A method for Bayesian regression modelling of composition data

Sean van der Merwe — University of the Free State, South Africa

What is composition data?

- Vector observations
- All positive values
- Sum to one, or less than one

Examples of composition data

- Proportion of votes going to each major party in an election
- Proportion of employee time spent on different activities
- Composition of inputs to a manufacturing process (chemical, mineral, *etc.*)
- Dietary preferences of people or animals under different circumstances
- **Composition of foods:**

	Carbs	+	Fibre	+	Fat	+	Protein	+	Water & Other	=	Total
	11%	+	0%	+	0%	+	0%	+	89%	=	100%
	58%	+	3%	+	5%	+	8%	+	26%	=	100%
	1%	+	0%	+	33%	+	25%	+	41%	=	100%
	2%	+	0%	+	10%	+	12%	+	76%	=	100%
⋮	⋮		⋮		⋮		⋮		⋮		⋮

Standard analysis

1. Pick a reference category
2. Apply transformation to each other category relative to reference
3. Analyse data on free scale

Advantages of standard analysis

- Popular multivariate analysis techniques can be used
- Popular multivariate visualisation works
- Extremely easy to implement **IF** there is an obvious reference category

Disadvantages of standard analysis

- Can't transform back — not 1-to-1
- So all interpretations are **relative**
- Gives different results if you change reference category
- We want to do inference on all categories simultaneously

Regression for compositions

Let $Y = (\mathbf{y}_{1\cdot}; \mathbf{y}_{2\cdot}; \dots; \mathbf{y}_{n\cdot})$ be a sample of vector observations arranged in rows of the matrix Y. Let $X = (\mathbf{x}_{1\cdot}; \mathbf{x}_{2\cdot}; \dots; \mathbf{x}_{n\cdot})$ be Q explanatory variables arranged the same way. $\sum_{j=1}^{P} y_{ij} = 1,\ y_{ij} > 0$; while the values of X could be anything. Campbell and Mosimann (1987), Hijazi and Jernigan (2009) and Carmargo et al. (2012) apply the Dirichlet distribution and model each parameter as a function of the explanatory variables. They use an identity link and describe procedures to estimate these parameters under the constraints that all parameters $\alpha_{ij} > 0$. Gueorguieva et al. (2008) propose using a log link in each dimension to eliminate those constraints. Maier (2014) applies a multivariate transformation to the parameters of the Dirichlet distribution, arriving at an alternative formulation that has the advantage of modelling the expected value of an observation separately from its precision, which he defines as $\phi = \alpha_0$.

The problem is that each coefficient β_{kj} does not have a clean interpretation in the above models as $E[Y_{ij}]$ is a function of all β_{kj}.

Campbell, G. and Mosimann, J. E. (1987), "Multivariate methods for proportional shape", *ASA Proceedings of the Section on Statistical Graphics*, vol. 1, pp. 10–17.

Carmargo, A. P., Stern, J. M., and Lauretto, M. S. (2012), "Estimation and Model Selection in Dirichlet Regression", *Conference proceedings* 1143, pp. 206–213.

Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008), "Dirichlet component regression and its applications to psychiatric data", *Computational statistics & data analysis* 52.12, pp. 5344–5355.

Hijazi, R. H. and Jernigan, R. W. (2009), "Modeling Compositional Data Using Dirichlet Regression Models", *Journal of Applied Probability & Statistics* 4.1, pp. 77–91.

Maier, M. J. (2014), *DirichletReg: Dirichlet Regression for Compositional Data in R*, Technical Report.

Sturtz, S., Ligges, U., and Gelman, A. (2005), "R2WinBUGS: A Package for Running WinBUGS from R", *Journal of Statistical Software* 12.3, pp. 1–16.

My solution: Bayes

- Specify the model in hierarchical form
- Easy to understand and modify

General model specification

$$\mathbf{y}_{i\cdot} \sim Dirichlet(\boldsymbol{\alpha}_{i\cdot})$$

$$\ln \alpha_{ij} \sim N\left(\ln \mu_{ij} + \ln \phi_i\ ,\ \frac{1}{\xi^*}\right)$$

$$\ln \phi_i = g\left(\mathbf{w}_{i\cdot}\boldsymbol{\beta}_{\cdot\phi}\right)$$

$$\text{logit}(\mu_{ij}) = h\left(\mathbf{x}_{i\cdot}\boldsymbol{\beta}_{\cdot j}\right)$$

$$\sum_{j=1}^{P} \mu_{ij} \sim N\left(1, \frac{1}{\xi}\right)$$

$$\beta_{ij}\ ,\ \beta_{i\phi} \sim N(0, 10000)$$

$$\xi \sim Exp\left(\tfrac{P}{1000}\right)\ ,\ \xi^* \sim Exp\left(\tfrac{P}{100}\right)$$

$f(\mathbf{y}) = \frac{\prod_{i=1}^{P}\Gamma(\alpha_i)}{\Gamma(\alpha_0)}\left\{\prod_{i=1}^{P} y_i^{\alpha_i - 1}\right\}$

Decompose means and precisions

Models for means and precisions on free scale, very flexible, can include random effects

Replaced constraint with penalty term.
This is the key value-adding change I introduced 🙂

Vague priors for coefficients

Priors for flexibility parameters

Implementation and results

- Implemented via the R2OpenBUGS system (Sturtz et al., 2005)
- Very flexible system that allows for most scenarios
- Simulations studies were performed to assess the new methodology:

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

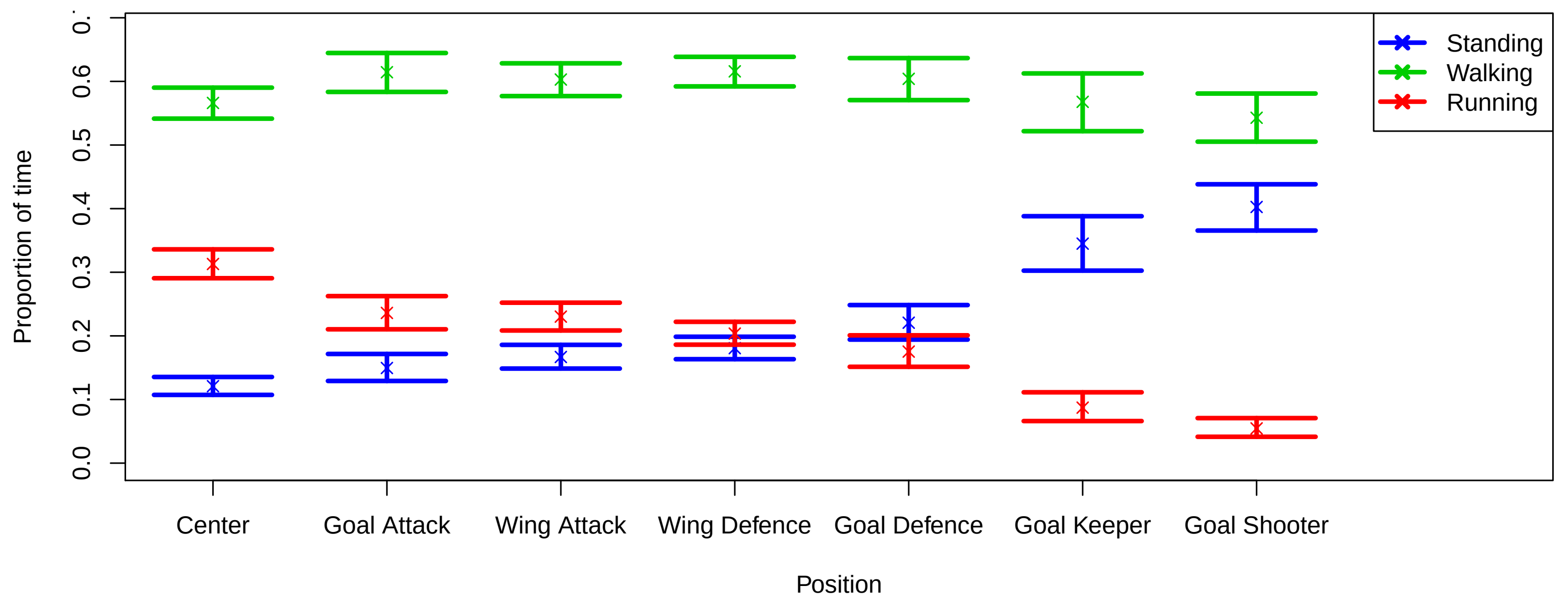
Scenario B is Scenario A + linear terms in means and precision.
Here we also consider inference — can the models correctly detect the linear relationships, measured by the median p-values?

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

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

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 taking the player effect into account
- The new approach allows for random effects modelling
- We found significant differences between playing positions in all dimensions:



Summary

I developed 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

Highlights

- The new method is more accurate and more flexible than previous methods
- It is also easier to understand, use, and interpret