Regression Problems for Compositional Data

Guillaume Franchi

ENSAI, Bruz

7 February 2022

Outlines

Outlines

Definition

Compositional data consist of vectors whose components are the proportions of some whole.
Mathematically, those vectors are elements of the **simplex**

$$\mathcal{S}_{d-1} = \left\{ (x_1, \ldots, x_d) \in]0;1[^d \;\middle|\; \sum_{i=1}^{d} x_i = 1 \right\}.$$

Example *Arctic Lake*

In sedimentology, specimens of sediments are traditionally separated into three mutually exclusive and exhaustive constituents: sand, silt and clay.

The table below records some compositions of sediment samples at different water depths in an Arctic lake.

Sediment	Sand	Silt	Clay	Water Depth
1	0.78	0.20	0.03	10.40
2	0.72	0.25	0.03	11.70
3	0.51	0.36	0.13	12.80
4	0.52	0.41	0.07	13.00
5	0.70	0.26	0.04	15.70

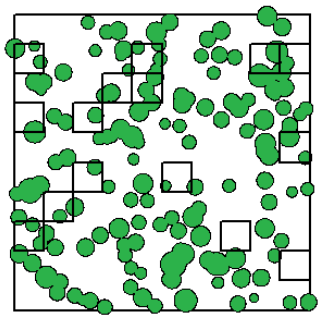
When $d = 2$, there is only one proportion to study, which leads to a regression problem for real values.However, the constraints due to the proportional nature of the values must be taken into account.

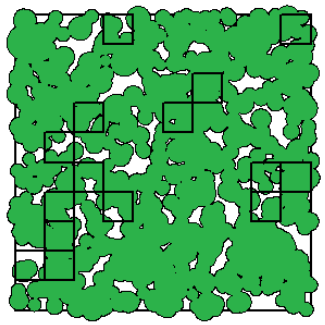
Example *Forest Cover*

Consider a forest represented here by a square of one hectare in area. We focus here on the ground cover of the forest explained by the annual rainfall.

Twenty forests are simulated, each with a mean annual precipitation (MAP) ranging from 125 to 2,500 mm per year. In practice, it is impossible to measure the ground cover of the whole forest.Thus for each forest, the percentage of ground cover is calculated for fifteen randomly positioned and non-overlapping quadrats of 10×10 m^2.This leads to 300 observations of ground cover percentage.

Example





(a) Ground cover of a forest with a MAP of 375 mm.

(b) Ground cover of a forest with a MAP of 2000 mm.

Outlines

Outlines

One can immediately guess why applying any classical regression model on the raw data is a bad idea for compositional problems.

It is indeed impossible to ensure that the constraints on the simplex will be satisfied for the predicted values.

In the following, we assume that we obtain a sample $y^{(1)}, \dots y^{(N)}$ of compositional data, which are the realizations of $Y^{(1)}, \dots, Y^{(N)}$ random variables valued in the simplex.

For all $1 \leqslant t \leqslant N$:

$$y^{(t)} = \left(y_1^{(t)}, \dots, y_d^{(t)} \right).$$

For each $1 \leqslant t \leqslant N$, we also consider K explanatory variables $x_{t,1}, \dots x_{t,K}$, valued in $\mathbb{R}$.

Outlines

The general idea here is to apply a one-to-one mapping

$$g : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^k$$

and to apply any classical regression method on the transformed data $z^{(1)} = g(y^{(1)}), \ldots, z^{(N)} = g(y^{(N)})$, with covariates $x_1, \ldots, x_K$.

Then the prediction $\hat{y}$ is obtained with $\hat{y} = g^{-1}(\hat{z})$ where $\hat{z}$ is the prediction on the space of transformed data.

In this presentation, we will consider **additive logratio**:

$$\begin{aligned} alr: \quad & \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^{d-1} \\ & y \longmapsto \left(\log\left(\frac{y_1}{y_d}\right), \ldots, \log\left(\frac{y_{d-1}}{y_d}\right) \right) \end{aligned}$$

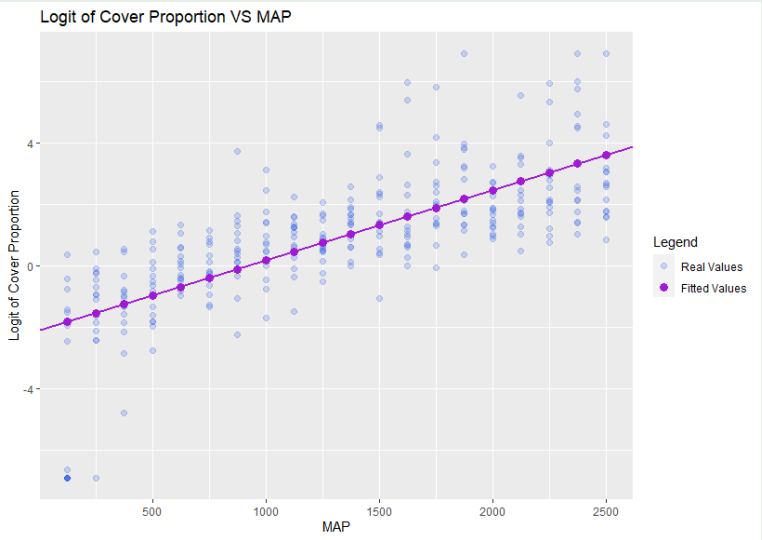
with inverse transformation:

$$alr^{-1}(z_1, \ldots, z_{d-1}) = \left(\frac{\exp(z_i)}{1 + \sum_{j=1}^{d-1} \exp(z_j)} \right)_{1 \leqslant 1 \leqslant d-1}.$$

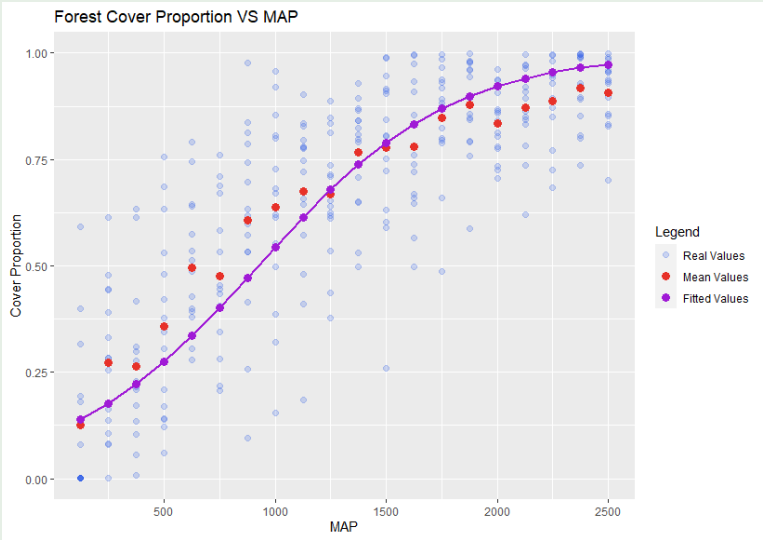
Note that when $d = 2$, i.e. we study one proportion, the alr transform is actually a **logit** transformation.

Example *Forest cover*

Here an ordinary least square regression is applied to the data transformed with a logit function.The fitted values are subsequently backtransformed to the original scale.



(a) Fitted values on the transformed scale.



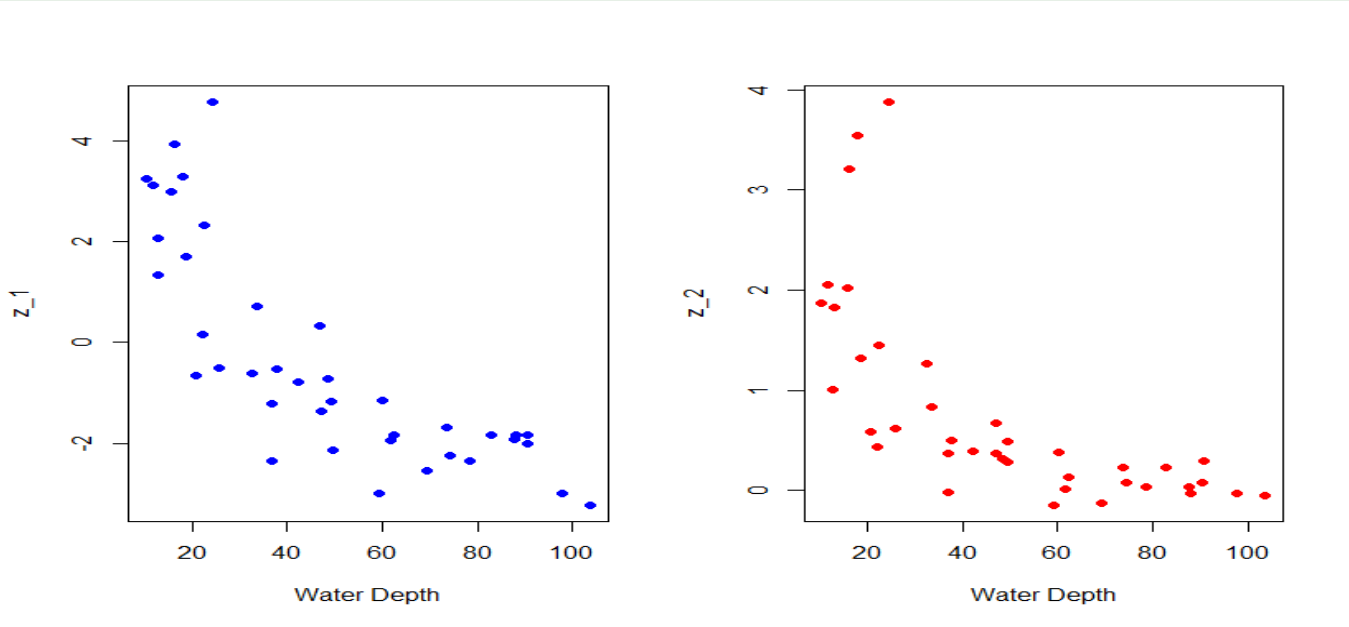
(b) Fitted values on the original scale.

Regression for the forest cover proportion with a logit transform

Example *Arctic Lake*

The question of interest is to predict the composition (Sand, Silt and Clay) of a sediment from the depth at which it was taken.

Here we apply the alr transformation to the data. The transformed data are vectors in $\mathbb{R}^2$ denoted $z^{(t)} = (z_1^{(t)}, z_2^{(t)})_{1 \leqslant t \leqslant 39}$.

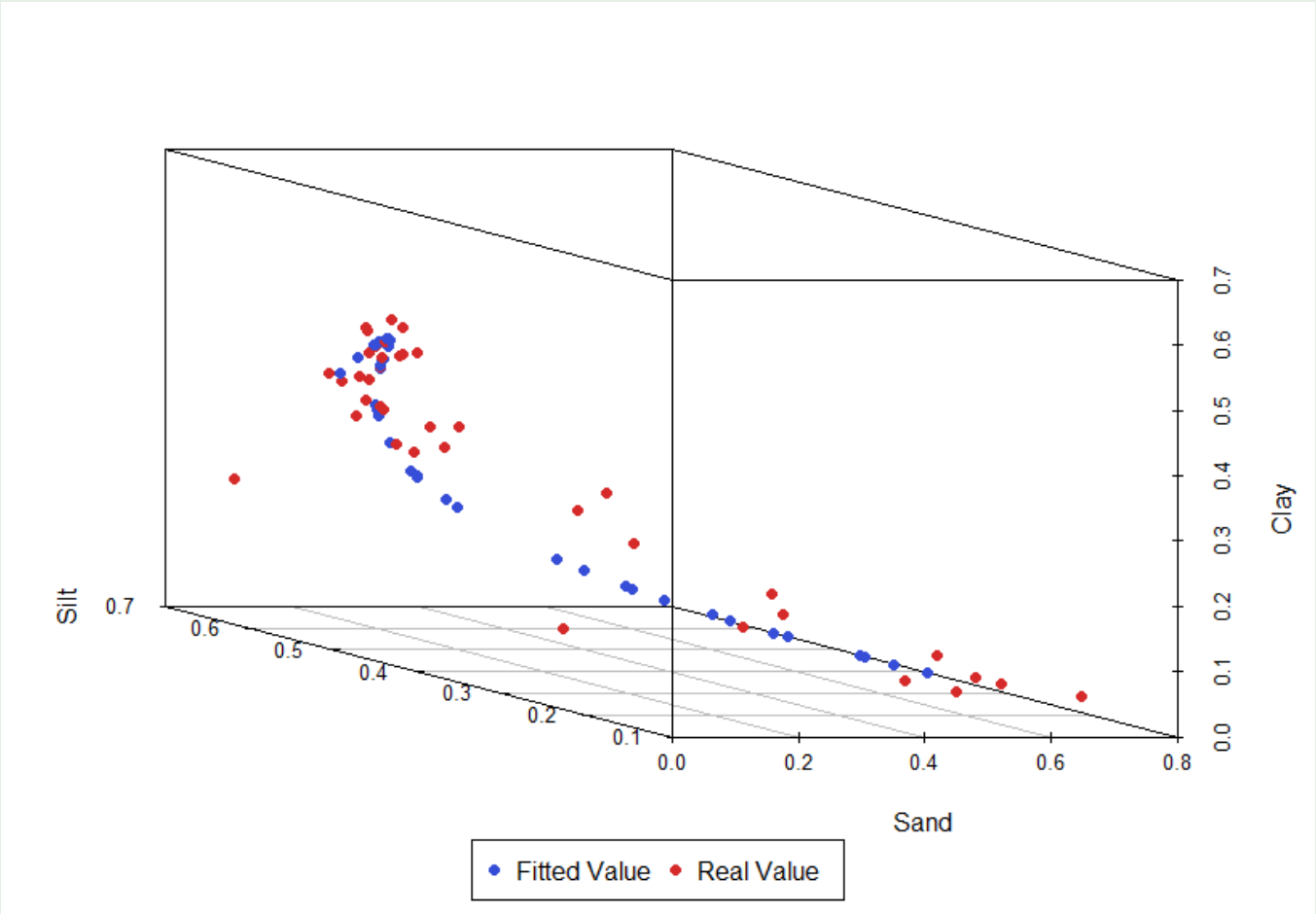


Scatterplots of the transformed data in the Arctic Lake Regression.

Example

Due to the shape of the scatterplots, we decide to model these data by a polynomial of order 2. We thus apply a non-linear least square regression on the transformed data.
We finally get back to the original scale by applying the inverse transform of the alr on the fitted values in the transformed scale.The figure below presents the estimated compositions in comparison to the true ones.

Example



Regression for the Arctic Lake Sediments compositions with an alr transformation.

The main issue of this technique is that, even if the estimates on the transformed scale are unbiased, it might not be the case on the original scale.

This issue arises due to Jensen's inequality. Since the back-transformation is, in general, not linear, the bias on the original scale might be greater than the one on the transformed scale, resulting sometimes in major discrepancies in the fitted values.

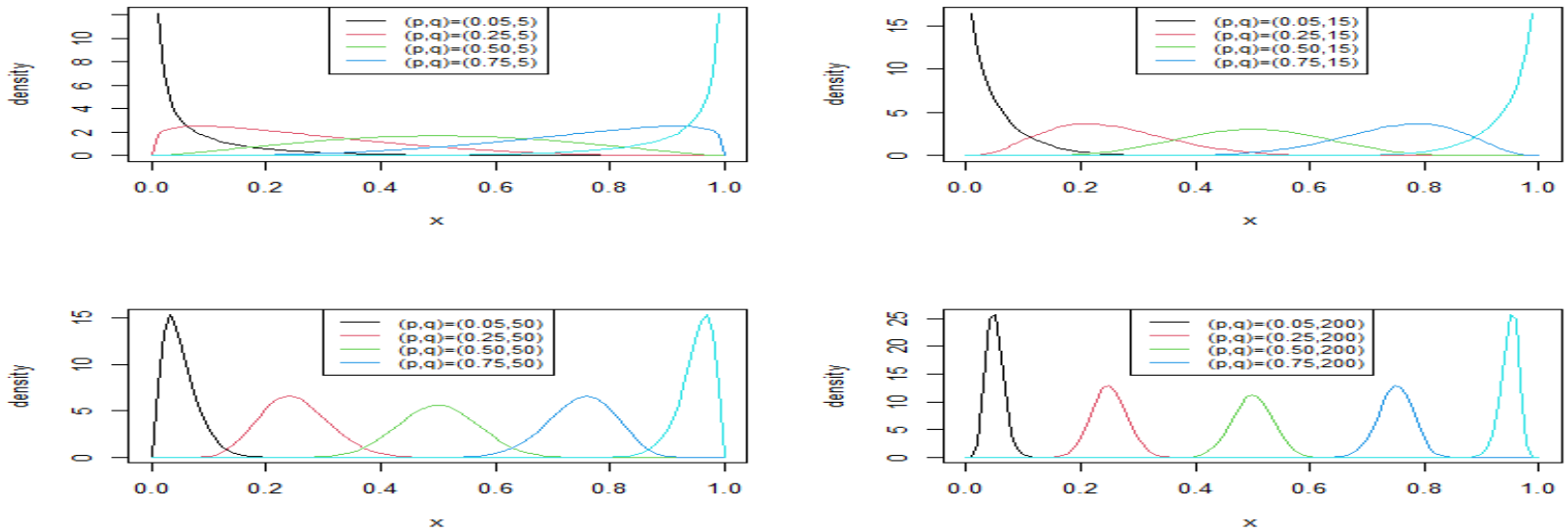
Outlines

The principle of this approach is to assume that each random variable $Y^{(t)}$ has a specific distribution supported by the simplex.

Outlines

We first consider the case $d = 2$, where there is actually only one proportion to study.

In Beta regression, we will assume that each random variable $Y^{(t)}$ folllows a Beta distribution $B(p_t, q_t)$.Varying values of p_t and q_t allows indeed to obtain very different densities, which makes the Beta distribution very flexible and well suited for modelization.



Beta densities for different combinations of (p, q).

Definition

The **Beta distribution** $B(p, q)$ is the distribution whose density is given by:

$$f_{(p,q)}(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \mathbb{1}_{[0;1]}(y) y^{p-1}(1-y)^{q-1}$$

where $\Gamma(\cdot)$ is the gamma function.

Remark

Recall that the Gamma function is defined on $]0;+\infty[$ by:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) \mathrm{d}t$$

and satisfies for all $x > 0$: $\Gamma(x+1) = x\Gamma(x)$.

The Beta function is defined for all $a, b > 0$ by:

$$\beta(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} \mathrm{d}t$$

and satisfies the relation:

$$\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Proposition

Let Y be a random variable with distribution $B(p, q)$.The mean and variance of Y are given by:

$$\mathbb{E}(Y) = \frac{p}{p+q}$$

and

$$\mathrm{Var}(Y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

For regression purpose, the parameter of interest is often the mean of the response variable.In Beta regression, we propose a different parametrization of the Beta density given previously, so that the model we build focuses on the mean of the response with a dispersion parameter.

Proposition

Let Y be a random variable with distribution $\beta(p, q)$ and denote:

$$\mu = \frac{p}{p+q} \quad \text{and} \quad \phi = p + q.$$

We have:

$$\mathbb{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\mu(1-\mu)}{1+\phi}$$

and the density of Y can be written with this new parametrization:

$$f_{(\mu,\phi)}(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\mathbb{1}_{]0;1[}(y)$$

where $0 < \mu < 1$ and $\phi > 0$.

Remark

We can interprete ϕ as a dispersion parameter, since $\mathrm{Var}(Y)$ decreases as ϕ increases.

The model we build in Beta regression is obtained by assuming that each independent variable $Y^{(t)}$ follows a Beta distribution with mean μ_t and unknown dispersion ϕ.
Furthermore, we assume that the mean μ_t satisfies the equation:

$$g(\mu_t) = \sum_{k=1}^{K} x_{t,k}\beta_k := \eta_t$$

where $\beta = (\beta_1, \ldots, \beta_k)'$ is a vector of unknown regression parameters of $\mathbb{R}^K$, $x_{t,1}, \ldots, x_{t,K}$ are observations of K covariates, and g is a **link function**, strictly monotonic and twice differentiable, mapping $]0;1[$ into $\mathbb{R}$.

Remark

There are of course many possibilities in the choice of the link function g.Among the most popular is the logit function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

In this case, we can write:

$$\mu_t = \frac{\exp(x_t' \cdot \beta)}{1 + \exp(x_t' \cdot \beta)}$$

where $x_t' = (x_{t,1}, \ldots, x_{t,K})$, so that the regression parameters have an important interpretation.

Remark

Suppose that the value of the k^{th} regressor is modified by c units, and all other independent variables remain unchanged.Let ν_t denote the mean of $Y^{(t)}$ under the new covariate values, whereas μ_t denotes the mean of $Y^{(t)}$ under the original covariate values. One can show that:

$$\exp(c\beta_k) = \frac{\nu_t/(1-\nu_t)}{\mu_t/(1-\mu_t)},$$

meaning that $\exp(c\beta_k)$ equals the odds ratio.

The regression parameter $\beta = (\beta_1, \ldots, \beta_K)$, as well as the dispersion parameter ϕ are estimated with their maximum likelihood estimators $\hat{\beta}$ and $\hat{\phi}$.

Remark

Note that $\hat{\beta}$ and $\hat{\phi}$ do not have a closed form, hence they are obtained by numerical optimization algorithms such as a Newton algorithm.

Example *Forest cover*

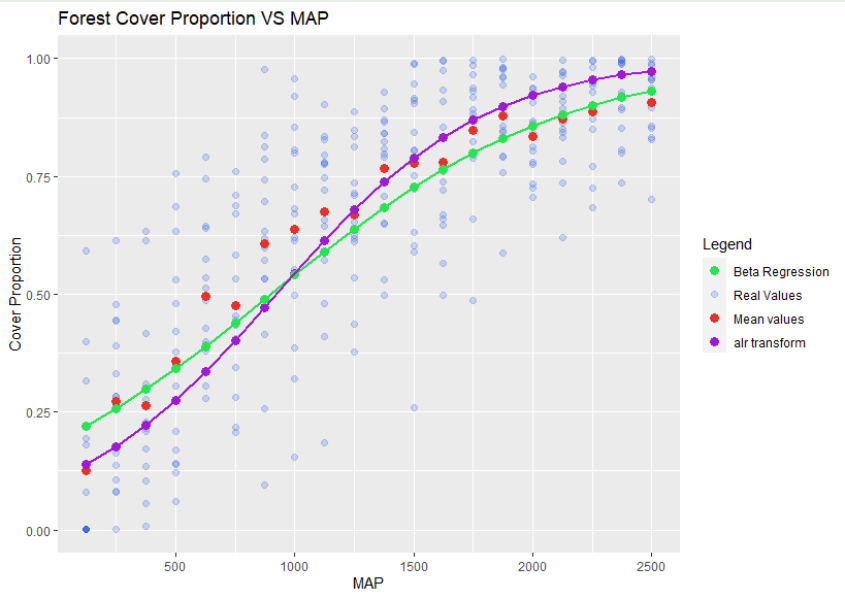
This time, we assume that the response variable $Y^{(t)}$, i.e. the proportion of ground cover follows a beta distribution with mean μ_t satisfying:

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta x_t + \alpha$$

where x_t is the mean annual precipitation, in mm per year. In our model, an intercept is added in the covariates.
The maximum likelihood estimators obtained are
$\hat{\alpha} = -1,4697$, $\hat{\beta} = 0,0016$ and $\hat{\phi} = 4,5377$.

Example

We can visualize below the fitted values with Beta regression, compared to the ones obtained by data transformation.



Comparison of regression models on the Forest Cover example.

Example

In this example, the goodness of fit for both models is quite similar, as the MSE suggest. However, the Beta regression model seems to fit slightly better the observed mean cover proportion.

	Data Transformation	Beta Regression
MSE	0.0054	0.0035

MSE of the different models in the Forest Cover example.

Outlines

We now consider the multivariate case where there are more than two proportions to study: $d \geqslant 3$.A generalization of the Beta distribution is the Dirichlet distribution.

Definition

The **Dirichlet distribution** $\mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$ is the distribution $\mu_\alpha^{(d)}$ such that for any Borel set in $\mathbb{R}^d$ we have:

$$\mu_\alpha^{(d)}(A) = \int \cdots \int \mathbb{1}_A(x_1, \ldots, x_d) \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_d)} \prod_{i=1}^d x_i^{\alpha_i - 1} \mathbb{1}_{B_{d-1}}(x_1, \ldots, x_{d-1}) \, \delta_{\left\{1 - \sum_{i=1}^{d-1} x_i\right\}}(\mathrm{d}x_d) \mathrm{d}x_1 \ldots \mathrm{d}x_{d-1}$$

where $B_{d-1} = \left\{(x_1, \ldots, x_{d-1}) \in \mathbb{R}_+^{d-1} \;\middle|\; 0 < \sum_{i=1}^{d-1} x_i < 1\right\}$.

The Dirichlet distribution is indeed a generalization of the Beta distribution.

Proposition

Let $Y = (Y_1, \ldots, Y_d)$ a random vector with distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$, and let $\phi = \sum_{i=1}^{d} \alpha_i$.

Then, for all $1 \leqslant i \leqslant d$, we have $X_i \sim \beta(\alpha_i, \phi - \alpha_i)$.

In particular $\mathbb{E}(X_i) = \dfrac{\alpha_i}{\phi}$ and $\mathrm{Var}(X_i) = \dfrac{\alpha_i(\phi - \alpha_i)}{\phi^2(\phi + 1)}$.

By analogy with the Beta regression, our parameter of interest will be the mean vector of the response variable, along with a dispersion parameter.Thus, we propose a different parametrization of the Dirichlet distribution to fulfill those requirements.

Proposition

Let $Y = (Y_1, \ldots, Y_d)$ be a random variable with distribution $\mathrm{Dir}(\alpha_1, \ldots \alpha_d)$ and denote:

$$\mu = (\mu_1, \ldots \mu_d) = \left(\frac{\alpha_1}{\phi}, \ldots, \frac{\alpha_d}{\phi} \right) \quad \text{and} \quad \phi = \sum_{i=1}^{d} \alpha_i.$$

We have:

$$\mathbb{E}(Y) = \mu \quad \text{and} \quad \mathrm{Var}(Y_i) = \frac{\mu_i(-\mu_i)}{\phi + 1}$$

Proposition

The density of Y can be written with this new parametrization:

$$f_{(\mu,\phi)}(y_1, \dots y_d) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu_1)\dots\Gamma(\phi\mu_d)} \prod_{i=1}^{d} y_i^{\phi\mu_i - 1} \mathbb{1}_{B_{d-1}}(y_1, \dots, y_{d-1})$$

where $0 < \mu_i < 1$ for all $1 \leqslant i \leqslant d$ and $\phi > 0$.

Following the steps of the Beta regression, the model we build in Dirichlet regression is obtained by assuming that each independent variable $Y^{(t)}$ follows a Dirichlet distribution, with mean μ_t and unknown dispersion ϕ.

Furthermore, by analogy with the multinomial regression, we assume that the mean μ_t satisfies the equations:

$$\forall i \in \{1, \ldots, d-1\}, \ \mu_{t,i} = \frac{\exp(\beta^{(i)} \cdot x_t)}{1 + \sum_{j=1}^{d-1} \exp(\beta^{(j)} \cdot x_t)}$$

and

$$\mu_{t,d} = \frac{1}{1 + \sum_{j=1}^{d-1} \exp(\beta^{(j)} \cdot x_t)}$$

where the $\beta^{(i)}$'s are vectors of unknown regression parameters of $\mathbb{R}^K$ and x_t is the observation of K covariates.

Remark

Similarly to the Beta regression, the vectors $\beta^{(i)}$ have an interpretation in terms of odds ratio.

The regression vectors $\beta^{(i)}$ and the dispersion parameter ϕ are estimated with their maximum likelihood estimators $\widehat{\beta^{(i)}}$ and $\hat{\phi}$.

Remark

Once again, $\widehat{\beta^{(i)}}$, $1 \leqslant i \leqslant d$ and $\hat{\phi}$ have no closed form, hence they are obtained by a numerical optimization algorithm.

Example *Arctic Lake*

We assume that the composition $Y^{(t)}$ of the t-th sediment follows a Dirichlet distribution with mean μ_t satisfying for $i \in \{1, 2\}$:

$$\mu_{t,i} = \frac{\exp(\alpha^{(i)} + \beta_1^{(i)} x_t + \beta_2^{(i)} x_t^2)}{1 + \sum_{j=1}^{2} \exp(\alpha^{(j)} + \beta_1^{(j)} x_t + \beta_2^{(j)} x_t^2)}$$

and :

$$\mu_{t,3} = \frac{1}{1 + \sum_{j=1}^{2} \exp(\alpha^{(j)} + \beta_1^{(j)} x_t + \beta_2^{(j)} x_t^2)}$$

where x_t is the water depth of the sediment, in meters.
Here, we actually consider two covariates in addition to an intercept: the water depth and the square of the water depth.

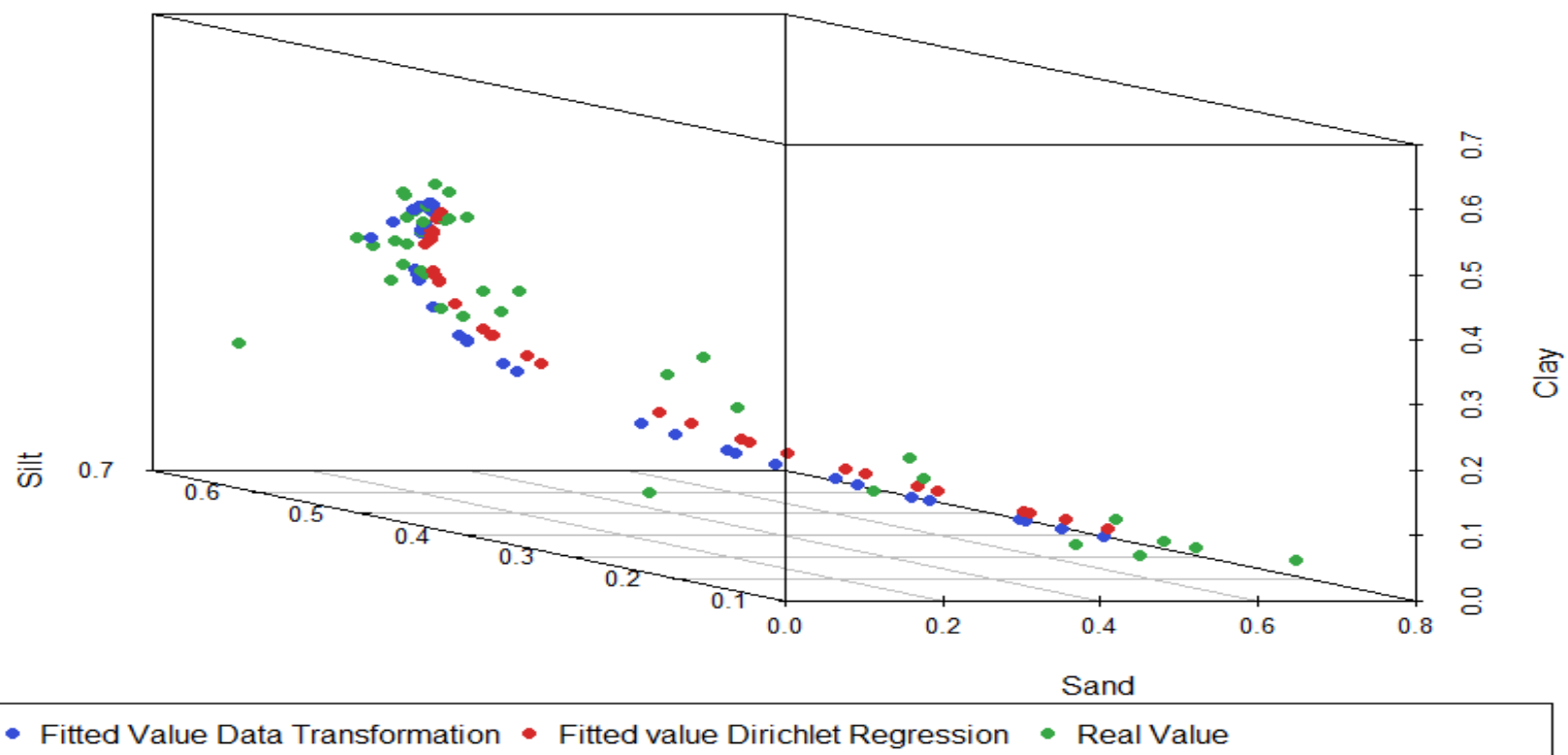
Example

The maximum likelihood estimators obtained are given in the table below.

$\widehat{\alpha^{(1)}}$	$\widehat{\alpha^{(2)}}$	$\widehat{\beta_1^{(1)}}$	$\widehat{\beta_2^{(1)}}$	$\widehat{\beta_1^{(2)}}$	$\widehat{\beta_2^{(2)}}$	$\hat{\phi}$
4,1566	2,4091	-0,1552	0,0010	-0,0602	0,0040	19,0410

Example

We can visualize below the fitted values with Dirichlet regression compared to the ones obtained by data transformation.



Example

We also propose in the table below the values of the MSE for the two regression models.

	Data Transformation	Dirichlet Regression
MSE	0.0254	0.0244

Outlines

In both our models, we assume that each response variable Y_j follows a Beta distribution.In the case of the Dirichlet model:

$$Y_j \sim B(\alpha_j, \phi - \alpha_j).$$

If the model is correct, then by denoting F_j the cumulative distribution function of $B(\alpha_j, \phi - \alpha_j)$, we have:

$$F_j(Y_j) \sim \mathcal{U}(]0;1[).$$

and:

$$R_j := \Phi^{-1}\left(F_j(Y_j)\right) \sim \mathcal{N}(0,1)$$

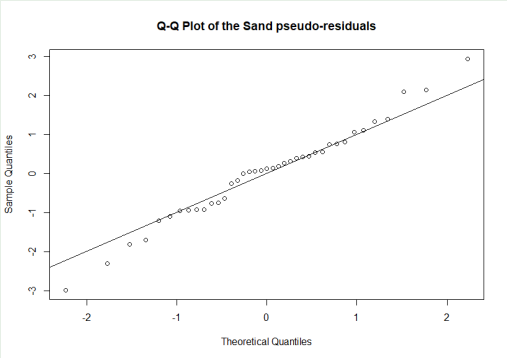
where Φ^{-1} is the inverse cumulative distribution function of a standard gaussian distribution.

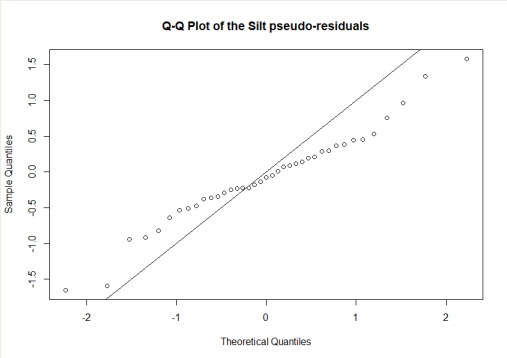
Definition

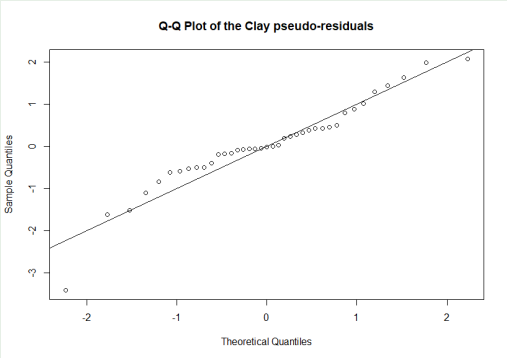
The random variables R_j are called the **pseudo-residuals** of the regression.

Example *Arctic Lake*

For the Arctic Lake example, we plotted below the normal QQ-plots of the pseudo-residuals.Those residuals are calculated with the estimates obtained previously.







Normal QQ-Plots of the pseudo residuals in the Arctic Lake Regression.

Outlines

With the transformation of the data, we already mentioned that issues can arise due to Jensen's inequality.

Furthermore, interpretation of the coefficient estimates is difficult, since the estimation is done on the transformed scale.

On the other hand, interpretation is very easy with the Beta or Dirichlet regression.

However, the Dirichlet distribution implies a very strong structure of independence in the compositions.

Proposition

Let $\alpha_1, \ldots, \alpha_d$ be positive real numbers, and $U_1, \ldots, U_d$ be independent random variables with gamma distributions:

$$U_i \sim \gamma(\alpha_i, 1).$$

Then, if we denote $V = \sum_{i=1}^{d} U_i$, we have:

$$\left(\frac{U_1}{V}, \ldots, \frac{U_d}{V} \right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d).$$

Remark

Let us recall that a random variable U follows the gamma distribution $\gamma(a, p)$ if it admits the density:

$$f(u) = \mathbb{1}_{\mathbb{R}_+}(u) \frac{p^a}{\Gamma(a)} \mathrm{e}^{-pu} u^{a-1}.$$