A method for Bayesian regression modelling of composition data

Sean van der Merwe

University of the Free State, South Africa

4 February 2019

DON'T PANIC

Overview

Bad example

- We don't report enough "no result" results
- Take the "election" data from "robCompositions" package for R

District	CDU CSU	SDP	GRUENE	FDP	DIE LINKE	other parties	unemployment	income
SH	638756	513725	153137	91714	84177	146781	6.9	3157
HH	285927	288902	112826	42869	78296	82009	7.4	3835
NI	1825592	1470005	391901	185647	223935	348180	6.6	3229
HB	96459	117204	40014	11204	33284	31247	11.1	3505
NW	3776563	3028282	760642	498027	582925	851718	8.3	3547
HE	1232994	906906	313135	175144	188654	331258	5.8	3729
RP	958655	608910	169372	122640	120338	234582	5.5	3356
BW	2576606	1160424	623294	348317	272456	660922	4.1	3664
BY	3243569	1314009	552818	334158	248920	887281	3.8	3525
SL	212368	174592	31998	21506	56045	66051	7.3	3293
BE	508643	439387	220737	63616	330507	224831	11.7	3294
BB	482601	321174	65182	35365	311312	172728	9.9	2742
MV	369048	154431	37716	18968	186871	100709	11.7	2601
SN	994601	340819	113916	71259	467045	345012	9.4	2627
ST	485781	214731	46858	30998	282319	118128	11.2	2648
TH	477283	198714	60511	32101	288615	174469	8.2	2580

- The numbers are distracting, it's the proportions that matter.

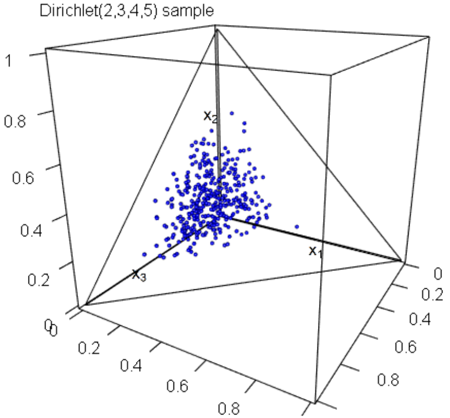
Proportions

District	CDU CSU	SDP	GRUENE	FDP	DIE LINKE	other parties
SH	0.392	0.315	0.094	0.056	0.052	0.090
HH	0.321	0.324	0.127	0.048	0.088	0.092
NI	0.411	0.331	0.088	0.042	0.050	0.078
HB	0.293	0.356	0.121	0.034	0.101	0.095
NW	0.398	0.319	0.080	0.052	0.061	0.090
HE	0.392	0.288	0.099	0.056	0.060	0.105
RP	0.433	0.275	0.076	0.055	0.054	0.106
BW	0.457	0.206	0.110	0.062	0.048	0.117
BY	0.493	0.200	0.084	0.051	0.038	0.135
SL	0.378	0.310	0.057	0.038	0.100	0.117
BE	0.285	0.246	0.123	0.036	0.185	0.126
BB	0.348	0.231	0.047	0.025	0.224	0.124
MV	0.425	0.178	0.043	0.022	0.215	0.116
SN	0.426	0.146	0.049	0.031	0.200	0.148
ST	0.412	0.182	0.040	0.026	0.239	0.100
TH	0.388	0.161	0.049	0.026	0.234	0.142

- The proportions are multivariate — They must be analysed as vectors
- Remembering that they must be positive and sum to one

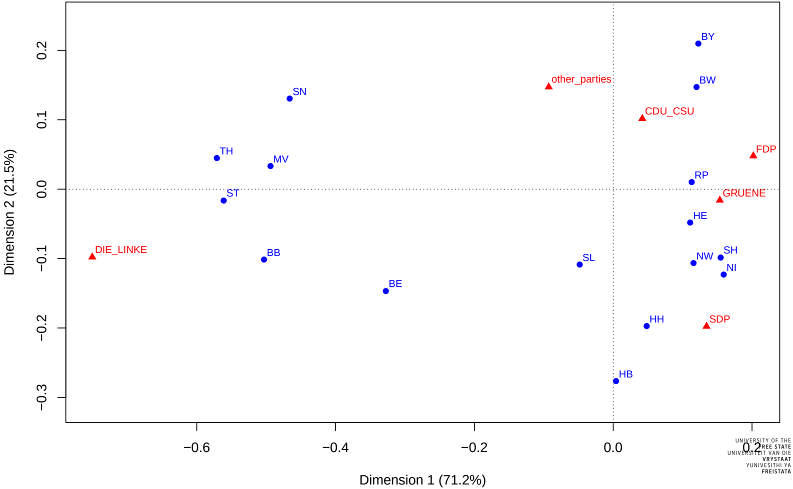
Visualising proportions

- Proportions are restricted to a unit simplex (triangle, pyramid, hyper-pyramid)



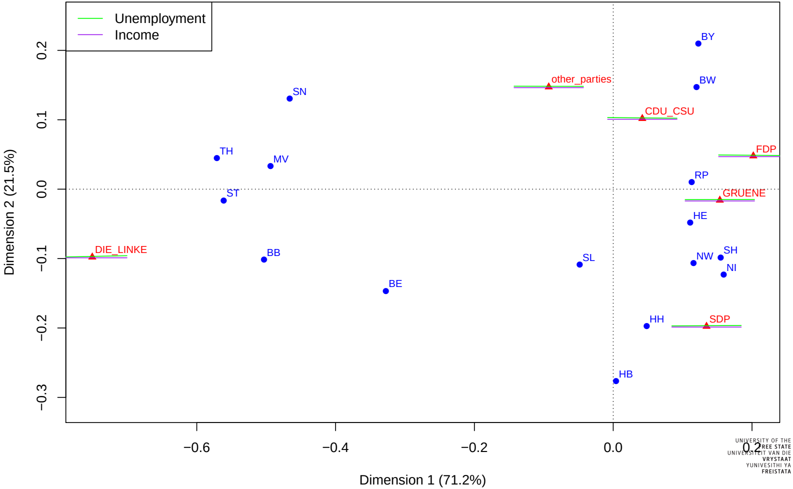
Correspondence analysis

- The traditional use of this example is for correspondence analysis



Correspondence analysis plus regression

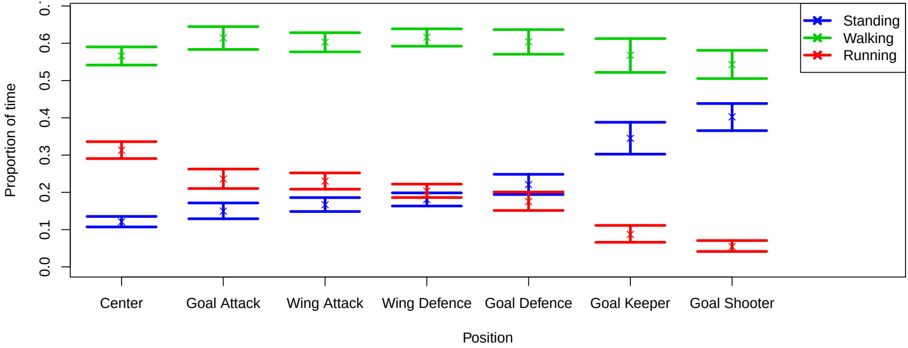
- What about the unemployment and income variables?



Example: Netball players

- Movement speeds of players during a school tournament were tracked, and classified as Standing or Walking or Running
- The goal is to compare the playing positions, while accounting for differing fitness levels of players
- This means mixed effects modelling where the dependent variable is vectors of proportions
- I found significant differences between playing positions in all dimensions

Results: Netball players



Analysing the proportions

- We can try to model the proportions jointly using the Dirichlet distribution:
- $f(\mathbf{y}) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{P} \alpha_j} \prod_{j=1}^{P} y_j^{\alpha_j - 1} \ , \qquad \alpha_0 = \sum_{j=1}^{P} \alpha_j \ , \qquad \alpha_j > 0$
- Has the natural restriction $\sum_{i=1}^{P} y_i = 1$
- Why Dirichlet? Because it's parsimonious! Only P parameters

Overview

Regression

- Regression with the Dirichlet distribution for dependent variables is tricky
- There is not a straightforward relationship between the parameters and the mean
- $E[Y_j] = \frac{\alpha_j}{\sum \alpha_j}$
- There also isn't a neat relationship with the variances, besides the general notion that higher α values result in less variation
- A transformation is required to enable regression models
- Let's review the literature

Direct approach

- One observation is arranged in a row as $\mathbf{y}_{i\cdot} = (y_{i1}, y_{i2}, \ldots, y_{iP}) \sim D(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iP})$
- A sample is then a matrix of n rows $Y = (\mathbf{y}_{1\cdot}; \mathbf{y}_{2\cdot}; \ldots; \mathbf{y}_{n\cdot})$
- Let $X = (\mathbf{x}_{1\cdot}; \mathbf{x}_{2\cdot}; \ldots; \mathbf{x}_{n\cdot})$ be Q explanatory variables arranged the same way (can be anything)
- Model each parameter as a linear function of the explanatory variables, $\alpha_{ij} = x_{i1}\beta_{1j} + \cdots + x_{iq}\beta_{qj} = \mathbf{x}_{i\cdot}\boldsymbol{\beta}_{\cdot j}$
- Introduced by Campbell and Mosimann (1987). Worked on by Hijazi and Jernigan (2009). Best explained in Carmargo et al. (2012).
- Gueorguieva et al. (2008) propose using a log link in each dimension to reduce the number of imposed constraints.

Doesn't solve the interpretation problem

Maier (2014) approach

- A Marco J Maier in Vienna figured out a parameterisation that allows for a dual regression model, where one can specify a model on the mean vector and a model on the precision
  - See http://epub.wu.ac.at/4077/1/Report125.pdf for a their full explanation.
  - He also made an R package (DirichletReg) to help with this
- Define new parameters $\boldsymbol{\mu}_i = E[Y_{i\cdot}]$ and $\phi_i = \alpha_{i0}$, then $\alpha_{ij} = \mu_{ij}\phi_i$.
- These parameters are still restricted positive, so take logs all round, *i.e.* $\log(\alpha_{ij}) = \log(\mu_{ij}) + \log(\phi_i)$
- Define regression models $\log(\boldsymbol{\mu}_{\cdot j}) = f\left(\mathbf{X}, \boldsymbol{\beta}_j\right)$ and $\log(\boldsymbol{\phi}) = g\left(\mathbf{Z}, \boldsymbol{\delta}\right)$

Multivariate logit

- The restricted space is ever present:
- $\sum \alpha_j = \alpha_0 \Rightarrow \sum \mu_j = 1$ and
  $\sum Y_j = 1 \Rightarrow \sum E[Y_j] = 1 \Rightarrow \sum \mu_j = 1$
- To accommodate this they set one dimension as reference by making all coefficients zero ($\boldsymbol{\beta}_b = \mathbf{0}$)
- The parameters are then modelled like so:
- $\boldsymbol{\mu}_{\cdot j} = \frac{\exp(\mathbf{X}\boldsymbol{\beta}_j)}{\sum_{a=1}^{Q} \exp(\mathbf{X}\boldsymbol{\beta}_a)}$
- $\boldsymbol{\mu}_{\cdot b} = \frac{\mathbf{1}}{\sum_{a=1}^{Q} \exp(\mathbf{X}\boldsymbol{\beta}_a)}$
- The parameter estimates can be interpreted as odds ratios after you exponentiate them

Doesn't solve the interpretation problem

My approach

- I don't want a reference category
- I want exactly 1 coefficient connecting 1 explanatory variable to 1 dependent category (with 1 inference)
- I want to be able to fit complicated models, including mixed effects models

Solution:

- Instead of the multivariate logit, I use multiple individual logits for all $\boldsymbol{\mu}_{.j}$
- $\boldsymbol{\mu}_{.j} = \frac{\exp(\mathbf{X}\boldsymbol{\beta}_j)}{1+\exp(\mathbf{X}\boldsymbol{\beta}_j)} \qquad \forall\, j \in 1, \ldots, Q$
- Replace the restriction $(\sum_j \mu_{ij} = 1 \ \forall i)$ with a penalty added to the likelihood: $L^* \propto L * \exp\left\{-\rho \sum_i (\sum_j \mu_{ij} - 1)^2\right\}$

Model definition

I define the model in a hierarchical fashion:

$$\begin{aligned}
\mathbf{y}_{i\cdot} &\sim Dirichlet(\boldsymbol{\alpha}_{i\cdot}) \\
\ln \alpha_{ij} &\sim N\left(\ln \mu_{ij} + \ln \phi_i \ , \frac{1}{\xi^*}\right) \\
\ln \phi_i &= \text{some model for precision} \\
\text{logit}(\mu_{ij}) &= \text{some model for each expected value} \\
\sum_{j=1}^{P} \mu_{ij} &\sim N\left(1, \frac{1}{\xi}\right) \\
\beta_{ij} \ , \ \beta_{i\phi} &\sim N(0, 10000) \\
\xi &\sim Exp\left(\tfrac{P}{1000}\right) \ , \ \xi^* \sim Exp\left(\tfrac{P}{100}\right)
\end{aligned}$$

Overview

Simulation example 1

- Implemented via the R2OpenBUGS system (Sturtz et al., 2005)
- Simulations studies were performed to assess the new methodology:

Scenario A is the MANOVA problem for proportions.
I consider a factor with 3 levels in each of 3 dimensions, ($n = 60$).
Samples are generated according to Maier (2014).
I calculate the average sum of composition errors over hundreds of samples, as well as the prediction interval coverage:

Scenario A	Target	Maier	Me
Error	0	19.59	18.38 (better)
Coverage	0.95	0.87	0.94 (better)

Higher dimensions favour the new approach even further.

Simulation example 2

Scenario B is Scenario A + linear terms in means and precision.

Here I also consider inference — can the models correctly detect the linear relationships, measured by the median p-values?

Scenario B	Target	Maier	Me
Error	0	19.19	18.81 (better)
Coverage	0.95	0.85	0.86 (better)
p-value β_ϕ	0	0.001	0.000 (better)
p-value β_2	0	0.50	0.01 (better)
p-value β_3	0	0.24	0.001 (better)
p-value β_1	0	N/A	0.004 (better)

UNIVERSITY OF THE FREE STATE UNIVERSITEIT VAN DIE VRYSTAAT YUNIVESITHI YA FREISTATA UFS UV

Overview

The point

- I presented a new approach for regression modelling of composition data (vectors of proportions)
- This method combines the best parts of previous (non-Bayes) approaches, and incorporates some modern Bayes ideas
- If you value all dimensions, or you have explanatory factors, or you have random effects, then try this approach
- The new method is more accurate and more flexible than previous methods
- It is also **easier to interpret**