The Simplicial Generalized Beta distribution - R-package SGB and applications

Monique Graf
Université de Neuchâtel, Switzerland; *monique.p.n.graf@bluewin.ch*
Elpacos Statistics, La Neuveville, Switzerland

Summary

A generalization of the Dirichlet and the scaled Dirichlet distributions is given by the simplicial generalized Beta, SGB (Graf, 2017). In the Dirichlet and the scaled Dirichlet distributions, the shape parameters are modeled with auxiliary variables (Maier, 2015, R-package **DirichletReg**) and Monti et al. (2011), respectively. On the other hand, in the ordinary logistic normal regression, it is the scale composition that is made dependent on auxiliary variables. The modeling of scales seems easier to interpret than the modeling of shapes. Thus in the SGB regression:

- The *scale compositions* are modeled in the same way as for the logistic normal regression, i.e. each auxiliary variable generates $D-1$ parameters, where D is the number of parts.
- The *D Dirichlet shape parameters*, one for each part in the compositions, are estimated as well.
- An additional *overall shape parameter* is introduced in the SGB that proves to have important properties in relation with non essential zeros.
- Use of survey weights is an option.
- Imputation of missing parts is possible.

An application to the United Kingdom Time Use Survey (Gershuny and Sullivan, 2017) shows the power of the method. The R-package **SGB** (Graf, 2019) makes the method accessible to users. See the package vignette for more information and examples.

Key words: Dirichlet distribution, simplicial Generalized Beta, maximum likelihood estimation, imputation, R-package, Time Use survey.

1 SGB distribution

The Dirichlet distribution can be viewed as the distribution of $\mathbf{U} = \mathcal{C}(\mathbf{Y})$, where $\mathbf{Y} = (Y_j)_{j=1,...,D}$ is a vector of independent $Gamma(p_j)$ components and $\mathcal{C}(.)$ is the closure operation (i.e. $U_j = Y_j / \sum_{i=1}^{D} Y_i$). The SGB distribution follows the same construction, with the Gamma distribution replaced by the generalized Gamma, that is the underlying Y_i are independent $GG(a, c\, b_j, p_j)$, c being an arbitrary positive constant. The parameters are all positive and $\mathbf{b} = (b_1, ..., b_D)$ is itself a composition, the *scale composition*. The SGB can also be generated from the Dirichlet:

Definition *Suppose that* $\mathbf{Z} = (Z_1, ..., Z_D)$ *follows a Dirichlet*$(p_1, ..., p_D)$ *distribution. Then the random composition* $\mathbf{U} = (U_1, ..., U_D)$, $(D \geq 2)$, *given by*

$$U_j = \frac{b_j\, Z_j^{1/a}}{\sum_{i=1}^{D} b_i\, Z_i^{1/a}}, \quad j = 1, ..., D \quad or \quad \mathbf{U} = \mathcal{C}[\mathbf{b}\, \mathbf{Z}^{1/a}]$$

follows a $SGB(a, \{b_j, p_j, j = 1, ..., D\})$ *distribution.*

All parameters are positive; a *is an overall shape parameter,* $\mathbf{b} = (b_1, ..., b_D)$ *a scale composition and* $\mathbf{p} = (p_1, ..., p_D)$ *the vector of Dirichlet shape parameters.*

Conversely, the random composition $\mathbf{Z}$ can be written in function of $\mathbf{U}$,

$$Z_j = \frac{(U_j/b_j)^a}{\sum_{i=1}^{D}(U_i/b_i)^a}, \quad j = 1, ..., D \quad \text{or} \quad \mathbf{Z} = \mathcal{C}[(\mathbf{U}/\mathbf{b})^a]. \tag{1}$$

Because $U_D = 1 - \sum_{j=1}^{D-1} U_j$, there are only $D-1$ variables in the composition $\mathbf{U}$. The L_a-norm of the vector $(\mathbf{u}/\mathbf{b})$ is

$$\|\mathbf{u}/\mathbf{b}\|_a = \left[\sum_{k=1}^{D-1}(u_k/b_k)^a + \left((1 - \sum_{j=1}^{D-1} u_j)/b_D\right)^a\right]^{1/a}.$$

The probability density of the $SGB(a, \{b_j, p_j, j = 1, ..., D\})$ distribution is obtained as

$$f_{\mathbf{U}}(\mathbf{u}_{-D}) = \frac{\Gamma(P)a^{D-1}}{\prod_{j=1}^{D}\Gamma(p_j)} \prod_{k=1}^{D-1}\left\{\frac{u_k/b_k}{\|\mathbf{u}/\mathbf{b}\|_a}\right\}^{ap_k} \left\{\frac{\left(1 - \sum_{j=1}^{D-1} u_j\right)/b_D}{\|\mathbf{u}/\mathbf{b}\|_a}\right\}^{ap_D} \frac{1}{\prod_{k=1}^{D-1} u_k \left(1 - \sum_{j=1}^{D-1} u_j\right)},$$

$$u_k > 0,\ k = 1, ..., D-1, \qquad 1 - \sum_{j=1}^{D-1} u_j > 0.$$

Craiu and Craiu (1969) derived this density. The fitted compositions are defined as the estimated value of the so called Aitchison's expectation $\mathrm{E}_A(\mathbf{U}) = \mathcal{C}[\exp(\mathrm{E}\log(\mathbf{U}))]$, i.e. the image in the simplex of the expectation at log-scale, that is for the SGB, with $\psi(.)$ the digamma function,

$$\mathrm{E}_A(U_k) = \frac{b_k \exp\{\psi(p_k)/a\}}{\sum_{j=1}^{D} b_j \exp\{\psi(p_j)/a\}} \quad k = 1, ..., D.$$

In the R-package **SGB** regression models can be set up for the scale composition $\mathbf{b}$. The shape parameters a and $\mathbf{p}$ are estimated as well, but are supposed constant across compositions.

2 SGB regression model

2.1 Model

The SGB regression models follow the principles of log-ratio analysis advocated by Aitchison (1986). We define a general $D \times (D-1)$ contrast matrix $\mathbf{V}$, such that

$$\mathbf{1}_D^t \mathbf{V} = \mathbf{0}_{D-1}^t,$$

where $\mathbf{1}_D$ is a D-vector of ones and $\mathbf{0}_{D-1}$ is a $(D-1)$-vector of zeros. The model for scales is the general linear model. Let $\mathbf{X}$ be a $n \times p$ matrix of explanatory variables, where n is the sample size. Let $\mathbf{u}_i, i = 1, ..., n$ be the composition associated to $\mathbf{x}_i^t$, the i-th row of $\mathbf{X}$. Then the scales are modeled by

$$\log(\mathbf{b}_i^t)\mathbf{V} = \mathbf{x}_i^t \mathbf{B}, \tag{2}$$

where

$$\mathbf{B} = \left(\boldsymbol{\beta}_1 ... \boldsymbol{\beta}_{D-1}\right)$$

is the $p \times (D-1)$ - matrix of regression parameters for the log-ratio transforms, i.e. the $(D-1)$ columns of $\log(\mathbf{u}_i^t)\,\mathbf{V}$, $i = 1, ..., n$.

2.2 Fitting procedure

There is the possibility to introduce sampling weights into the procedure. These weights $w_i, i = 1, ..., n$ are scaled to sum to n.

The pseudo-log-likelihood is the weighted version of the log-likelihood and is given by

$$\begin{aligned} & \ell\left(a, (b_1, p_1), ..., (b_D, p_D) | \mathbf{u}_{i,-D}, i = 1, ..., n\right) \\ = \; & n\left[(D-1)\log(a) + \log\Gamma(P) - \sum_{k=1}^{D}\log\Gamma(p_k)\right] + \sum_{i=1}^{n} w_i \sum_{k=1}^{D} p_k \log z_k(\mathbf{u}_i) \\ & - \text{ terms not depending on parameters.} \end{aligned}$$

with $z(\mathbf{u}_i) = (z_1(\mathbf{u}_i), ..., z_D(\mathbf{u}_i))$ given at Equation (1) and $P = \sum_{j=1}^{D} p_j$.

The model is estimated by maximizing the pseudo-log-likelihood using a constrained optimization method, the augmented Lagrangian, see e.g. Madsen et al. (2004), and implemented in the `R`-package **alabama** as function `auglag` (Varadhan, 2015). The gradient is computed analytically and the Hessian numerically. The default constraints are

$$\begin{aligned} a &> 0.1 \quad \text{(to avoid numerical problems)} \\ p_j &> 0, \quad j = 1, ..., D \\ a\,p_j &> \texttt{bound}, \quad \text{by default, } \texttt{bound} = 2.1. \end{aligned}$$

Moments of ratios of parts following the SGB distribution only exist up to $(a\,p_j)$. Thus `bound` $= 2.1$ guarantees the existence of variances of all ratios of parts. Notice that the most important variables, the log-ratios of parts, possess moment of all orders.

A very handy feature of `alabama::auglag` is that the initial values do not need to satisfy the constraints, and that general (twice derivable) constraints on parameters can be introduced. The price to pay is the speed.

3 United Kingdom time use survey 2014-2015

3.1 Data-set

The time diary files of the United Kingdom time use survey 2014-2015 (Gershuny and Sullivan, 2017) provide activities and corresponding level of enjoyment reported over 24 hour period (from 4am to 4am) on business days and week end days. A household file and a individual file contain data collected during the household (individual) interviews. Extrapolation weights at the individual and day levels are given.

In the following application, the total time spent doing an activity during a 24 hour period (`eptime`) was extracted from the diaries. The activities are recorded by 10 min time span. Only the primary activity is considered here. In order to avoid too many zeroes, activities were grouped into 8 categories:

- `y0` Personal care,
- `y12` Employment grouped with study,
- `y34` Household and family care grouped with voluntary work,
- `y5` Social life and entertainment,
- `y6` Sports and outdoor activities,
- `y7` Hobbies, games and computing,
- `y8` Mass media,
- `y9` Travel and unspecified time use.

We consider here the 3,393 (out of 13,603) person-days with zero time spent on `y12`, `y6` and `y7`. For other activities not done that day, the rounded zero technique was used.

3.2 Analysis of one group

The explanatory variables are a weekend indicator; enjoyment data (levels 1 to 7, zero if missing); indicators of missing response on enjoyment; an indicator of "in employment" (`dilodefr=1`); `DVAge` age; `DMSex=2` an indicator of "woman"; indicators of missing `y34`, `y5`, `y8`, `y9`. 16 cases with missing `dilodefr` were deleted. The file with explanatory variables is then `dnot1267b`. The corresponding compositions are in `unot1267b`. The weights are given by `wnot1267b`. The log-ratio transform is alr with reference part `y0` which is never missing. It is specified by the matrix $\mathbf{V}$ = `Vmat2` (see Equation 2). The alr transforms are denoted `a34`, `a5`, `a8` and `a9`. The regression model is specified in the `Formula` `Fnot1267b`, following the syntax of Zeileis and Croissant (2010).

```
Fnot1267b <- Formula(a34 | a5 | a8 | a9 ~
                weekend + enjoy0 +  enjoy34t + enjoy5t +  enjoy8t + enjoy9t +
                I(is.na(enjoy34)) +  I(is.na(enjoy5)) + I(is.na(enjoy8)) +
                I(is.na(enjoy9)) + DVAge + I(DMSex==2) + I(dilodefr==1) +
                ymiss34 + ymiss5  + ymiss8 + ymiss9 )

regnot1267b <- regSGB(Fnot1267b, data = list(dnot1267b, unot1267b, Vmat2),
                weight = wnot1267b, bound = 1.7, shape10 = 0.15,
                control.optim = list(trace = 0, fnscale = -1))

round(table.regSGB (regnot1267b),3)
```

Table 1 shows that the algorithm converged properly to a true maximum of the likelihood: `convergence` equals zero, `kkt1` and `kkt2` (first and second Karush-Kuhn-Tucker conditions) equal one.
`value` is the value of the objective function, minus the pseudo-log-likelihood.
The 78 parameters (`n.par`) are (1+17)*4 regression parameters (intercept and 17 explanatory variables for each alr), one overall shape parameter and 5 Dirichlet shape parameters.
There is the possibility to fix some parameters, but the option was not used (`n.par.fixed` = 0).
`AIC` is Aikaike's criterion.
`Rsquare` was defined by Hijazi and Jernigan (2009) as the ratio of the total variance (Aitchison, 1986) of the fitted compositions to the observed compositions.
`counts.function` and `counts.gradient` give the number of times the objective function and the gradient were evaluated.

More interpretation will be given during the talk.

Table 1: Overall results, output of `table.regSGB`

	statistics
value	-16843.051
n.par	78.000
n.par.fixed	0.000
AIC	33842.101
Rsquare	0.763
convergence	0.000
kkt1	1.000
kkt2	1.000
counts.function	3715.000
counts.gradient	673.000

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd (reprinted 2003 with additional material by the Blackburn Press, London (UK).

Craiu, M. and V. Craiu (1969). Repartitia Dirichlet generalizatá. *Analele Universitatii Bucuresti, Mathematicá-Mecanicá 18*, 9–11.

Gershuny, J. and O. Sullivan (2017). *United Kingdom Time Use Survey, 2014-2015. [data collection]*. UK Data Service. http://doi.org/10.5255/UKDA-SN-8128-1.

Graf, M. (2017). A distribution on the simplex of the Generalized Beta type. In J. A. Martín-Fernández (Ed.), *Proceedings CoDaWork 2017*. University of Girona (Spain).

Graf, M. (2019). *SGB: Simplicial Generalized Beta Regression*. R package version 1.0.

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

Madsen, K., H. Nielsen, and O. Tingleff (2004). Optimization With Constraints. Informatics and Mathematical Modelling, Technical University of Denmark.

Maier, M. J. (2015). *DirichletReg: Dirichlet Regression in R*. R package version 0.6-3.1.

Monti, G. S., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2011). Notes on the scaled Dirichlet distribution. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional data analysis. Theory and applications*. Wiley.

Varadhan, R. (2015). *alabama: Constrained Nonlinear Optimization*. R package version 2015.3-1.

Zeileis, A. and Y. Croissant (2010). Extended model formulas in R: Multiple parts and multiple responses. *Journal of Statistical Software 34* (1), 1–13.