Dirichlet Regression in R

the DirichletReg package

Marco Maier

WU Vienna

25. Februar 2011

1 Compositional Data ...

are composed of a set of variables whose contents are in a certain interval and sum up to a constant for each observation, e.g. the composition of the sediments in a lake which could be partitioned in sand, silt, and clay:

obs.	sand	silt	clay	$\sum$
1	.50	.25	.25	1
2	.10	.40	.50	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i3}	y_{i3}	y_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Because of the constraint, any variable can be omitted and represented by $y_j = 1 - \sum_{i \neq j} y_i$.

Compositional data reflect – as the name suggests – the 'compositional structure' of something across all variables. It can be applied in fields as diverse as medicine (toxins etc. in blood samples), geology, psychology, ...

As the beta distribution is the continuous version of the binomial dist., the Dirichlet dist. is a continuous multinomial distribution. This allows for nominal items without coercing respondents to select only one category, e.g.:

Which party would you vote for?

	Grüne	SPÖ	ÖVP	FPÖ
multinomial	0	1	0	0
Dirichlet	.45	.50	.05	0

If the 'probability' of answering in a certain cateogory is spread across the choices, a Dirichlet approach is more informative.

This package aims at implementing a Dirichlet-regression using two different parameterizations along with a strong focus on graphical representation of the data and models, model tests and model selection.

2 The Dirichlet Distribution

The Dirichlet distribution is a generalization of the beta dist. for more than 2 variables (of which one is usually omitted, because it is redundant; $y_1 = 1 - y_2$ and vice versa). These k variables have to lie in the interval $(0, 1)$ and sum up to 1 for each observation.

$$f(\boldsymbol{y}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k} y_i^{\alpha_i - 1} \tag{1}$$

Normalization is provided by $B(\boldsymbol{\alpha})$, the multinomial beta-function, which can be expressed as:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} \tag{2}$$

Each component is governed by a shape parameter $\alpha > 0$ which are in and of itself not very informative. Their sum $\alpha_0 = \sum_i \alpha_i$ can be interpreted as a 'precision parameter'.

With this precision parameter, we can calculate the means

$$\mathrm{E}(y_i) = \frac{\alpha_i}{\alpha_0}$$

and also the variances

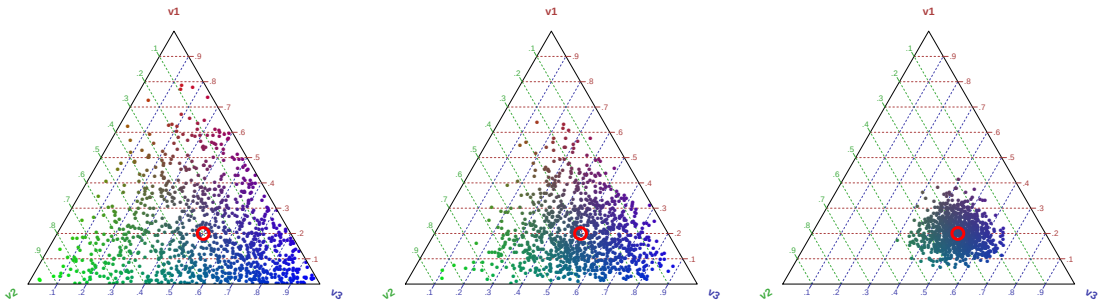
$$\mathrm{VAR}(y_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

and covariances of the variables

$$\mathrm{COV}(y_i, y_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}; \quad i \neq j$$

3 Data in the Simplex

Because one variable can always be represented as the difference between the constant and the sum of the other variables, the data lose a degree of freedom. Practically, this means that if we have k variables, the data lie on a $k-1$-dimensional simplex. With 3 variables we can do a so-called 'ternary plot', i.e. the data lie on a triangle.



These data all have an expected value of (.2, .3, .5) but with precisions of 5, 10 and 50 so the alphas are (1, 1.5, 2.5), (2, 3, 5) and (10, 15, 25).

4 Regression Models – Parameterization 1

In, what I call the 'common parameterization', we try to predict the alphas for each component by a set of variables. Because α must be greater than 0, we can conveniently use a log-link for this parameterization. So for each component y_i there is a vector of regression coefficients $\boldsymbol{\beta}$ along with an appropriate design matrix $\boldsymbol{X}$.

$$\log\left(\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_k \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{\beta}_1 \\ \boldsymbol{X}_2\boldsymbol{\beta}_2 \\ \boldsymbol{X}_3\boldsymbol{\beta}_3 \\ \vdots \\ \boldsymbol{X}_k\boldsymbol{\beta}_k \end{bmatrix} \tag{3}$$

Because all α parameters are modeled individually, heteroskedasticity is accounted for implicitly.

5 Regression Models – Parameterization 2

If we want a kind of mean/dispersion model, we can take an approach as in `betareg` where $\alpha = \mu\phi$. The precision paramter ϕ can be predicted using a log-link, for example.

For the means we have to make sure that they always sum up to 1, so a strategy as in multinomial regression models is employed.

$$\mu_c = \frac{\exp(\boldsymbol{X}^{[c]}\boldsymbol{\beta}^{[c]})}{1 + \sum \exp(\boldsymbol{X}\boldsymbol{\beta}_j)} \qquad c \neq b \tag{4}$$

$$\mu_b = \frac{1}{1 + \sum \exp(\boldsymbol{X}\boldsymbol{\beta}_j)} \tag{5}$$

6 Pros and Cons

The common parameterization is more flexible, especially concerning model selection whereas the reparameterization might be more appealing to practitioners due to the interpretability as in multinomial logistic regression.

7 Model Specification

Depending on the parameterization there are two ways of setting up the model formulae. All dependent variables are first prepared by `DR.data(y1,y2,y3)` (this normalizes and transforms the data if necessary).

For the common parameterization we have

```
DirichReg(DV ~ x1 * x2, data = some.data)
```

or

```
DirichReg(DV ~ x1 + x2 | x1 * x2 | x1, data = some.data)
```

The reparameterization contains only one set of predictors for the means and one for the precision.

```
DirichReg(DV ~ x1 * x2 | phi ~ x1 + x2, data = some.data)
```

8 Estimation

The log-likelihood functions have been adapted and simplified for both parameterizations and the gradient-vectors were derived analytically to improve and speed up optimization.

As of now, the BFGS algorithm as implemented in `optim` is used for optimization. To compute the parameters' standard errors, the Hessian resulting from the optimization process is used.

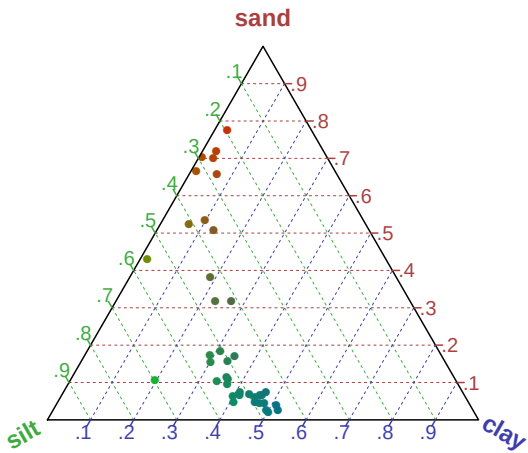
9 Model Selection

Regardless of the parameterization, an `anova` function is implemented to compare and select models.

In the long run, an iterative algorithm for model selection would be 'nice to have', probably involving selection strategies as in graphical models. This would be especially interesting for the common parameterization, because each component is modeled by a completely independent set of predictors.

10 Example – Arctic Lake

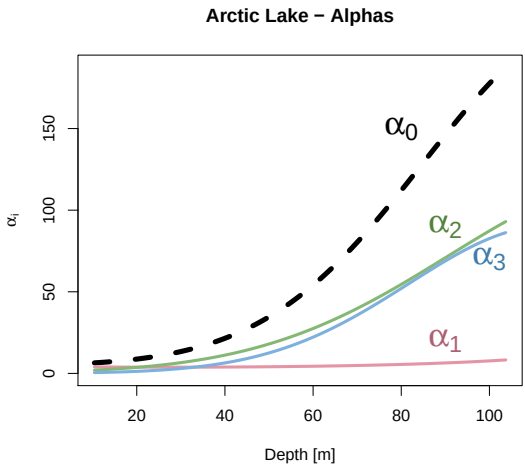
The ground composition of an arctic lake was partitioned into sand, silt, and clay. We want to find out, if the composition can be predicted by the depth. First a ternary plot:

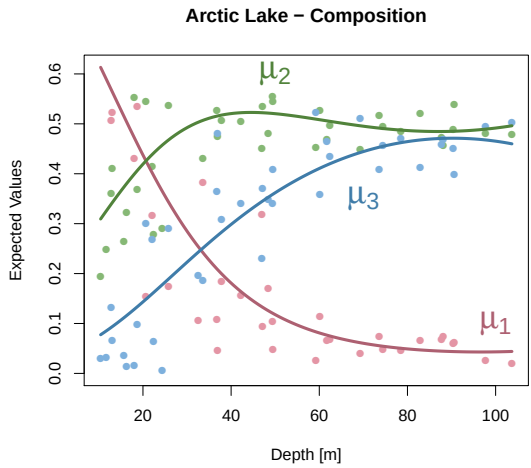


Fitting a model in R:

```
> AL <- DR.data(ArcticLake[,1:3])
> res <- DirichReg(AL ~ depth + I(depth^2), ArcticLake)
> summary(res)

Call:
DirichReg(formula = AL ~ depth + I(depth^2), data = ArcticLake)

RESIDUALS WILL BE IMPLEMENTED SOON! :)

--------------------------------------------------------------------------------
Coefficients for variable no. 1: sand
             Estimate Std. Error z-Value p-Value
(Intercept)  1.4361854  0.8022580    1.79  0.0734 .
depth       -0.0072376  0.0329250   -0.22  0.8260
I(depth^2)   0.0001324  0.0002760    0.48  0.6314
--------------------------------------------------------------------------------
Coefficients for variable no. 2: silt
             Estimate Std. Error z-Value p-Value
(Intercept) -0.0259884  0.7595826  -0.034  0.9727
depth        0.0717460  0.0342953   2.092  0.0364 *
I(depth^2)  -0.0002679  0.0003088  -0.868  0.3856
--------------------------------------------------------------------------------
Coefficients for variable no. 3: clay
             Estimate Std. Error z-Value p-Value
(Intercept) -1.7931592  0.7360825  -2.436 0.01485 *
depth        0.1107914  0.0357608   3.098 0.00195 **
I(depth^2)  -0.0004872  0.0003307  -1.473 0.14074
--------------------------------------------------------------------------------
Signif. codes: '***' < .001, '**' < 0.01, '*' < 0.05, '.' < 0.1

Log-likelihood: 81.96 on 9 df (30 iterations)
Link: Log
Parameterization: common
```

Graphics and Interpretation:





Apart from the depth-related changing composition, we can see from α_0 that the precision increases with depth.

11 To do & Conclusion

- Full implementation of the alternative paramterization.
- Good starting values.
- Various residuals.
- Generic plotting routines.

- Allows for the collection of multinomial data in an uncommon and potentially more informative way.
- Applicable in many fields.
- User-friendly modeling and presentation of results.

Thank you!

marco.maier@wu.ac.at

http://r-forge.r-project.org/projects/dirichletreg/