be used in volume as they are closed at the end.

Thus, if we let

$$A_{jt}^{CODA} = a_j \prod_{k=1}^{K} \prod_{l=1}^{D} X_{klt}^{b_{kjl}} = \exp\left(\log(a_j) + \sum_{k=1}^{K} \sum_{l=1}^{D} b_{kjl} \log(X_{klt})\right) \tag{8}$$

then we have:

$$\mathbb{E}^{\oplus} S_{jt} = \frac{A_{jt}^{CODA}}{\sum_{m=1}^{D} A_{mt}^{CODA}} \tag{9}$$

Note that taking X as explanatory variable in (6) actually corresponds to using $\log(X)$ in the attraction formulation of the CODA model under the exponential form (8). This is similar to the MCI specification of the GMCI model, and different from the MNL model, the MNL specification of the GMCI and the DIR model.

3.3 Properties

We now discuss whether the properties that have been introduced and established in the literature for a given model are valid for the other ones.

IIA and subcompositional coherence In the econometric literature, an important question often discussed is whether or not a choice model satisfies the IIA (Independence from Irrelevant Alternatives) property. IIA means that the ratio of shares of an alternative j with respect to an alternative l only depends on the characteristics of j and l and is not affected by the presence or absence of irrelevant alternatives. This property allows to simplify the models but it is not always realistic (see the red bus - blue bus example of McFadden [14]). Without cross-effects, MNL, GMCI and Dirichlet models satisfy IIA but CODA models do not.

In the CODA literature, the subcompositional coherence property (see Pawlowsky-Glahn [18]) means that the results of an analysis made on a subcomposition (i.e. remove some alternatives) should not contradict the results of the analysis made on the whole composition. This is coming from the fact that compositional data analysis is based on the use of log-ratios. However, if we look at equation (8), we can see that the market-share of brand j is determined by the explanatory variables of all the brands. Thus, subcompositional coherence does not imply IIA, but the reciprocal is true. In the econometrics literature, it is considered that IIA can be a severe limitation, which is a positive point for CODA models.

Invariance The **scale invariance** is the fact that multiplying the count data by a constant does not affect the estimation results. It is a desirable property satisfied by the four models.
The **permutation invariance** is a desirable property corresponding to invariance through a permutation of the components of a composition. It is clearly satisfied by all the described models.
The **perturbation invariance** corresponds to coherence when performing a change of units possibly different for each component of a composition. For example, we can model brand market-shares in terms of sales volumes or in terms of sales values (that is sales volumes perturbed by the vector of prices). The estimated market-shares and parameters from the "volume" model should be equal to those of the "value" model after perturbation by the vector of prices. This property is satisfied by CODA and GMCI models. We can show empirically that it is not satisfied by MNL and DIR models.

3.4 Model complexity

In MNL, GMCI and DIR models, the deterministic attraction A_{jt} is a function of the explanatory variables characterizing alternative j only, leading to the absence of cross-effects. But in the Dirichlet model, parameters are alternative-specific, which increases the complexity of the model. In the CODA model, the attraction may depend on all alternative characteristics, inducing alternative-specific and cross-effect parameters. This is why CODA is the most complex model with the higher number of parameters.

It is not possible to take into account all cross effects in the MNL model (see So and Kuhfeld [20]). Cross effects can be incorporated in the GMCI model (see Cooper and Nakanishi [5]) and in the Dirichlet models but the number of parameters dramatically increases. CODA is relatively parsimonious in the sense that it allows to incorporate all cross effects with a number of parameters relatively lower than the other models ($(D-1) \times (D-1)$ versus $D \times D$ for others), thanks to the constraints on the B matrix of parameters.

It is interesting to see that using the same dependent and explanatory variables, the complexity is totally different from one model to another. For example (as in our application, see Section 4), if the number of components (shares) of the dependent variable is $D = 3$, explained by $K_X = 7$ compositions of size $D = 3$ and $K_W = 1$ time-dependent variable, the number of estimated parameters are the following: 11 for MNL, 13 for GMCI, 27 for DIR and 32 for CODA. With 32 parameters, the CODA model reflects all the cross-effects between shares whereas the DIR and the GMCI models with cross-effects would require 69 parameters ($D(1 + D \times K_X + K_W)$). Note also that the number of parameters increases dramatically with the number of components (brands), especially in the CODA model. For example if D becomes equal to 5 (with K_X and K_W fixed), the number of parameters become 15, 17, 45, and 120, which can be a serious limitation for the CODA model.

3.5 Compositional form of the GMCI model

Even though the GMCI estimation procedure uses a log-ratio transformation as the CODA model, the two models are different and we are now going to express the GMCI model in a compositional form, which will reveal this difference.

Wang et al. [25] propose a CODA regression model for the case when both dependent and explanatory variables are compositional which is simpler than the one presented in paragraph 2.4: instead of having a matrix of parameters for each compositional explanatory variable, the model has a unique real parameter for all components of the explanatory composition. This model does not include cross effects between components contrary to the usual CODA model.

Actually Wang et al.'s model is exactly similar to the MCI model proposed by Cooper and Nakanishi in 1988 [5], except that Wang et al. use ILR coordinates while CLR coordinates are used in the MCI model.

From this correspondence we derive a compositional form for the GMCI model:

$$\mathbf{S}_t = \mathbf{a} \bigoplus_{k=1}^{K} b_k \odot \mathbf{Z}_{\mathbf{k}t} \oplus \boldsymbol{\varepsilon}_t \tag{10}$$

$$\Leftrightarrow S_{jt} = \frac{a_j \cdot \prod_{k=1}^{K} X_{kjt}^{b_k} \cdot \varepsilon_{jt}}{\sum_{l=1}^{D} a_l \cdot \prod_{k=1}^{K} X_{klt}^{b_k} \cdot \varepsilon_{lt}} = \frac{\exp(\log a_j + \sum_{k=1}^{K} b_k \log X_{kjt} + \log \varepsilon_{jt})}{\sum_{l=1}^{D} \exp(\log a_l + \sum_{k=1}^{K} b_k \log X_{klt} + \log \varepsilon_{lt})}$$

Equation (10) highlights the similarities and differences between GMCI and CODA models: in place of the B_k matrix in Equation (6) of the CODA model, we now have a single b_k parameter in the GMCI model. We prove in Morais et al. [16] that the GMCI model is a particular case of the CODA model.

Name	Expected shares	Distribution	Estimation	Properties	Nb param.
MNL GLM type	$\mathbb{E}S_{jt} = \frac{\exp(a_j+\sum_{k=1}^{K_X} b_k X_{kjt}+\sum_{\kappa=1}^{K_W} b_{\kappa j} W_{\kappa t})}{\sum_{l=1}^{D}\exp(a_l+\sum_{k=1}^{K_X} b_k X_{klt}+\sum_{\kappa=1}^{K_W} b_{\kappa l} W_{\kappa t})}$ with $a_1 = 0$ for identifiability reasons.	$(N_{1t}, \dots N_{Dt}) \sim \mathcal{MN}(N_t, s_{1t}, \dots, s_{Dt})$ Indep. distributed over t	Maximum Likelihood	Permutation invariance, Scale invariance, Random utility model, IIA	$(D-1)(1+K_W)+K_X$
GMCI TRM type	Share: $\mathbb{E}^{\oplus}S_{jt} = \frac{\exp(a_j+\sum_{k=1}^{K_X} b_k X_{kjt}+\sum_{\kappa=1}^{K_W} b_{\kappa j} W_{\kappa t})}{\sum_{l=1}^{D}\exp(a_l+\sum_{k=1}^{K_X} b_k X_{klt}+\sum_{\kappa=1}^{K_W} b_{\kappa l} W_{\kappa t})}$ Equivalently in terms of CLR coordinate: $\mathbb{E}\log\left(\frac{S_{jt}}{\widetilde{S}_t}\right) = a_1 + \sum_{j'=2}^{D} a'_{j'} d_{j'} + \sum_{k=1}^{K_X} b_k (X_{kjt} - \overline{X}_{kt}) + \sum_{\kappa=1}^{K_W}\left(b_{\kappa 1} W_{\kappa t} + \sum_{j'=2}^{D} b'_{\kappa j'} W_{\kappa t} d'_j\right)$	$clr(\mathbf{S}_t) \sim \mathcal{N}_D(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ with $\mathcal{N}_D$ the multivariate normal distribution (degenerate here). Indep. distributed over t	OLS on coordinates	Permutation invariance, Scale invariance, Perturbation invariance, IIA	$D(1+K_W)$ $+K_X$
DIR GLM type	With the common parametrization: $\mathbb{E}S_{jt} = \frac{\exp(a_j+\sum_{k=1}^{K_X} b_{kj} X_{kjt}+\sum_{\kappa=1}^{K_W} b_{\kappa j} W_{\kappa t})}{\sum_{l=1}^{D}\exp(a_l+\sum_{k=1}^{K_X} b_{kj} X_{klt}+\sum_{\kappa=1}^{K_W} b_{\kappa l} W_{\kappa t})}$ $\log\alpha_{jt} = a_j + \sum_{k=1}^{K_X} b_{kj} X_{kjt} + \sum_{\kappa=1}^{K_W} b_{\kappa j} W_{\kappa t}$	$(S_{1t}, \dots S_{Dt}) \sim \mathcal{D}(\alpha_{1t}, \dots, \alpha_{Dt})$ Indep. distributed over t	Maximum likelihood	Permutation invariance, Scale invariance, IIA	$(1+K_X+K_W)\times D$
CODA TRM type	Composition in the simplex: $\mathbb{E}^{\oplus}\mathbf{S}_t = \mathbf{a}\bigoplus_{k=1}^{K_X}\mathbf{B_k}\boxdot\mathbf{X_{k}}_t\bigoplus_{\kappa=1}^{K_W} W_{\kappa t}\odot\mathbf{b}_\kappa$ Equivalently in terms of share in the simplex: $\mathbb{E}^{\oplus}S_{jt} = \frac{a_j\cdot\prod_{l=1}^{D}\prod_{k=1}^{K_X} X_{klt}^{b_{kjl}}\cdot\prod_{\kappa=1}^{K_W} b_{\kappa j}^{W_{\kappa t}}}{\sum_{m=1}^{D} a_m\cdot\prod_{l=1}^{D}\prod_{k=1}^{K_X} X_{klt}^{b_{kml}}\cdot\prod_{\kappa=1}^{K_W} b_{\kappa m}^{W_{\kappa t}}}$ Equivalently in terms of ILR coordinates: $\mathbb{E}S_{jt}^{*} = a_j + \sum_{k=1}^{K_X}\sum_{m=1}^{D-1} b_{kjm}^{*} X_{kmt}^{*} + \sum_{\kappa=1}^{K_W} b_{\kappa,j}^{*} W_{\kappa t}$	$\mathbf{S}_t \sim \mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ with $\mathcal{N}_{\mathcal{S}}$ the normal distribution on the simplex, μ a mean vector, $\boldsymbol{\Sigma}$ a variance matrix. $\mathbf{S}^{*} = ilr(\mathbf{S}_t) \sim \mathcal{N}_{D-1}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$	OLS on coordinates	Sub-compositional coherence, Permutation invariance, Scale invariance, Perturbation invariance	$((D-1)\times K_X+K_W+1)\times(D-1)$

Table 2: Benchmark of models for explaining shares
(GLM: Generalized linear model ; TRM: Transformation model)

4 Empirical comparison of share models

In this section, we use the MNL, GMCI, DIR and CODA models for a concrete case study in order to demonstrate that Dirichlet and compositional models can perform better that usual market-share models. After presenting the application and the data of our illustrative example, a cross-validation process is proposed based on quality measures adapted for shares models on the four types of models. Finally, we compare the interpretation of the parameters of the four models in terms of elasticities.

4.1 Application and data

The main objective of this application is to understand the impact of media investments on brand market-shares controlling for other factors like price and scrapping incentive. In each model specification, the interest is on the marginal impact of each media channel on relative sales, that is on the elasticities of market-shares to media investments by channel.

We focus here on the B segment[8] of the French automobile market, which represents half of the sales in France in terms of volume. More precisely, following the subcompositional coherence property of CODA, we focus on 3 brands of this segment: Renault, Nissan and Dacia ($D = 3$).

The studied period, running from June 2005 to August 2015, is characterized by the birth of Dacia on the French automobile market, a low-cost brand belonging to Renault, at the beginning of 2005. It is also characterized by the economic crisis which has hurt the French automobile market a lot from 2008 to 2012. The French government tried to help this market setting up a scrapping incentive[9] which has "artificially" boosted the sales during 2009 and 2010. Note that Dacia increased a lot its market-share during the crisis thanks to its low price. These facts have to be kept in mind in order to understand the evolution of market-shares, and it justifies the use of a scrapping incentive dummy as control variable.

The four models are applied to an automobile market data set coming from Renault containing for each brand of the B segment the sales volume in units N_{jt}, the catalog price in euros P_{jt}, the media investments by canal in euros M_{cjt} (TV, press, radio, outdoor, digital, cinema), and the periods of scrapping incentive I_t (dummy variable), monthly from June 2005 to August 2015 ($T = 123$ periods of observation).

The ternary diagram allows to represent compositions of 3 components in the simplex (see Van den Boogaart and Tolosana-Delgado [24]). Figure 1 represents for example the annual market-shares of Dacia, Nissan and Renault from 2005 to 2014. We can see easily that Dacia increases its market-share easily at the expense of Renault from 2005 to 2010.

According to the marketing literature, it is preferable to use the logarithm of price instead of the raw price[10]. Indeed, for our four models, using the log of price instead of the price gives best in-sample fits. The media investments have to be considered with a lag with respect to sales.

[8] Segments of the automobile market are determined according to the size of the chassis. Segment B corresponds to small mainstream vehicles like the Renault Clio which is the most famous of this segment in France.

[9] A scrapping incentive is an incentive given by a government to promote the replacement of old vehicles with modern vehicles.

[10] The reason of that is linked to the shape of the elasticity of market-shares to the price. Moreover, to keep the market-shares equal, the logged variables have to increase in the same proportion while the non-logged variables have to increase by the same amount.

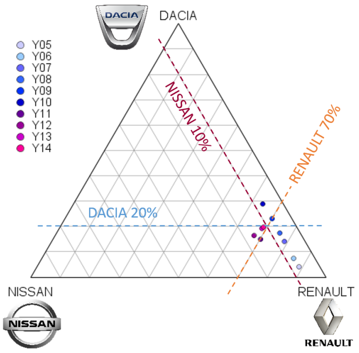

Figure 1: Ternary diagram of annual market-shares of Dacia, Nissan and Renault

Statistically in this application, a lag of four months gave the best results on the four models[11]. To avoid the problem of zeros due to the use of logarithm, when media investments at time t are equal to zero, we replace them by one euro, which is a very small amount compared to the non-zero investments.

4.2 A cross-validation comparison

A cross-validation process is used to compute out-of-sample goodness-of-fit measures on the four considered models, in order to avoid an over-fitting effect and in order to compare the considered models which do not have the same number of parameters.

1. Randomly draw a sub-sample of 100 observations among 123[12], resulting in 81% (100) in-sample observations and 19% (23) out-of-sample observations
2. Fit the 4 models to the sub-sample, store the fitted parameters
3. Apply the 4 models to the out-of-sample observations, store the fitted values of the shares
4. Compute the quality measures using the out-of-sample predicted share values
5. Iterate 100 times steps 1 to 4
6. Compute the average quality measures using the out-of-sample predicted share values over the 100 iterations

4.3 Quality measures

The out-of-sample accuracy of the four models is compared according to a list of different indicators adapted to shares that we found in the literature. Two categories of measures are detailed: the R^2-type measures which are based on the notion of explained variability, and the distance-type measures which evaluate how far are the fitted values from the true values. Table 3 presents the

[11] In forthcoming work, we consider using an adstock function, which is a cumulative function of actual and past investments.

[12] Here we want to have an efficient model all along the studied period, the aim is not to have a good predictive model for the future. Moreover the presented models are not taking into account the potential auto-correlation of error terms. That is the reason why the cross-validation can be made on randomly drawn dates and not on a split of the studied period according to time.

out-of-sample average quality measures for our four models (for each measure, the best model is in bold), for the following measures:

- R^2_T: R-squared based on the total variability, widely used in the compositional literature.
- R^2_A: R-squared based on Aitchison distance, used in Hijazi [9] and Monti et al. [15]. Warning: it can be smaller than 0 and larger than 1.
- KL_C: the compositional Kullback-Leibler divergence (see Martin-Fernandez et al. [13]).

	MNL		GMCI		DIR		CODA	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
R^2_T	0.425	0.164	0.462	0.179	0.622	0.224	**0.647**	0.227
R^2_A	0.196	0.270	0.155	0.325	**0.373**	0.235	0.084	0.433
KL_C	0.139	0.034	0.137	0.032	**0.117**	0.071	0.134	0.034

Table 3: Out-of-sample quality measures

The out-of-sample average quality measures suggest that DIR is the most adapted model to fit our data (27 parameters). However, according to the R^2 based on total variability (R^2_T), CODA (32 parameters) is better than the Dirichlet model. The GMCI model and the MNL model without cross-effects are almost systematically the worst models, certainly due to their simplicity and low number of parameters.

4.4 Interpretation of parameters

MNL and GMCI models are usually interpreted in terms of direct and cross elasticities (see Cooper and Nakanishi [5]). In Morais et al. [16] we adapt this notion to Dirichlet and compositional models using the attraction formulations presented in section 3.2. The (direct) elasticity of the share S_{jt} relative to the media X_{kjt} is equal to $(1 - S_{jt})b_k X_{kjt}$ in the MNL model, DIR model and MNL specification of the GMCI models, whereas it is equal to $(1 - S_{jt})b_k$ in the MCI specification for the GMCI, and to $b_{kjj} - \sum_{m=1}^{D} S_{mt} b_{kmj}$ for the CODA model.
For example, the direct elasticities of market-shares of the three considered brands are computed for the TV channel, for the 123 observed periods, and the average is presented in Table 4. They correspond to the average relative impact on the market-share of brand j, S_j, of a 1% increase of the TV investment of brand j .

	MNL	GMCI	DIR	CODA
DACIA	0.0019	0.0028	-0.0068	-0.0046
NISSAN	**0.0101**	**0.0152**	**0.0389**	**-0.0022**
RENAULT	0.0058	0.0088	0.0145	-0.0038

Table 4: Average direct elasticities for TV investments

We observe that elasticities are not the same across models, and can even be of opposite sign. For example, the DIR model concludes that, on average over the period 2005-2015, if Nissan increases its TV investment by 1% , it will increase its market-share by 0.04%, whereas in CODA, it will have a small negative impact. The CODA model, which includes all cross effects, suggests that the impacts of TV investments of Dacia, Nissan and Renault tend to “cancel each other”, in the sense that all impacts are very close to zero. However, all models except CODA agree on the fact that Nissan has the highest TV's elasticities (in bold in the table).

5 Conclusion

Because of the constraints of shares data, classical regression models cannot be used directly to model market-shares. Market-share models have been developed in the marketing literature, but they fail in estimating brand-specific and cross-effect parameters in a parsimonious fashion.

In this paper, we show that the Dirichlet model (DIR) and the linear compositional regression model (CODA), which are not usually used in this context, can perform better than usual market-share models, thanks to their higher flexibility. We express all these models in attraction form to ease their comparison, and we propose to interpret them in terms of elasticities. We also prove that the generalized multiplicative competitive interaction model (GMCI) can be written as a particular compositional model. We highlight the similarities and the differences of these models. The multinomial logit model (MNL) and DIR are generalized linear models estimated by maximum likelihood and centered on the arithmetic mean shares, whereas GMCI and CODA are transformation models estimated by OLS, centered on the geometric mean shares. MNL and GMCI models without cross-effects are very simple and parsimonious models but fail to capture the variability of the data in our application. The CODA model is the most complex model but it manages to capture all cross effects with a relative parsimony, compared to other models thanks to constraints on parameters, resulting in a good fitting quality. The DIR model is very flexible and it successfully fits the data with less parameters than the CODA model. All these models are implemented in R, and can be interpreted in terms of elasticities.

We use the four models to understand the impact of media investments by channel on brand market-shares in the automobile market, controlling for price and scrapping incentive. We base our choice of models on cross-validation using quality measures adapted for shares data. In our application, the Dirichlet model gives the best out-of-sample results, followed by the CODA model.

We intend to focus in further work on the interpretability of the CODA model. More precisely, direct and cross elasticities have to be deeply interpreted in order to check that the models make sense for the considered application, and to be able to use them to help decision making in practice. Concerning our particular application, the observations are across time. Thus, the potential auto-correlation of error terms should be tested and taken into account if necessary. Moreover, as we measure the impact of media investments on market-shares, considering "adstock function" of media investments instead of pointwise media investments might be more relevant. Adstock functions are often used in the marketing literature, they are cumulative value of past and present advertising expenditures, corresponding to the "carry-over effect" over time. Furthermore, the introduction of random coefficients can be discussed. Such models are considered by Berry, Levinsohn and Pakes [2] in the aggregated MNL framework in econometrics.

Acknowledgment: We would like to thank BVA and the Marketing Direction of Renaultfor sharing valuable data with us, and for their support during the elaboration of the data base and the model specification. We would also like to thank Vera Pawlowsky-Glahn, Marina Vives-Mestres, Juan Jose Egozcue, Josep Antoni Martin Fernandez and Karel Hron for their help to get familiar with the compositional data analysis.

Funding: This work was supported by the market research agency BVA and the ANRT.

References

[1] Aitchison, J. *The statistical analysis of compositional data.* Monographs on statistics and applied probability. Chapman and Hall, 1986.

[2] Berry, S., Levinsohn, J., and Pakes, A. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* (1995), 841–890.

[3] Campbell, G., and Mosimann, J. E. Multivariate analysis of size and shape: modelling with the dirichlet distribution. In *ASA Proceedings of Section on Statistical Graphics* (1987), pp. 93–101.

[4] Chakir, R., Laurent, T., Ruiz-Gazen, A., Thomas-Agnan, C., and Vignes, C. Spatial scale in land use models: application to the teruti-lucas survey. *Spatial Statistics* (2016).

[5] Cooper, L., and Nakanishi, M. *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness.* International Series in Quantitative Marketing. Springer, 1988.

[6] Elff, M. Social divisions, party positions, and electoral behaviour. *Electoral Studies 28*, 2 (2009), 297–308.

[7] Elff, M. *mclogit: Mixed Conditional Logit*, 2014. R package version 0.3-1.

[8] Green, P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), 149–192.

[9] Hijazi, R. H. Residuals and diagnostics in dirichlet regression. *ASA Proceedings of the General Methodology Section* (2006), 1190–1196.

[10] Hijazi, R. H., and Jernigan, R. W. Modelling compositional data using dirichlet regression models. *Journal of Applied Probability and Statistics 4*, 1 (2009), 77–91.

[11] Kynclova, P., Filzmoser, P., and Hron, K. Modeling compositional time series with vector autoregressive models. *Journal of Forecasting 34*, 4 (2015), 303–314.

[12] Maier, M. J. Dirichletreg: Dirichlet regression for compositional data in r. Research Report Series/ Department of Statistics and Mathematics 125, WU Vienna University of Economics and Business, Vienna, January 2014.

[13] Martin-Fernandez, J. A., Bren, M., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. A measure of difference for compositional data based on measures of divergence. In *Proceedings of IAMG* (1999), vol. 99, pp. 211–216.

[14] McFadden, D. L. Econometric analysis of qualitative response models. *Handbook of econometrics 2* (1984), 1395–1457.

[15] Monti, G., Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. Shifted-dirichlet regression vs simplicial regression: a comparison. *Welcome to CoDawork 2015* (2015).

[16] Morais, J., Thomas-Agnan, C., and Simioni, M. Interpreting the impact of explanatory variables in compositional models. *TSE Working Paper* (2017).

[17] Nakanishi, M., and Cooper, L. G. Simplified estimation procedures for mci models. *Marketing Science 1*, 3 (1982), pp. 314–322.

[18] Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, 2015.

[19] Peyhardi, J., Trottier, C., and Guédon, Y. A new specification of generalized linear models for categorical data. *arXiv preprint arXiv:1404.7331* (2014).

[20] So, Y., and Kuhfeld, W. F. Multinomial logit models. In *in SUGI 20 Conference Proceedings, Cary, NC: SAS Institute Inc* (1995).

[21] Solano-Acosta, W., and Dutta, P. K. Unexpected trend in the compositional maturity of second-cycle sand. *Sedimentary Geology 178*, 3 (2005), 275–283.

[22] Templ, M., Hron, K., and Filzmoser, P. *robCompositions: an R-package for robust statistical analysis of compositional data*, 2011.

[23] van den Boogaart, K. G., Tolosana, R., and Bren, M. *compositions: Compositional Data Analysis*, 2014. R package version 1.40-1.

[24] Van Den Boogaart, K. G., and Tolosana-Delgado, R. *Analysing Compositional Data with R*. Springer, 2013.

[25] Wang, H., Liu, Q., Mok, H. M., Fu, L., and Tse, W. M. A hyperspherical transformation forecasting model for compositional data. *European journal of operational research 179*, 2 (2007), 459–468.