Thesis summary

"Extending Bayesian Analysis of Circular Data to comparison of Multiple Groups"

Kees Mulder

May 29, 2015

Researchers often analyze data that is either *numerical*, e.g. length, or is divided into (ordered) *categories*, e.g. level of education. However, what should researchers do if their data consists of angles, measured in degrees? In that case, the data is called *circular* data. This type of data, in some form or another, actually pops up in a large variety of scientific disciplines. For example:

- In biology, the direction of animal movement (Cochran et al., 2004).
- In political science, the time of day at which crimes are committed (Gill & Hangartner, 2010).
- In social sciences, measurements on a circumplex model, such as Leary's rose for interpersonal behaviour (Leary, 1957).
- In medicine, the structure of proteins and DNA (Harder et al., 2010).
- In colour vision, the hue of an object (Hanbury, 2003).
- In geography, migration flows (Faggian et al., 2013).

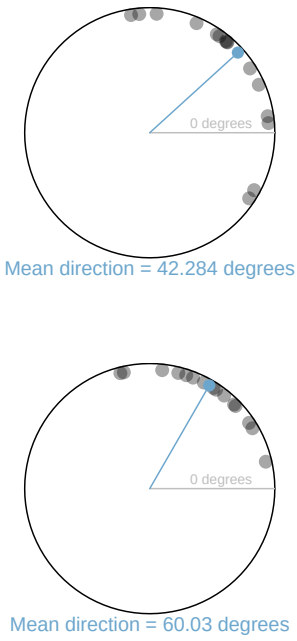

Figure 1: Examples of circular datasets.

Circular data analysis requires completely different models from linear data analysis. To see why circular data is so different, we can take a look at the difference between two points of circular data, 5° and 355°. Because 360° = 0°, we can see that they are both 5° away from 0°, so they differ by 10°. If we pretend that the data here are simply linear, we would have gotten $355 - 5 = 350$, which does not accurately describe the distance between the two datapoints. What is underlying here is the fact that the circular sample space is *periodical*. As a result, we can't talk about one angle being "larger" or "smaller" than another. That, in turn, is a requirement of all linear models, such as regression and ANOVA, that are commonly used to conduct statistical inference. So, we must use specialized methods.

The circular data models we must resort to, however, are somewhat underdeveloped in the literature. Theoretical work into this field is limited, and applications employ comparatively simplistic statistical methods. It has been mathematically challenging to develop new statistical methods for circular data in the past, and in fact it still is.

Bayesian analysis is a framework for statistical sciences that differs fundamentally from the traditional frequentist framework, which is by far more prominent in the literature. Bayesian methods offer an alternative to frequentist methods. In the past two decades, computer algorithms developed for Bayesian analysis, called *MCMC* methods, have become much more feasible as computational power has increased dramatically. This allows us to revisit difficult statistical problems, and attempt to solve them using recent computational and statistical advances.

The model we focus on here is the model where a researcher has a circular outcome variable, and would like to compare multiple groups on their mean. This model is essentially a circular ANOVA-model. An important assumption for this model is *homogeneity of variance*, which means that the variance in each group is equal. An example of a research scenario in which the current model would be useful is a study where a biologist wants to compare the escape directions of three species of birds. Another would be a medical scientist assessing whether patients given medication go to bed later than a control group. Compared to the simple methods that are often employed in these kinds of situations, the model being developed here would be much more informative about the hypotheses of the researcher.

Three new Bayesian circular ANOVA-models were developed in this study: A version of the *Gibbs sampler* extending Damien & Walker (1999), a *Metropolis-Hastings sampler*, and a *rejection sampler* extending Forbes & Mardia (2014). The samplers were implemented in R and C++ via Rcpp, and are easily available online at `https://github.com/keesmulder/BayesianMultigroupCircularData`.

In an extensive simulation study, the three methods were compared on biasedness, precision, ease of use and efficiency. The Gibbs sampler was complex, very slow, and imprecise. The Metropolis-Hastings sampler performed adequately, but performed a bit worse when the data was strongly concentrated. The rejection sampler also performed well, and was slightly more efficient than the Metropolis-Hastings sampler.

In sum, we have added a new tool to the limited toolbox of circular data analysis. It is recommended that researchers apply the rejection sampler developed here for circular ANOVA contexts because it shows the best performance among the available methods. We feel that Bayesian methods offer a flexible approach to circular data analysis, using which it may be possible to develop new models that are lacking in the frequentist framework. Future models may extend the current circular ANOVA model to more complex models, for example to include covariates, or to incorporate repeated measures designs.

References

Cochran, W. W., Mouritsen, H., & Wikelski, M. (2004). Migrating songbirds recalibrate their magnetic compass daily from twilight cues. *Science*, *304*(5669), 405–408.

Damien, P., & Walker, S. (1999). A full Bayesian analysis of circular data using the von mises distribution. *Canadian Journal of Statistics*, *27*(2), 291–298.

Faggian, A., Corcoran, J., & McCann, P. (2013). Modelling geographical graduate job search using circular statistics. *Papers in Regional Science*, *92*(2), 329–343.

Forbes, P. G., & Mardia, K. V. (2014). A fast algorithm for sampling from the posterior of a von mises distribution. *arXiv preprint arXiv:1402.3569*.

Gill, J., & Hangartner, D. (2010). Circular data in political science and how to handle it. *Political Analysis*, *18*(3), 316–336.

Hanbury, A. (2003). Circular statistics applied to colour images. In *8th computer vision winter workshop* (Vol. 91, pp. 53–71).

Harder, T., Boomsma, W., Paluszewski, M., Frellsen, J., Johansson, K. E., & Hamelryck, T. (2010). Beyond rotamers: a generative, probabilistic model of side chains in proteins. *BMC bioinformatics*, *11*(1), 306.

Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald Press.