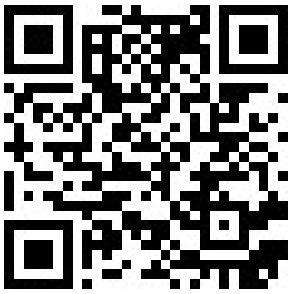
Pak.j.stat.oper.res. Vol.18 No.4 2022 pp 867-882 DOI: http://dx.doi.org/10.18187/pjsor.v18i4.3969

Pakistan Journal of Statistics and Operation Research

f-divergence and parametric regression models for compositional data



Abdulaziz Alenazi[1*]

*Corresponding author

1. Department of Mathematics, College of Science,
Northern Border University, Saudi Arabia, a.alenazi@nbu.edu.sa

Abstract

The paper considers the class of f-divergence regression models as alternatives to parametric regression models (such as Dirichlet regression) for compositional data. The special cases examined in this paper include the Jensen-Shannon, Kullback-Leibler, Hellinger, χ^2 and total variation divergence. Strong advantages of the proposed regression models are a) the absence of parametric assumptions and b) the ability to treat zero values (which commonly occur in practice) naturally. Extensive Monte Carlo simulation studies comparatively assess the performance of the models in terms of bias and an empirical evaluation using real data examining further aspects, such as predictive performance and computational cost. The results reveal that Kullback-Leibler and Jensen-Shannon divergence regression models exhibited high quality performance in multiple directions. Ultimately, penalized versions of the Kullback-Leibler divergence regression are introduced and illustrated using real data rendering this model the optimal model to utilise in practice.

Key Words: compositional data; f-divergence regression models; penalized regression models; Dirichlet regression.

Mathematical Subject Classification: 62H99, 62J02.

1. Background

Compositional data are non-negative multivariate vectors whose variables (typically called components) conveying only relative information. When the vectors are normalized to sum to 1, their sample space is the standard simplex

$$\mathbb{S}^{D-1} = \left\{ (y_1, ..., y_D)^T \middle| y_i \geq 0, \sum_{i=1}^{D} y_i = 1 \right\}, \tag{1}$$

where D denotes the number of components.
Examples of such data may be found in many different fields of study and the extensive scientific literature that has been published on the proper analysis of this type of data is indicative of its prevalence in real-life applications[1].
Analysis of compositional data may be carried out in various ways which can be summarized in two directions: either by transforming the data or not. In the first direction, the most popular transformations are the logarithms of ratios formed by the components (Aitchison, 1982, Aitchison, 2003), some properties of which can be found in Jooa and Lee (2021). It must be stressed though that this school of thought is not the only one and has been heavily criticised by Scealy and Welsh (2014). Other approaches include taking the square root of the data, resulting in data which lie on the hypersphere (Stephens, 1982, Scealy and Welsh, 2011) or Box-Cox type transformations (Greenacre, 2009, Tsagris et al., 2011, Tsagris and Stewart, 2020). The second direction is to apply transformation free parametric models.

[1]For a substantial number of specific examples of applications involving compositional data see Tsagris and Stewart (2020).

These include the Dirichlet distribution (Gueorguieva et al., 2008, Tsagris and Stewart,2018), the simplex distribution (Barndorff-Nielsen and Jørgensen, 1991), the Liouville distribution (Rayens and Srinivasan,1994) and a generalized Dirichlet distribution (Graf, 2020a).

The widespread occurrence of this type of data in numerous scientific fields that involve covariates, has necessitated the need for valid regression models which in turn has led to several developments in this area, many of which have been proposed recently. The first regression model for compositional response data was developed by Aitchison (2003) and was based on the additive log-ratio transformation. Tolosana-Delgado and von Eynatten (2009) also used the additive log-ratio and Egozcue et al. (2012) extended Aitchison's regression model by using a transformation similar to the isometric log-ratio transformation (Egozcue et al., 2003) but instead of the popular Helmert sub-matrix (Tsagris et al., 2011), they employed a data dependent orthogonal matrix. Moving along the same lines, Tsagris (2015b) generalized Aitchison's regression model via the α-transformation. Morais et al. (2018) studied the automobile market shares as a function of media investments using several regression models, whereas Katz and King (1999) studied the effect of predictor variables on the predicted vote using a Bayesian approach. Fry et al. (2000) proposed a new methodology to model Australian household expenditure data and Larrosa (2003) studied the capital composition across different country economics. Iyengar and Dey (2002) investigated the generalized Liouville family of distributions that admits negative or mixed correlation and also contains non-Dirichlet distributions with non-positive correlation. Wang et al. (2007) used a hyper-spherical transformation to model the three components of the primary, secondary, and tertiary sectors in the Chinese economy in order to predict their future sector compositions. Gueorguieva et al. (2008), Hijazi and Jernigan (2009) and Melo et al. 2009) modelled compositional data using Dirichlet regression (DR) while more recently, Graf (2020a) introduced a generalized DR model.

An important issue in compositional data is the presence of zeros that prohibit the use of the logarithmic transformations, an issue that is not addressed in most papers. In the case of the log-ratio and DR models, either zero imputation techniques must be applied prior to fitting the models or proper adjustments must be made to the regression models (Tsagris and Stewart, 2018). Various other alternative strategies have been proposed in the literature. Scealy and Welsh (2011) transformed the data onto the unit hyper-sphere and introduced the Kent regression which treats zero values naturally. Leininger et al. (2013) modelled spatial compositional data with zeros from the Bayesian stance. Mullahy (2015) estimated regression models of economic share data where the shares could take zero values with nontrivial probability. Tsagris (2015a) suggested a non-parametric regression model utilising the Jensen-Shannon (JS) divergence, examined in the present paper, while the α-regression Tsagris (2015b) is another possibility. Murteira and Ramalho (2016) discussed alternative regression models, also applicable when zero values are present, in the field of econometrics, such as the Kullback-Leibler (KL) divergence regression that is also examined in this paper and non-linear least squares regression. More recently, Tsagris and Stewart (2018) introduced the Zero Adjusted Dirichlet Regression (ZADR).

This paper examines from an empirical point of view, the f-divergence regression models and compares them to the parametric DR and ZADR models. The family of f-divergence regression models is suggested for compositional data because they are distribution free and are thus not affected by the true data distribution, they are computationally more efficient, scalable to large sample sizes, they can estimate regression coefficients with low bias and treat zero values naturally. However they are not fully non-parametric as they specify the form of the link function between the compositional data and the covariates. Instead of maximising the likelihood of a parametric model, the proposed regression models rely on some distance minimization, and specifically minimisation of the f-divergence measure. Some of the f-divergence measures examined in this paper satisfy the metric properties, such as the Jensen-Shannon divergence[2], but not all of them[3]. The family of f-divergence regression models is significantly large and hence some regression models will be presented here, namely those that can treat zero values naturally. To the best of our knowledge, not all of the regression models presented here have been considered in the literature.

A Monte-Carlo simulation study investigates the bias of the divergence regression coefficients and compares them to the parametric DR and ZADR models. Secondly, using real data the bias of the regression coefficients of each model is estimated using non-parametric bootstrap. The predictive performance of each regression model is estimated in order to give more insight into their properties and finally, their computational cost is also evaluated. The bias is a statistical measure describing, to some degree, the correctness of the regression estimates. Predictive performance and computational cost are two important aspects in the data science field where practitioners are more interested in fast and accurate predictions, rather than statistical inference.

The paper further introduces penalized regression using the KL divergence. Penalized versions of the KL regression,

[2]Endres and Schindelin (2003) and Österreicher and Vajda (2003), independently, proved that the JS divergence satisfies the triangle inequality.

[3]The KL divergence for instance is not a metric.

such as ridge, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic net are introduced and illustrated using real data. To the best of our knowledge, no one has studied penalized compositional regression without employing a log-ratio transformation. Shimizu (2019) for instance explored elastic net regression but using a log-ratio transformation, which evidently is not applicable when zero values are present. Note also that the main strand of research on penalized regression is directed toward the case where compositional data are predictor variables and not response variables (Xia et al., 2013, Lin et al., 2014, Shi et al., 2016).

The rest of the paper is structured as follows. Section 2 reviews the background of f-divergence measures and presents the relevant f-divergence regression models and two parametric regression models. Section 3 contains simulation studies comparing the aforementioned regression models and empirically evaluating them using real data. Section 4 introduces penalized versions of the KL regression and illustrates their performance using real data, while the last section concludes the paper.

2. f-divergence based regression models and parametric regression models

2.1. f-divergence measures

Let P and Q denote two probability measures. A widely studied family of f-divergences between probability measures is Csiszar's f-divergence Csiszár, 1967 and Csiszár (1974), defined as $D_f(P,Q) = \int_M \frac{dP}{dQ} Q$, where M is a measurable space, $f : [0,\infty) \to (-\infty,\infty]$ is a convex function such that $f(1) = 0$, $f''(1) > 0$ and P is absolutely continuous with respect to Q. It is known that for all probability measures P and Q, the $D_\phi(P,Q)$ is non-negative and when $P = Q$, $D_\phi(P,Q) = 0$.

The choices of f and the corresponding f-divergences considered in the current paper are presented in Table 1. These divergences have been extensively examined and a summary of relationships can be found in Jain and Srivastava (2007), Jain and Saraswat (2012) and Melbourne et al. (2019). It should be noted that the choice of $f(t) = (t-1)\log t$ leads to the Jeffreys divergence $E_f\left\{\left(1 - \frac{g(x)}{h(x)}\right)\log\left(\frac{h(x)}{g(x)}\right)\right\}$, which was not considered here because it is not defined with zeros.

Table 1: A list of the f-divergences considered in the present work.

Type	$f(t)$	$D_f(h,g)$
Kullback–Leibler divergence	$t\log t$	$E_h\left\{\log\left(\frac{h(x)}{g(x)}\right)\right\}$
Jensen-Shannon divergence	$t\log\frac{2t}{1+t} + \log\frac{2}{1+t}$	$E_h\left\{\log\frac{2h(x)}{h(x)+g(x)}\right\} + E_g\left\{\log\frac{2g(x)}{h(x)+g(x)}\right\}$
Hellinger distance	$\frac{1}{2}(\sqrt{t}-1)^2$	$\frac{1}{2}E_h\left\{\left(1 - \sqrt{\frac{h(x)}{g(x)}}\right)^2\right\}$
χ^2-distance	$(t-1)^2$	$E_h\left\{\frac{(h(x)-g(x))^2}{h(x)g(x)}\right\}$
Total variation distance	$\lvert t-1\rvert$	$E_h\left\{\left\lvert 1 - \frac{g(x)}{h(x)}\right\rvert\right\}$

2.2. Minimum f-divergence regression models

The minimum f-divergence regression models are drawn from Table 1. For the regression models considered here, the fitted compositional vectors will be related to the predictor variables ($\mathbf{x}$), via the parametric form $\mathbf{h}_s(\mathbf{B};\mathbf{x})$ defined as

$$\mathbf{h}_s(\mathbf{B};\mathbf{x}) = \left\{ \begin{array}{cc} \frac{1}{1+\sum_{k=2}^{D} e^{\mathbf{x}^T\boldsymbol{\beta}_k}} & \text{if } s = 1 \\ \frac{e^{\mathbf{x}^T\boldsymbol{\beta}_s}}{1+\sum_{k=2}^{D} e^{\mathbf{x}^T\boldsymbol{\beta}_k}} & \text{if } s = 2,...,D, \end{array} \right\} \tag{2}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d)$ with $\boldsymbol{\beta}_s = (\beta_{0s},\beta_{1s},...,\beta_{ps})^T$, $s = 1,\ldots,d$ denoting the set of regression coefficients. This functional form is by no means restrictive but was chosen as it is the most popular in these circumstances.

The KL regression was examined by Murteira and Ramalho (2016)[4], where the matrix of the regression coefficients

[4]Murteira and Ramalho (2016) termed it multinomial logit regression.

$\mathbf{B}$ is estimated via minimization of the KL divergence. Tsagris (2015a) proposed the use of the JS regression as a better alternative to the log-ratio regressions proposed by Aitchison (2003). JS regression springs from minimizing the JS metric between the observed and the fitted compositions with respect to the regression coefficients. The other three competing models are the Hellinger (HG) regression, χ^2 (CS) regression and the Total Variation (TV) regression, where each of them estimates the matrix of the regression coefficients $\mathbf{B}$ via minimization of the HG, the χ^2 and the TV distance, respectively.

- **KL regression**: $\min_{\mathbf{B}} \sum_{i=1}^{n} \mathbf{y}_i^T \log \frac{\mathbf{y}_i}{\mathbf{h}_i(\mathbf{B};\mathbf{x})}$.
- **JS regression**: $\min_{\mathbf{B}} \sum_{i=1}^{D} \left(\mathbf{y}_i^T \log \frac{2\mathbf{y}_i}{\mathbf{y}_i+\mathbf{h}_i(\mathbf{B};\mathbf{x})} + \mathbf{h}_i\left(\mathbf{B};\mathbf{x}\right)^T \log \frac{2\mathbf{h}_i(\mathbf{B};\mathbf{x})}{\mathbf{y}_i+\mathbf{h}_i(\mathbf{B};\mathbf{x})} \right)$.
- **HG regression**: $\min_{\mathbf{B}} \sum_{i=1}^{n} \mathbf{j}_D^T \left(\sqrt{\mathbf{y}_i} - \sqrt{\mathbf{h}_i\left(\mathbf{B};\mathbf{x}\right)} \right)^2$.
- **CS regression**: $\min_{\mathbf{B}} \sum_{i=1}^{n} \mathbf{j}_D^T \frac{\left(\mathbf{y}_i-\mathbf{h}_i(\mathbf{B};\mathbf{x})\right)^2}{\mathbf{h}_i(\mathbf{B};\mathbf{x})}$.
- **TV regression**: $\min_{\mathbf{B}} \sum_{i=1}^{n} \mathbf{j}_D^T \left|\mathbf{y}_i - \mathbf{h}_i\left(\mathbf{B};\mathbf{x}\right)\right|$.

The n denotes the sample size of the observations and $\mathbf{j}_D = (1, \ldots\ldots, 1)^T$ is the D-dimensional vector of $1s$.

2.3. Dirichlet regression with and without zero values

Dirichlet distribution is a natural parametric model for compositional data, since its support is the simplex whose probability density function is

$$f\left(\mathbf{y}\right) = \frac{\Gamma\left(\phi\right)}{\prod_{s=1}^{D} \Gamma\left(\phi\mu_s\right)} \prod_{s=1}^{D} y_s^{\phi\mu_s - 1}, \tag{3}$$

where $\phi > 0$ is the concentration parameter and $0 < \mu_s < 1$ denotes the mean vector of the distribution satisfying $\sum_{s=1}^{D} \mu_s = 1$. This parametrization allows to link μ_s to some covariates using the link function in Eq. (2). DR was extensively examined by Gueorguieva et al. (2008), where the Dirichlet distributed adjusted for covariates yields the following log-likelihood to be maximized with respect to $\mathbf{B}$ and ϕ

$$\ell\left(\mathbf{y};\mathbf{B},\phi\right) = n\log\Gamma\left(\phi\right) - \sum_{i=1}^{n}\sum_{s=1}^{D} \log\Gamma\left(\phi\mathbf{f}_{si}\left(\mathbf{B};\mathbf{x}\right)\right) + \sum_{i=1}^{n}\sum_{s=1}^{D} \left(\phi\mathbf{f}_{si}\left(\mathbf{B};\mathbf{x}\right) - 1\right)\log y_{si}. \tag{4}$$

To tackle the problem of zero values, Tsagris and Stewart (2018) proposed the ZADR model. There are g populations (or groups) corresponding to each observed subset of non-zero components of the compositional vector $\mathbf{Y}$. Let $\mathbf{G}$ denote the vector indexing the non-zero components of $\mathbf{Y}$. Further let $\theta_b = P\left[\mathbf{G} = \mathbf{g}_b\right]$ (the marginal probability that an observation comes from population b) where $\mathbf{g}_b$ is the vector with 1s and 0s corresponding to population b and $\sum_{b=1}^{B} \theta_b = 1$. In general, the density of $\mathbf{Y}$ with non-zero components corresponding to population b^* is then

$$f_{\mathbf{Y}}\left(\mathbf{y}\right) = \sum_{b=1}^{B} f_{\mathbf{Y},\mathbf{G}}\left(\mathbf{y},\mathbf{g}_b\right) = f_{\mathbf{Y},\mathbf{G}}\left(\mathbf{y},\mathbf{g}_{b^*}\right) \tag{5}$$

where $\mathbf{g}_{b^*}$ is the vector of indices corresponding to the nonzero components of $\mathbf{y}$. Note that for $b \neq b^*$, $f_{\mathbf{Y},\mathbf{G}}(\mathbf{y},\mathbf{g}_b) = 0$. Now let $\mathbf{y}_{b^*}$ of length D_{b^*} denote the vector containing the non-zero components of $\mathbf{y}$ and $h_{b^*}(\mathbf{y}_{b^*})$ the density of $\mathbf{Y}_{b^*}$. Eq. (5) can be written as

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y},\mathbf{G}}(\mathbf{y},\mathbf{g}_{b^*}) = f_{\mathbf{Y}|\mathbf{G}}(\mathbf{y}|\mathbf{g}_{b^*})\theta_{b^*} = f_{b^*}(\mathbf{y}_{b^*})\theta_{b^*} \tag{6}$$

While in Stewart and Field (2011) $h_{b^*}(\mathbf{y}_{b^*})$ was chosen to be the multiplicative logistic normal distribution, Tsagris and Stewart (2018) considered the Dirichlet distribution (3). Let S_b denote the set of observations with zeros indexed

by $\mathbf{g}_b$. Each set therefore contains n_b observations having zeros in the same components. The Dirichlet log-likelihood function in (4) when adjusted to accommodate zero values becomes

$$\begin{aligned}\ell\left(\mathbf{y};\mathbf{B},\phi\right) =& n_1\log(\theta_1)+n_1\log\Gamma\left(\phi\right)-\sum_{\{i:\mathbf{y}_i\in\ S_1\}}\sum_{s=1}^{D_1}\log\Gamma\left(\phi\mathbf{h}_{1si}\left(\mathbf{B};\mathbf{x}\right)\right)\\ &+\sum_{\{i:\mathbf{y}_i\in S_1\}}\sum_{s=1}^{D_1}\left(\phi a^{*}_{1si}-1\right)\log y_{1si}+\ldots\\ &+n_0\log(\theta_0)+n_0\log\Gamma\left(\phi\right)-\sum_{\{i:\mathbf{y}_i\in S_0\}}\sum_{s=1}^{D_0}\log\Gamma\left(\phi\mathbf{h}_{0si}\left(\mathbf{B};\mathbf{x}\right)\right)\\ &+\sum_{\{i:\mathbf{y}_i\in S_0\}}\sum_{s=1}^{D_0}\left(\phi\mathbf{h}_{0si}\left(\mathbf{B};\mathbf{x}\right)-1\right)\log y_{0si}\end{aligned} \tag{7}$$

where for $b=1,\ldots,B$, D_b denotes the number of non-zero components in population b, $\sum_{b=1}^{B}n_b=n$, $\sum_{b=1}^{B}\theta_b=1$, D_1 and D_0 denote the number of components without zero values and S_1 and S_0 their associated sample spaces.

2.4. Comments on these regression models

The f-divergence regression models require the correct form of a covariance matrix (Murteira and Ramalho, 2016) which unfortunately is not available in closed form and for this reason non-parametric bootstrap must be employed. Further, the asymptotic properties of their regression coefficients with compositional responses have not been studied. On the contrary, parametric regression models have the benefit of the log-likelihood which implies asymptotic normality of the estimated parameters. However, under model miss-specification, this property holds true only if the distribution is a member of the exponential family (Gourieroux et al., 1984) and hence the Dirichlet class of distributions does not meet this property. These theoretical properties will be examined via both Monte Carlo simulation studies and real data analysis attempting to shed more light on the empirical behavior of the models with respect to their regression coefficients and predictive performance. Finally, the empirical evaluations will provide guidance as to the suitability of each regression model under realistic conditions.

3. Comparison of the regression models

The empirical comparison of the regression models will be based on both simulation studies and real data analysis, covering multiple directions. Those experiments will explore the behavior of the regression models under different circumstances and subsequently will narrow down the available choices. All computations took place using the *R* package *Compositional* (Tsagris et al., 2022).

3.1. Simulation studies

Simulation studies were conducted to illustrate the performance of the f-divergence regression models with the axis of comparison being the bias of the estimated coefficients. Two cases were considered, with and without outliers, irrespective of zero values being present or not. The hard cases will prove more beneficiary as they will provide evidence as to which regression model should be used under more difficult circumstances. The simulation studies will further depict the performance of the (parametric) DR and ZADR models.
For a range of sample sizes $n=(100,300,500,1000,2000,3000,5000)$, the design matrix $\mathbf{X}$, composed of 1 or 3 predictor variables, was generated from a normal distribution with the response compositional data $(\mathbf{Y})$ consisting of D = 4, 7 or 10 components. At first, Euclidean data were generated as

$$\mu_i=\left(e^{\mathbf{x}_1^T\boldsymbol{\beta}_1+e_1},\ldots,e^{\mathbf{x}^T\boldsymbol{\beta}_D+e_D}\right), \tag{8}$$

where $i=1,\ldots,n$, $e_j\sim N(0,1)$, $\ j=2,\ldots,D$. The matrix of regression coefficients $\mathbf{B}=\left(\boldsymbol{\beta}_2^T,\ldots,\boldsymbol{\beta}_D^T\right)$ was generated from a normal distribution, with the constant terms being generated from $N(-2,1)$ and the slopes from

$N(2,1)$. The data were then mapped onto the simplex using the inverse of the additive log-ratio transformation

$$\mathbf{y} = \left(\frac{1}{1+\sum_{j=2}^{D}\mu_j}, \frac{\mu_2}{1+\sum_{j=2}^{D}\mu_j}, \ldots, \frac{\mu_D}{1+\sum_{j=2}^{D}\mu_j} \right)^T.$$

The bias measured as the difference between the true ($\mathbf{B}$) and the estimated ($\widetilde{\mathbf{B}}$) regression coefficients from either regression model was computed via the Frobenius norm $\left\|\widetilde{\mathbf{B}} - \mathbf{B}\right\|_F$ averaged over 500 repetitions. The data were intentionally not generated from a Dirichlet distribution since DR is tested and because a question of interest is to examine the behavior of the DR under model miss-specification.

The aforementioned scenario was repeated by adding zeros, completely at random, to the data. A random sample of 25% of the observations was selected and for each compositional vector, 40% of the components were set to zero. For example, when $D = 4$, 1 component was set to zero, when $D = 7$ there were 2 components and when $D = 10$, 3 components were set to zero. Those vectors were again normalized to sum to 1.

3.1.1. The case of no outliers

The results of the estimated bias with 1 and 3 predictor variables are presented in Tables 2 and 3 respectively. Table 2 contains the results with 1 predictor variable. When $D = 4$ components, JS regression produces the most accurate regression coefficients, regardless of zero values being present or not. KL and TV regression models are the second best options, whereas HG and DR, ZADR perform the worst and CS regression is somewhere in the middle. With $D = 7$ components the results are nearly similar. The exception is that in the presence of zero values, the CS does surprisingly well with large sample sizes. JS and TV perform satisfactorily well, relative to their competitors. Finally, with $D = 10$ components, the conclusions have changed and the DR and ZADR are the most accurate models. When no zero values are present, DR estimates coefficients with the lowest bias, while with zero values present CS regression is preferred with ZADR being second best, and JS ranking third best. The overall picture is not clear as there is no unanimously best regression for either case of zero values being present or not.

Moving on to Table 3 which contains the results with 3 predictor variables, the conclusions change. The HG regression model still remains the least accurate, but the DR and ZADR models now perform poorly in comparison to their competitors. With no values present, JS, KL and TV regression models are to be preferred, regardless of the number of components with their performance varying according to the sample size and the number of components. With zero values present though, the optimal regression models seem to be TV and JS, with the latter performing substantially better with increasing sample sizes. The KL regression is the third best option, with ZADR and CS performing three times poorer than KL regression.

Summing up, there are two distinct directions. With no zero values, the choice between KL, JS or TV regression is not clear enough. With zero values present though, TV regression has evidently outperformed KL and JS, in almost all cases.

3.1.2. The case of outliers

Robustness properties of f-divergence estimators have been studied but mainly in the univariate space and not in the multivariate space as is the focus of the current paper. For instance, Jiménz and Shao (2001) showed that no minimum divergence estimator is better than the minimum Hellinger distance estimator, in terms of both second-order efficiency and robustness. However, it has already been observed that the HG regression model produced the worst coefficients in terms of bias. Hence, in order to empirically assess the effect of the outliers on the bias of the regression models, the second simulation study scenario involves outliers.

Euclidean data were generated using Eq. (8), their norm is computed and a fraction equal to 5% of those with the highest norm is selected and their values are tripled. This way those observations become very influential and deviate from the rest of the data. Further, zero values are very likely to occur thus making the estimation even harder. In order to further increase the difficulty, this scenario was implemented with 3 predictor variables only. The case of 1 predictor variable seemed rather easy and was excluded. Since the DR/ZADR models and the HG regression were

Table 2: Relative bias $\left(\left\|\widetilde{\mathrm{B}}-\mathrm{B}\right\|_F\right)$ of the regression coefficients for each regression model with 1 predictor variable. Values greater than 1 indicate higher bias than the KL regression model, whereas values less than 1 indicate lower bias.

	No zero values present						Zero values present					
	D = 4 components											
Sample size	KL	JS	HG	CS	TV	DR	KL	JS	HG	CS	TV	ZADR
n=100	1.000	0.811	6.297	1.310	1.143	1.614	1.000	0.690	4.107	1.311	1.079	2.170
n=300	1.000	0.694	6.597	1.418	0.953	1.334	1.000	0.589	6.015	1.955	1.129	3.632
n=500	1.000	0.681	6.613	1.415	0.933	1.489	1.000	0.567	7.126	2.485	1.244	4.419
n=1000	1.000	0.659	6.654	1.426	0.885	1.050	1.000	0.521	8.503	2.949	1.322	5.254
n=2000	1.000	0.654	6.612	1.427	0.864	1.152	1.000	0.407	9.716	3.576	1.434	6.089
n=3000	1.000	0.653	6.704	1.428	0.869	1.301	1.000	0.397	10.387	3.830	1.550	6.466
n=5000	1.000	0.655	6.626	1.427	0.865	1.208	1.000	0.302	10.520	3.929	1.611	6.612
	D = 7 components											
Sample size	KL	JS	HG	CS	TV	DR	KL	JS	HG	CS	TV	ZADR
n=100	1.000	0.634	3.451	1.595	0.785	1.036	1.000	0.675	5.268	1.316	0.972	0.917
n=300	1.000	0.602	3.303	1.635	0.705	1.109	1.000	0.760	4.917	1.195	1.182	1.254
n=500	1.000	0.600	3.287	1.638	0.710	1.097	1.000	0.768	5.158	1.065	1.174	1.348
n=1000	1.000	0.594	3.269	1.646	0.710	1.106	1.000	0.828	5.708	0.865	1.299	1.547
n=2000	1.000	0.593	3.256	1.648	0.703	1.102	1.000	0.818	5.527	0.701	1.253	1.484
n=3000	1.000	0.595	3.252	1.645	0.707	1.103	1.000	0.817	5.509	0.619	1.257	1.497
n=5000	1.000	0.595	3.249	1.643	0.705	1.103	1.000	0.822	5.549	0.525	1.261	1.461
	D = 10 components											
Sample size	KL	JS	HG	CS	TV	DR	KL	JS	HG	CS	TV	ZADR
n=100	1.000	0.601	1.920	1.650	0.685	0.288	1.000	0.607	12.220	1.387	0.866	0.387
n=300	1.000	0.574	1.787	1.720	0.664	0.305	1.000	0.692	2.677	1.351	1.048	0.562
n=500	1.000	0.573	1.746	1.726	0.652	0.298	1.000	0.749	2.869	1.206	1.124	0.601
n=1000	1.000	0.571	1.745	1.730	0.653	0.306	1.000	0.805	2.142	1.012	1.239	0.695
n=.2000	1.000	0.571	1.736	1.737	0.651	0.306	1.000	0.801	2.152	0.803	1.222	0.688
n=3000	1.000	0.569	1.752	1.740	0.646	0.308	1.000	0.800	2.134	0.683	1.221	0.699
n=5000	1.000	0.568	1.735	1.742	0.646	0.307	1.000	0.804	2.145	0.616	1.219	0.695

shown to be highly biased in comparison to the other regression models, it was decided to exclude them from the outliers investigation scenario.

Table 4 contains the results of the estimated coefficient bias for each of the 4 f-divergence regression models, namely KL, JS, CS and TV regression. Regardless of the number of components the CS regression model produced the most biased regression coefficients. On the antipode, KL and JS regression models were the most robust with KL being the most robust. The TV regression produced coefficients whose bias is twice and thrice the bias of the JS and of the KL regression models respectively. Evidently, the KL regression model has outperformed the other 3 f-divergence regression models, namely, JS, CS and TV.

3.2. Empirical comparison

Table 5 contains information on 13 real datasets, most of which do not contain zero values. The datasets with zero values present may contain either a small or a very large portion of compositional vectors with at least one zero element. Both the sample size and number of components vary greatly, but the number of covariates is either 1, 2 or 8.

3.2.1. Estimated bias

The first axis of comparison is the estimated bias of regression coefficient when applied to those 13 real datasets. The models were be fitted to the datasets and their coefficients were extracted and were treated as the true ones. The

Table 3: Bias $\left(\left\|\widetilde{\mathbf{B}}-\mathbf{B}\right\|_F\right)$ of the regression coefficients for each regression model with 3 predictor variables. Values greater than 1 indicate higher bias than the KL regression model, whereas values less than 1 indicate lower bias.

	No zero values present						Zero values present					
	D = 4 components											
Sample size	KL	JS	HG	CS	TV	DR	KL	JS	HG	CS	TV	ZADR
n=100	1.000	0.996	18.765	1.482	1.744	8.382	1.000	0.608	3.351	1.769	0.779	1.903
n=300	1.000	0.758	22.854	2.083	1.233	11.748	1.000	0.270	2.305	2.049	0.222	1.635
n=500	1.000	0.646	22.645	2.179	0.984	11.887	1.000	0.281	1.868	2.021	0.183	1.534
n=1000	1.000	0.522	25.279	2.402	0.951	13.466	1.000	0.278	2.047	2.121	0.112	1.470
n=2000	1.000	0.431	25.034	2.414	0.815	13.414	1.000	0.331	2.141	2.253	0.075	1.693
n=3000	1.000	0.393	24.788	2.417	0.751	13.340	1.000	0.294	2.269	2.226	0.114	1.662
n=5000	1.000	0.375	24.542	2.398	0.706	13.254	1.000	0.291	2.287	2.297	0.078	1.652
	D = 7 components											
Sample size	KL	JS	HG	CS	TV	DR	KL	JS	HG	CS	TV	ZADR
n=100	1.000	1.099	24.590	1.138	1.549	3.966	1.000	0.670	4.186	1.493	0.573	0.881
n=300	1.000	1.068	36.785	1.372	1.536	4.937	1.000	0.451	3.481	1.895	0.374	1.168
n=500	1.000	1.042	44.174	1.509	1.492	4.075	1.000	0.492	3.590	2.073	0.341	1.063
n=1000	1.000	1.022	65.585	1.905	1.477	4.695	1.000	0.411	3.180	2.150	0.178	0.915
n=2000	1.000	1.082	89.163	2.236	1.714	4.688	1.000	0.407	4.053	2.229	0.206	1.137
n=3000	1.000	0.995	102.149	2.685	1.674	4.471	1.000	0.366	3.339	2.520	0.173	0.956
n=5000	1.000	1.042	138.104	3.538	1.989	4.852	1.000	0.377	3.991	2.423	0.127	1.210
	D = 10 components											
Sample size	KL	JS	HG	CS	TV	DR	KL	JS	HG	CS	TV	ZADR
n=100	1.000	0.942	20.434	1.097	1.456	0.964	1.000	0.362	6.354	1.661	0.729	0.389
n=300	1.000	0.818	21.979	1.224	1.227	1.019	1.000	0.321	5.625	2.246	0.469	0.453
n=500	1.000	0.769	21.588	1.269	1.161	0.949	1.000	0.316	7.055	2.323	0.390	0.231
n=1000	1.000	0.742	22.413	1.296	1.143	1.165	1.000	0.249	4.515	2.934	0.183	0.768
n=2000	1.000	0.732	22.267	1.302	1.149	1.200	1.000	0.267	4.275	2.832	0.106	0.753
n=3000	1.000	0.730	22.099	1.313	1.132	1.063	1.000	0.220	5.087	2.867	0.163	0.500
n=5000	1.000	0.723	21.996	1.322	1.124	1.185	1.000	0.236	4.794	3.009	0.134	0.501

coefficients were then estimated again using non-parametric bootstrap and the Frobenius norm of their difference served as an estimate of their bias.

Table 6 contains the results that clearly show that HG regression produced highly biased estimates in comparison to its competitors, in some cases, regardless of zero values being present or not. When no zero values are present, the DR performed equally well to the the f-divergence regression models. For data with zero values though, ZADR produced really biased coefficients in 3 datasets, showing signs of failure. However, in 2 datasets, the reported bias of ZADR was the lowest. Further, it is worthy to remark that TV regression exhibited really high bias in 1 dataset.

3.2.2. Predictive performance

The previous axis of comparison can be characterised as a purely statistical one, while the predictive performance is closer to the machine learning field where it is considered of higher importance than the model's statistical inference.

The 10-fold cross-validation (CV) procedure was implemented in order to estimate the predictive performance of the regression models. The 10-fold CV pipeline splits the dataset into 10 mutually exclusive folds or sets at random. One fold is left aside playing the role of the test fold, while the other 9 are collected in what is termed the training fold. The regression model is applied to the training fold and using the values of the covariates of the test fold the regression models predict the compositional response values of the test fold. The KL and JS divergences evaluate the regression model's predictive performance. This process is repeated 10 times so that all folds have played the role of the test set. However, to avoid optimistic results due to random chance, the aforementioned 10-fold CV process was repeated 10 times with different folds (splits) each time and the models' predictive performance was calculated from aggregation over the 10 folds across all repetitions.

Table 4: Bias $\left(\left\|\widetilde{\mathbf{B}}-\mathbf{B}\right\|_F\right)$ of the regression coefficients for each regression model with 3 predictor variables, when outliers are present. Values greater than 1 indicate higher bias than the KL regression model, whereas values less than 1 indicate lower bias.

	D = 4 components				D = 7 components				D = 10 components			
	KL	JS	CS	TV	KL	JS	CS	TV	KL	JS	CS	TV
n=100	1.000	1.170	13.943	1.774	1.000	1.250	12.135	1.822	1.000	1.227	20.045	1.955
n=300	1.000	1.200	21.500	1.800	1.000	1.414	15.600	2.079	1.000	1.243	31.765	2.044
n=500	1.000	1.185	23.444	1.778	1.000	1.564	18.479	2.376	1.000	1.265	35.120	2.171
n=1000	1.000	1.273	28.227	1.909	1.000	1.814	22.031	2.876	1.000	1.286	43.242	2.440
n=2000	1.000	1.368	32.526	2.105	1.000	1.906	24.576	3.059	1.000	1.436	50.090	2.782
n=3000	1.000	1.421	32.632	2.158	1.000	1.898	23.818	3.011	1.000	1.544	57.279	3.176
n=5000	1.000	1.444	34.167	2.278	1.000	1.976	25.169	3.145	1.000	1.594	60.594	3.234

Table 5: Information on the datasets

Dataset	Dimensions ($n \times D$)	Proportion (%) of vectors with at least one zero value	Number of predictors	Reference
Bayesite	21×4	0.00	2	Aitchison (2003)
Box	25×5	0.00	1	Aitchison (2003)
Cars	144×5	0.00	8	Graf (2020b)
Cox	25×5	0.00	2	Aitchison (2003)
Firework	81×5	0.00	2	Aitchison (2003)
GDP	27×12	0.00	1	Eurostat
Lake	39×3	0.00	1	Aitchison (2003)
Fish	75×26	0.00	1	Greenacre (2018)
Data	89×9	89.90	1	Rozenas (2015)
Elections	67×10	34.32	8	Smith (2002)
Foraminifers	30×4	16.67	1	Aitchison (2003)
Gemas	2083×22	0.04	2	Templ et al. (2011)
Glacial	92×3	13.04	2	van den Boogaart et al. (2018)

Table 7 presents the KL and JS divergences of each regression model applied to each dataset. HG regression produced highly biased regression coefficients whose effect is reflected in the model's predictive performance. Its predictive performance was the worst observed and in some cases, it was multiple times higher. With the Box, Cars, Cox and Gemas datasets, for instance, the HG proved very weak in providing accurate predictions based on the KL divergence. When the predictive performance is measured via the JS divergence, the conclusions were similar, only this time the differences were less magnified. This behavior of the models' predictive performance is independent of zero values being present or not.

3.2.3. Computational cost

The computational cost is the third axis of comparison among the 6 candidate regression models. The results are presented in Table 8. With no zero values present, the KL regression was the faster overall, with DR coming last. On datasets with zero values present and with larger sample sizes and number of components the computational cost increased dramatically. Especially for the ZADR model which required minutes to fit in these 5 datasets. KL evidently outperformed all of its competitors. HG regression required a reasonable amount of time, but at the cost of high bias as already discussed.

The computational cost assists in making decisions as to which regression model to use and for what purpose. When the interest lies in obtaining accurate regression coefficients, the computational cost is not that important and the answer relies upon the results presented in Tables 2 and 3. In that instance, the safest choices are the KL, JS and TV regression models. If, on the contrary, the interest lies in accurate predictions the computational cost contributes significantly to the choice of the regression model. As depicted in Table 7, all regression models except for HG

Table 6: Bootstrap based estimated bias.

	Regression model					
Dataset	KL	JS	HG	CS	TV	DR
Bayesite	1.000	1.040	2.278	0.940	1.013	0.982
Box	1.000	1.000	2.000	1.000	1.267	0.967
Cars	1.000	1.000	1.000	1.000	1.000	1.011
Cox	1.000	1.000	1.969	1.000	1.021	1.010
Firework	1.000	1.089	2.285	0.905	1.386	0.892
GDP	1.000	0.992	2.210	1.039	1.069	0.946
Lake	1.000	1.085	2.463	0.902	1.183	1.000
Fish	1.000	1.000	1.817	1.000	1.000	0.986
Dataset	KL	JS	HG	CS	TV	ZADR
Data	1.000	0.671	12.630	0.671	14.616	0.808
Elections	1.000	0.999	1.050	0.998	1.000	0.976
Foraminimers	1.000	0.929	4.089	1.089	1.455	0.571
Gemas	1.000	1.053	1.003	1.057	1.003	1.087
Glacial	1.000	1.158	15.342	0.842	1.105	1.055

Table 7: Predictive performance measured by Kullback-Leibler and Jensen-Shannon divergences. Values greater than 1 indicate worst performance than the KL regression model, whereas values less than 1 indicate better performance.

	Kullback-Leibler divergence						Jensen-Shannon divergence					
	Regression model											
Dataset	KL	JS	HG	CS	TV	D	KL	JS	HG	CS	TV	D
Bayesite	1.000	1.006	1.392	0.994	1.006	0.994	1.000	1.010	1.709	0.981	1.019	0.990
Box	1.000	1.016	3.143	1.016	1.000	1.016	1.000	1.000	6.333	1.000	1.000	1.000
Cars	1.000	1.000	7.739	1.000	1.043	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Cox	1.000	1.000	12.500	1.000	1.111	1.000	1.000	1.000	8.000	1.000	1.000	1.000
Firework	1.000	1.010	1.087	0.995	1.026	1.031	1.000	1.000	1.160	1.011	1.043	0.995
GDP	1.000	1.000	3.588	1.000	1.000	1.000	1.000	0.938	6.312	1.000	1.000	0.938
Lake	1.000	1.030	1.881	0.978	0.970	1.037	1.000	1.000	1.978	1.022	1.011	1.011
Fish	1.000	1.000	15.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Dataset	KL	JS	HG	CS	TV	ZADR	KL	JS	HG	CS	TV	ZADR
Data	1.000	0.961	2.701	0.948	3.753	1.000	1.000	0.884	3.581	0.907	1.279	0.930
Elections	1.000	1.000	2.462	1.000	1.000	3.204	1.000	1.000	4.571	1.000	1.000	1.957
Foraminimers	1.000	1.012	2.406	1.056	1.056	0.981	1.000	0.985	2.149	1.179	0.993	0.985
Gemas	1.000	1.083	4.015	1.000	1.125	1.208	1.000	0.952	5.619	1.286	1.000	1.476
Glacial	1.000	1.006	3.579	1.003	1.088	1.118	1.000	1.007	1.287	1.009	1.178	1.206

produce predictions of nearly similar accuracy. Further, the computational cost plays an important role in selecting the regression model. KL regression is the apparent winner as it requires fractions of a second to fit to all 13 datasets.

4. Penalized KL regression

The 21st century signified the explosion of data, especially of big data. One such example is the case of regression with numerous possible predictor variables for which a variable selection must occur prior to fitting the regression model. The most famous variable selection algorithms rely on constrained optimization or penalization.

Ridge regression (Hoerl and Kennard, 1970) is the first penalized regression that poses a restriction on the square form of the coefficients. Ridge regression in some consists of minimizing the following function

$$\min_{\boldsymbol{\Theta}} D\left(\mathbf{y};\boldsymbol{\Theta}\right)+\frac{\lambda}{2}\left\|\mathbf{B}\right\|_{F}^{2}, \tag{9}$$

where $\boldsymbol{\Theta}=(\mathbf{B},\phi)$ denotes the parameter space in general, so as to account for both f-divergence regression models and for DR/ZADR models. The function $D\left(\mathbf{y};\boldsymbol{\Theta}\right)$ can be any of the aforementioned divergences or the negative of the log-likelihood of any parametric model (not just of DR or ZADR). The λ is the Lagrangian parameter measuring the weight of the constraint and $\left\|\mathbf{B}\right\|_{F}$ denotes the Frobenius norm of the matrix of regression coefficients $\mathbf{B}$. The choice

Table 8: Duration (in seconds) of the regression models.

	Regression model					
Dataset	KL	JS	HG	CS	TV	D
Bayesite	0.00	0.04	0.04	0.03	0.03	0.05
Box	0.00	0.02	0.02	0.01	0.03	0.03
Cars	0.00	0.17	0.97	0.24	0.19	4.73
Cox	0.00	0.01	0.04	0.06	0.08	0.09
Firework	0.00	0.08	0.06	0.06	0.11	0.18
GDP	0.00	0.19	0.34	0.14	0.17	0.37
Lake	0.00	0.01	0.01	0.01	0.01	0.02
Fish	0.04	0.39	3.39	1.86	0.74	6.01
Total	0.05	0.91	4.87	2.41	1.38	11.47
	KL	JS	HG	CS	TV	ZADR
Data	0.00	0.18	0.21	0.08	0.19	0.45
Elections	0.01	4.63	3.14	3.08	2.35	55.07
Foraminifers	0.00	0.01	0.02	0.01	0.02	0.03
Gemas	0.44	188.40	2.01	92.78	93.10	781.03
Glacial	0.00	0.04	0.02	0.02	0.02	0.11
Total	0.46	193.26	5.39	95.97	95.68	836.69

of λ is data dependent and is ordinarily made via cross-validation.

Least Absolute Selection and Shrinkage Operator (LASSO) on the other hand imposes the absolute norm constraint on the regression coefficients, that is, minimization of the following function

$$\min_{\boldsymbol{\Theta}} D\left(\mathbf{y};\boldsymbol{\Theta}\right)+\lambda\sum_{s=2}^{D}\left\|\boldsymbol{\beta}_s\right\|_1, \tag{10}$$

where $\left\|\boldsymbol{\beta}_s\right\|_1$ is the L_1 norm of the vector $\boldsymbol{\beta}_s$. Elastic net is a weighted combination of ridge (9) and LASSO (10)

$$\min_{\boldsymbol{\Theta}} D\left(\mathbf{y};\boldsymbol{\Theta}\right)+\lambda\left(\frac{1-\alpha}{2}\left\|\mathbf{B}\right\|_F^2+\alpha\sum_{s=2}^{D}\left\|\boldsymbol{\beta}_s\right\|_1\right), \tag{11}$$

where $\alpha\in[0,1]$. Evidently, when $\alpha=0$ one ends up with ridge regression, while when $\alpha=1$ LASSO emerges.

It is well known that LASSO is likelihood or distance-based dependent, meaning that LASSO solution is not generic and applicable to all models. For instance, LASSO implementation of ZADR differs from LASSO implementation of DR. The same is true for all f-divergence regression models. The *R* package *glmnet* (Friedman et al., 2010) implements LASSO multinomial logistic regression which is in fact a penalized minimization of the KL divergence. This enables its use in the compositional data regression setting and thus the resulting regression shall be termed LASSO-KL regression. KL regression is the only model, among those examined in this paper, that can be implemented with all three types of penalization, LASSO, ridge, and elastic net. Further, it is highly computationally efficient, rendering it currently the only available choice for penalized f-divergence regression.

The main difference however is that *glmnet* adopts a different link function. The link function in Eq. (2) is the inverse of the additive log-ratio transformation (Aitchison, 1982, Aitchison, 2003), whereas *glmnet* utilises the inverse of the centred log-ratio transformation and does not include the constant term.

$$\mathbf{h}_s\left(\mathbf{B};\mathbf{x}\right)=\frac{e^{\mathbf{x}^T\boldsymbol{\beta}_s}}{\sum_{k=1}^{D}e^{\mathbf{x}^T\boldsymbol{\beta}_k}}\quad\text{for }\ s=1,...,D. \tag{12}$$

The implications of (12) are straightforward. For compositional data with D components and p predictor variables, the f-divergence (and the parametric) regression models using (2) would estimate $(D-1)\times(p+1)$ regression coefficients. Their interpretation is the expected log change of each component relative to the first component. On the contrary, LASSO-KL regression, uses (12) and estimates $D\times(p+1)$ regression coefficients, whose interpretation is the logarithmic change of each component relative to the mean. This implies that the penalisation takes place over all D sets of regression coefficients, i.e. $s=1,\ldots,D$ and not $s=2,\ldots,D$ as in Eqs. (9)-(11). The L_1 norm constraint takes care of the redundancies (Friedman et al., 2010). Unfortunately, penalized f-divergence regression

models specifically for compositional data have not been explored and hence their theoretical properties are not known.

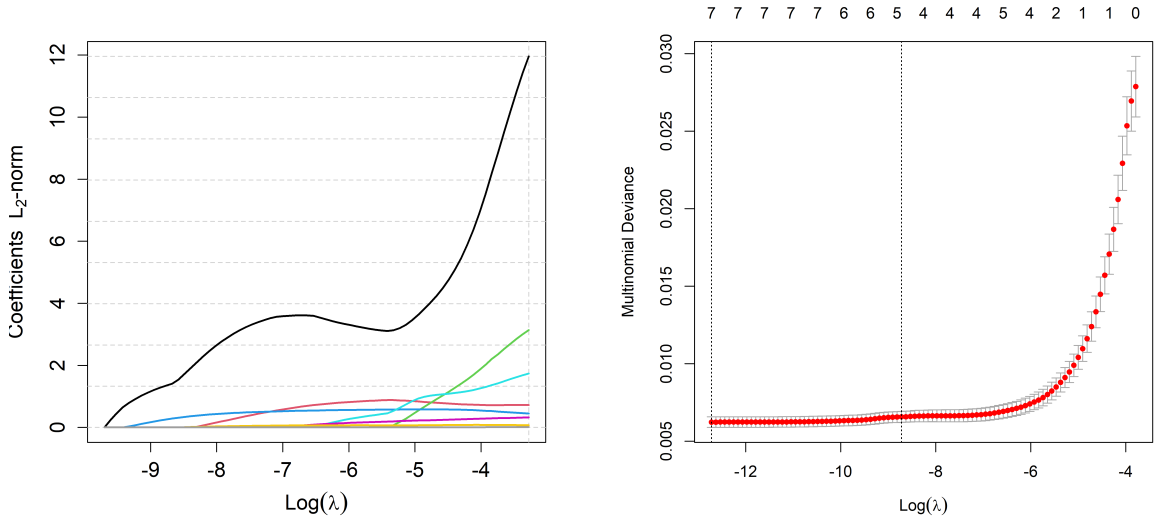
The computational cost revealed that in order to apply ridge, LASSO or elastic net, a computationally efficient regression model, regardless of zero values being present or not, must be employed. Evidently, DR and ZADR models and most f-divergence regression models can be computationally prohibitive, especially with large sample sizes. KL regression model is to be preferred as it is computationally efficient and its predictive performance was shown to be similar to that of JS, CS and TV regression. But, *glmnet*'s implementation allows only for LASSO-KL regression and this is why KL regression is selected for further evaluation and assessment.

An empirical evaluation will take place illustrating the performance of LASSO-KL regression. The most suitable datasets, those with more that two predictor variables, are *Cars* and *Elections*. The first dataset is zero value free, while the second contains a high proportion of zero values.

4.1. Cars dataset

The Cars dataset comprises of segment shares of 144 car sales in 5 categories according to the size of the car chassis, with 8 predictor variables. The dataset is publicly available via the *R* package *SGB* (Graf, 2020b).

Figure 1 illustrates the performance of LASSO-KL regression. In (a) the trace plot appears. That is, the Euclidean norm of each set of coefficients is plotted versus the logarithm of the λ values. Each coloured line corresponds to the Euclidean norm of the coefficients of each predictor variable. In (b) performance of the cross-validated LASSO-KL regression is depicted. The estimated KL divergence (with error bars) for is plotted for various values of λ. The minimum KL divergence (0.006113) is achieved when $\lambda = 2.99 \times 10^{-6}$ and is denoted with the left dotted vertical line. However, according to Hastie et al. (2009) the "one-standard-error" rule should be used with cross-validation. The value of λ corresponding to the most parsimonious model whose error is no more than one standard error above the error of the best model should be chosen. Based on this rule the value $\lambda = 1.236 \times 10^{-4}$ (highlighted with the right dotted line) is chosen and its corresponding KL divergence is equal to 0.006370. LASSO with $\lambda = 0$ is equivalent to the KL regression, that is no penalisation is applied. By examining Figure 1(b) is can be observed that a value of λ greater than zero does offers a significant improvement in the KL regression's predictive performance as the the KL divergence when $\lambda = 0$ is equal to 0.023340 which is 3.664 times the KL divergence when $\lambda > 0$.



(a) (b)

Figure 1: Cars dataset: (a) LASSO coefficients plot and (b) Cross-validated deviance as a function of $\log(\lambda)$.

4.2. Elections dataset

The Elections dataset contains the results of the 2000 U.S. presidential election in the 67 counties of Florida. The number of votes each of the 10 candidates received and were normalized to compositional and there are 8 predictor variables that can be linked to these compositional data. In 23 out of those 67 counties some candidates received no vote and hence zero values appear.

Figure 2 presents the results of LASSO-KL regression applied to this dataset. Figure 2(a) shows the trace plot of the set of coefficients of each predictor variable while and the results of the 10-fold cross-validation are presented in Figure 2(b). The chosen value of λ following the "one-standard-error" rule is 0.009006 with a KL divergence equal to 0.02717. This time the error bars are wider showing greater variability in LASSO-KL regression's predictive performance. KL regression (LASSO-KL regression with $\lambda = 0$) on the other hand produces a KL divergence equal to 0.02534, which is 0.933 the KL divergence estimated by LASSO-KL regression and this is evident in Figure 2(b).

This motivated the use of the elastic net as a means of improving the performance of LASSO-KL regression. 11 equidistant values of α, from 0 up to 1 were used and for each value of α cross-validation of LASSO-KL regression was implemented. Figure 2(c) visualises the results, where the "one-standard-error" rule was again followed. Evidently, the minimum KL divergence was observed when $\alpha = 0$, which corresponds to ridge regression. In that instance the KL divergence was equal to 0.02733, which is again higher than the KL divergence of the simple KL regression.

This example showed that neither LASSO-KL regression, ridge or elastic net yielded better performance than simple KL regression. This conveys the message that some times simpler regression models perform better than more complex or more advanced regression models.

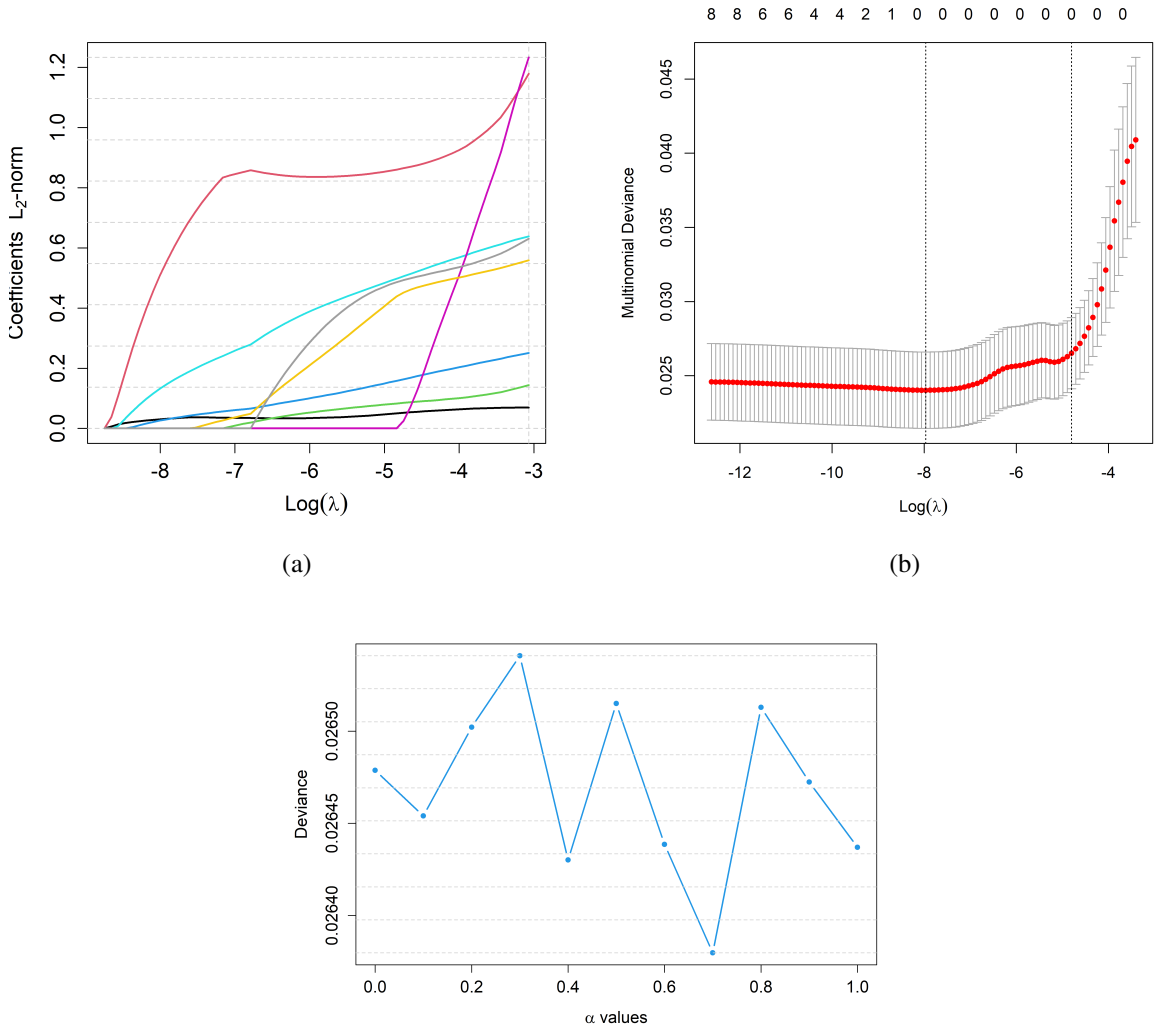
5. Conclusions

The perk of the divergence regression models is their ability to treat compositional data with zero values present, unlike parametric regression models for which adaptation is required, as in the case of the DR The downside of the first family of regression models though is the lack of straightforward statistical inference as they are likelihood-free. Secondly, they have not attracted research interest, especially in the field of compositional data. Parametric models enjoy nice theoretical properties as they are likelihood based, but the case of DR proved weak in the simulation studies. Since divergence regression models have not been widely studied in the context of compositional data we realized the need to perform a large scale comparison between them and a parametric competitor which is able to account for zero values. Based on simulation studies we concluded that the Hellinger divergence regression model should not be used as it produces highly unbiased regression estimates, regardless of zero values present in the data, and has low predictive performance. DR and ZADR models were also shown to produce biased estimates. With outliers present, KL and JS regression models seem to be the optimal models and they exhibited a comparable performance. This helps limit the choice of which regression model to use and when, to these two models. If the goal is to build a fast predictive model, then KL regression is apparently preferable to the JS regression model.

Based on the real data, DR performed satisfactorily, whereas ZADR failed in most cases. The performance of the f-divergence regression models was similar to that observed in the simulation studies. When it came to predictive performance, the results were similar, as HG had the smallest accuracy, while the other regression models produced similar predictions. Among them, the KL regression is to be selected as it can build a regression model in fractions of a second. Further, its penalized version, that it also extremely fast, renders the KL regression a highly attractive model for practical purposes.

References

1. Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B*, 44(2):139–177.
2. Aitchison, J. (2003). *The statistical analysis of compositional data*. New Jersey: Reprinted by The Blackburn Press.
3. Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39(1):106–116.
4. Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318.



(a) (b) (c)

Figure 2: Elections dataset: (a) LASSO coefficients plot and (b) Cross-validated deviance as a function of $\log(\lambda)$.

5. Csiszár, I. (1974). Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 73–86.
6. Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.
7. Egozcue, J. J., Daunis-I-Estadella, J., Pawlowsky-Glahn, V., Hron, K., and Filzmoser, P. (2012). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, 6(182):87–108.
8. Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.
9. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).
10. Fry, J. M., Fry, T. R., and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Applied Economics*, 32(8):953–959.
11. Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society*, 52(3):681–700.
12. Graf, M. (2020a). Regression for compositions based on a generalization of the Dirichlet distribution. *Statistical Methods & Applications*, 29(4):913–936.
13. Graf, M. (2020b). *SGB: Simplicial Generalized Beta Regression*. R package version 1.0.1.
14. Greenacre, M. (2009). Power transformations in correspondence analysis. *Computational Statistics & Data*

Analysis, 53(8):3107–3116.

15. Greenacre, M. (2018). *Compositional Data Analysis in Practice*. Chapman & Hall/CRC Press.
16. Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis*, 52(12):5344–5355.
17. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
18. Hijazi, R. and Jernigan, R. (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics*, 4(1):77–91.
19. Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12(1):69–82.
20. Iyengar, M. and Dey, D. K. (2002). A semiparametric model for compositional data analysis in presence of covariates on the simplex. *Test*, 11(2):303–315.
21. Jain, K. and Saraswat, R. (2012). A new information inequality and its application in establishing relation among various f-divergence measures. *Journal of Applied Mathematics, Statistics and Informatics*, 8(1):17–32.
22. Jain, K. and Srivastava, A. (2007). On symmetric information divergence measures of Csiszar's f-divergence class. *Journal of Applied Mathematics, Statistics and Informatics (JAMSI)*, 3(1):85–102.
23. Jiménz, R. and Shao, Y. (2001). On robustness and efficiency of minimum divergence estimators. *Test*, 10(2):241–248.
24. Jooa, J. Y. and Lee, S. (2021). Binary classification on compositional data. *Communications for Statistical Applications and Methods*, 28(1):89–97.
25. Katz, J. and King, G. (1999). A statistical model for multiparty electoral data. *American Political Science Review*, 93(1):15–32.
26. Larrosa, J. (2003). A compositional statistical analysis of capital stock. In *Proceedings of the 1st Compositional Data Analysis Workshop, Girona, Spain*.
27. Leininger, T. J., Gelfand, A. E., Allen, J. M., and Silander Jr, J. A. (2013). Spatial Regression Modeling for Compositional Data With Many Zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):314–334.
28. Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797.
29. Melbourne, J., Madiman, M., and Salapaka, M. V. (2019). Relationships between certain f-divergences. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1068–1073. IEEE.
30. Melo, T. F., Vasconcellos, K. L., and Lemonte, A. J. (2009). Some restriction tests in a new class of regression models for proportions. *Computational Statistics & Data Analysis*, 53(12):3972–3979.
31. Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Using compositional and Dirichlet models for market share regression. *Journal of Applied Statistics*, 45(9):1670–1689.
32. Mullahy, J. (2015). Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods*, 4(1):71–100.
33. Murteira, J. M. R. and Ramalho, J. J. S. (2016). Regression analysis of multivariate fractional data. *Econometric Reviews*, 35(4):515–552.
34. Österreicher, F. and Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653.
35. Rayens, W. S. and Srinivasan, C. (1994). Dependence properties of generalized Liouville distributions on the simplex. *Journal of the American Statistical Association*, 89(428):1465–1470.
36. Rozenas, A. (2015). *ocomposition: Regression for Rank-Indexed Compositional Data*. R package version 1.1.
37. Scealy, J. and Welsh, A. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society. Series B*, 73(3):351–375.
38. Scealy, J. and Welsh, A. (2014). Colours and cocktails: Compositional data analysis 2013 Lancaster lecture. *Australian & New Zealand Journal of Statistics*, 56(2):145–169.
39. Shi, P., Zhang, A., Li, H., et al. (2016). Regression analysis for microbiome compositional data. *Annals of Applied Statistics*, 10(2):1019–1040.
40. Shimizu, T. K. O. (2019). *Penalized regression methods for compositional data*. PhD thesis, Institute of

Mathematics and Computer Sciences.

41. Smith, R. L. (2002). A statistical assessment of Buchanan's vote in Palm Beach county. *Statistical Science*, 17(4):441–457.
42. Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69(1):197–203.
43. Stewart, C. and Field, C. (2011). Managing the Essential Zeros in Quantitative Fatty Acid Signature Analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(1):45–69.
44. Templ, M., Hron, K., and Filzmoser, P. (2011). *robCompositions: Robust estimation for compositional data*. R package version 0.8-4.
45. Tolosana-Delgado, R. and von Eynatten, H. (2009). Grain-size control on petrographic composition of sediments: compositional regression and rounded zeros. *Mathematical Geosciences*, 41(8):869.
46. Tsagris, M. (2015a). A novel, divergence based, regression for compositional data. In *Proceedings of the 28th Panhellenic Statistics Conference, April, Athens, Greece*.
47. Tsagris, M. (2015b). Regression analysis with compositional data containing zero values. *Chilean Journal of Statistics*, 6(2):47–57.
48. Tsagris, M., Athineou, G., Alenazi, A., and Adam, C. (2022). *Compositional: Compositional Data Analysis*. R package version 5.8.
49. Tsagris, M., Preston, S., and Wood, A. (2011). A data-based power transformation for compositional data. In *Proceedings of the 4rth Compositional Data Analysis Workshop, Girona, Spain*.
50. Tsagris, M. and Stewart, C. (2018). A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39(3):398–412.
51. Tsagris, M. and Stewart, C. (2020). A folded model for compositional data analysis. *Australian Journal of Statistics*, 62(2):249–277.
52. van den Boogaart, K., Tolosana-Delgado, R., and Bren, M. (2018). *compositions: Compositional Data Analysis*. R package version 1.40-2.
53. Wang, H., Liu, Q., Mok, H. M., Fu, L., and Tse, W. M. (2007). A hyperspherical transformation forecasting model for compositional data. *European Journal of Operational Research*, 179(2):459–468.
54. Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.