Research Master's programme METHODOLOGY AND STATISTICS FOR THE BEHAVIOURAL, BIOMEDICAL AND SOCIAL SCIENCES
Utrecht University, the Netherlands

MSc Thesis *Inge Jansen (3221318)*
TITLE: Bayesian Regression Analysis of Circular Data Using the Wrapping Approach
May 2017

Supervisors:
Prof. Dr. *I.G. Klugkist*
Jolien Cremers, MSc

Preferred journal of publication: JOURNAL OF MATHEMATICAL PSYCHOLOGY
Word count: 7785

Bayesian Regression Analysis of Circular Data Using the Wrapping Approach

I. Jansen
Utrecht University

Abstract

In this paper, a novel approach is introduced for circular regression based on the wrapping approach. Circular regression models are developed based on both the wrapped normal (WN) and wrapped Cauchy (WC) distribution. The scope of circular regression based on the wrapping approach is increased by making our methods applicable to models with an arbitrary number of continuous and dichotomous predictors. The resulting methods are validated by means of a simulation study, paying special attention to the differences between WN and WC. To demonstrate the use of our techniques, a regression model is fit to a real dataset from the field of Political Science. Diagnostic tools are provided to test reliability of the results and support inferences based on our methods.

Keywords: Circular data, Wrapping approach, Wrapped Normal distribution, Wrapped Cauchy distribution, Bayesian regression

1 Introduction

Circular data can be found both within and outside the social sciences. Examples are research on the human sense of direction (see for example Brunyé et al. (2015)) or eye-hand coordination (f.e. Rentsch & Rand, 2014). Circular data are not necessarily measured in angles, also clock times can best be represented on the circle. See for example Herrera et al. (2010) who studied music listening patterns or Kimpton et al. (2015) on crime timing. A circular data point is different from a linear data point in the sense that its natural representation is on the circle rather than on the real line. This is made clear in Figure 1. If we would place the two directions 15° and 345° on the real line (left), they appear at the extremities of the interval $[0, 360°]$. However, if we place the same two points on the circle (right), it is clear that the two measurements actually point in almost the same direction.

Conventional statistical methods cannot be used on circular data. As an example, we calculate the mean of the two data points mentioned above: $(15°+345°)/2 = 180°$. However, from Figure 1 it is clear that the *mean direction* is 0°.[1] In other words, circular data require the use of special statistical techniques. Over the last few decades, three different approaches regarding circular data were proposed. The *intrinsic* approach makes use of distributions which are intrinsically defined on the circle, the most common being the Von Mises distribution (Von Mises, 1918). Relevant work on the intrinsic approach includes Damien and Walker (1999); Nuñez-Antonio and Gutiérrez-Peña (2005b); Bhattacharya and SenGupta (2009) and Forbes and Mardia

[1]Note that in the remainder of this paper, angles will be represented in *radians (rad)* rather than degrees, where 1 rad equals $360/2\pi$ degrees.

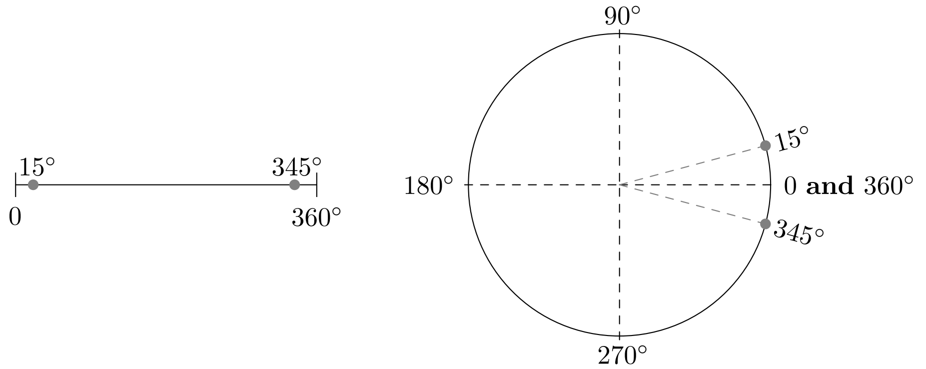


Figure 1. A linear (left) and circular (right) representation of the two data points 15° and 45°.

(2014). In the *embedding* approach points in a two-dimensional space are *projected* onto the circle, see for example Nuñez-Antonio and Gutiérrez-Peña (2005a); Nuñez-Antonio et al. (2011) and Wang and Gelfand (2013). The focus in this paper will be on yet another approach; the *wrapping* approach.

Background: the wrapping approach

The idea behind the wrapping approach is simple (Fisher, 1995; Mardia & Jupp, 1999). The real line from minus to plus infinity is wrapped around the circle in such a way that every two points that are a multiple of 2π apart end up on the same place on the circle. The same wrapping technique can also be applied to probability distributions defined on the real line.

Both frequentist and Bayesian linear regression techniques are often based on the assumption that data are normally distributed. The same idea can be used in circular regression, using the *wrapped Normal* distribution (WN). In Figure 2, the process of wrapping a normal distribution around the circle is illustrated. The first step is to wrap the interval $[-2\pi, 0]$ (corresponding to the striped area in Figure 2a), followed by wrapping the interval $[0, 2\pi]$ (the solid grey area in Figure 2b) and $[2\pi, 4\pi]$ (the dotted area in Figure 2c).

Another probability density of interest to circular regression is the *wrapped Cauchy* (WC) distribution. Neither the mean nor the variance of its linear counterpart are defined, which makes it unusable for linear regression. However, when wrapped around the circle a *close form* of this distribution can be derived with defined moments (Mardia & Jupp, 1999). In Figure 3, we compare the probability density function (pdf) of the WN (solid outlined, grey area) and the WC distribution (dashed outlined, dotted area) for increasing concentration. We observe that for very dispersed data (Figure 3a), both distributions resemble the wrapped Uniform distribution, whereas for more concentrated data (Figure 3b and Figure 3c) the WC tends to be more sharply peaked than the WN.

While in theory we should wrap from minus to plus infinity, in practice this is not feasible and it turns out to be unnecessary for both WN and WC. For WN, Kurz et al. (2014) showed that for sufficiently small values of the standard deviation σ, without loss of accuracy, the density can be approximated by wrapping only a finite number of times. Jona-Lasinio et al. (2013, p. 6) state that for values of the standard deviation $\sigma < 2\pi/3$, it is enough to wrap the density three times, which will result in a *truncated* version of WN. It is possible to directly use the closed form of WC, making it unnecessary to use an approximation.

Because of its intuitive relation to linear distributions, the wrapping approach could serve as a interesting starting point for circular regression. Yet so far its use in circular regression is hardly investigated. Bayesian regression methods based on *data augmentation* were proposed (Coles, 1998), but only applied to the case of circular mean estimation (Ferrari, 2009) or, combined with slice sampling (Neal, 2003), to uni-variable circular regression (Ravindran, 2002; Ravindran & Ghosh, 2011). The only multi-variable model that was derived so far (for WC) was based on maximum likelihood (ML) estimation (Abuzaid & Allahham, 2015). Furthermore, the methods developed so far were only validated by means of simulation studies for ideal data (e.g., perfectly distributed residuals, sufficiently large sample size, etc.) and for a limited set of parameter values.

Instead of using a data augmentation or a ML approach, we will directly use the truncated version of the WN of (Kurz et al., 2014) or the closed form of the WC distribution and combine

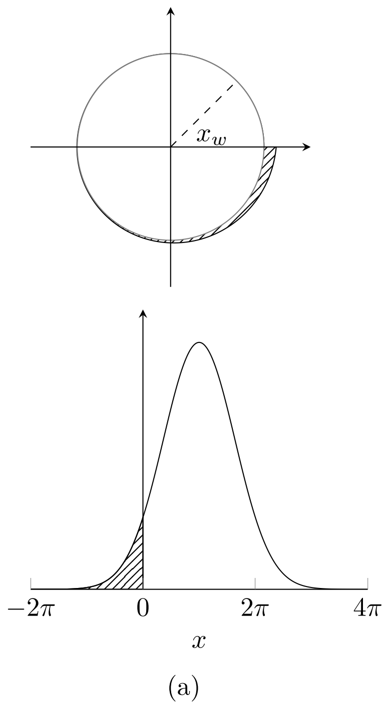


Figure 2. Three consecutive steps of wrapping a $N(\pi, 2)$ distribution around the circle. On the bottom row the probability density is shown on the real line, whereas the top row shows its circular counterpart.

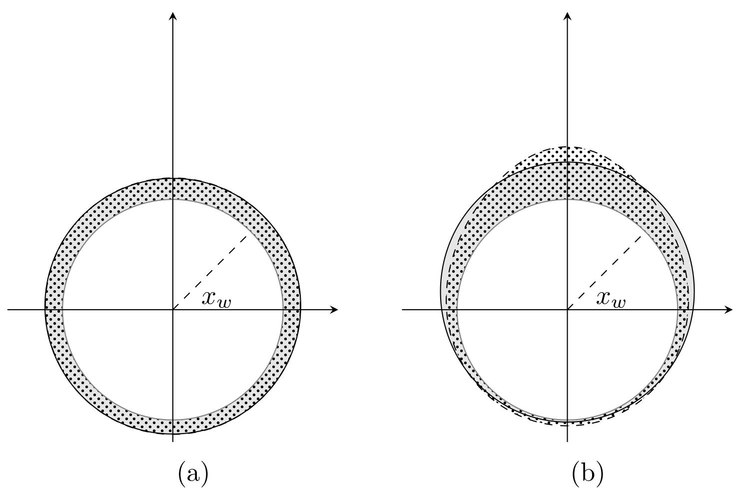


Figure 3. The pdf of the wrapped normal distribution (solid outlined, grey area) and the wrapped Cauchy distribution (dashed outlined, dotted area) for constant mean $\mu = \frac{\pi}{2}$ and increasing values of concentration.

these with slice sampling techniques to derive a multi-variable Bayesian circular regression model with a circular outcome measure and an arbitrary number of continuous and dichotomous predictors. This way we propose both an *adjustment* to existing methods as well as an *extension*. We will perform a extensive simulation study showing both the large scope of our methods as well as exposing its limitations in practice.

In Section 2 we will derive the posterior distribution based on both WN and WC and describe how to sample from it. The validation of our methods by means of a simulation study will be presented in Section 3. We will apply the new regression techniques to an existing dataset from Political Science in Section 4. Finally, in Section 5 the performance of our methods will be discussed.

2 Methods

In this section we will introduce a Bayesian circular regression technique based on WN and WC likelihoods for models with one circular outcome measure and an arbitrary number of dichotomous and continuous predictors. After defining the model, we will introduce prior distributions for the model parameters and derive the posterior distribution. Differences and similarities between the WN and WC method will be made explicit.

2.1 Model

We define a general circular regression model for a vector of circular outcomes $\mathbf{y}=(y_1, \cdots, y_n)$, with n being the sample size. We will predict the outcomes by the model matrix $\mathbf{X}=(\mathbf{1}, \mathbf{x}_1, \cdots, \mathbf{x}_d, \mathbf{x}_{d+1}, \cdots, \mathbf{x}_p)$, with d the number of dichotomous independent variables and $p - d = c$ the number of continuous predictors. The number of rows in $\mathbf{x}$ equals the sample size. Note that categorical variables having more than two levels can be incorporated in this model as dummy variables. The circular residual error terms $\mathbf{R}$, are assumed to have a wrapped distribution (WD) with mean zero and circular concentration measure ρ:

$$\mathbf{R} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} \mod 2\pi \sim \text{WD}(\mu = 0,\ \rho \cdot \mathbf{I}_n), \tag{1}$$

where $\hat{\mathbf{y}}=(\hat{y}_1, \ldots, \hat{y}_n)$ represent the vector of predicted values of $\mathbf{y}$ and $\mathbf{I}_n$ denotes an n-dimensional identity matrix. The wrapped distribution WD can either stand for WN or WC.

The predicted values $\hat{\mathbf{y}}$ can be written as:

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \boldsymbol{\beta} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \tag{2}$$

where $\mathbf{X}$ denotes the model matrix (of size n $\times$ p) and $\boldsymbol{\beta}$ together with ρ form the set of model parameters. If we split $\mathbf{X}$ in a dichotomous and a continuous part, we can rewrite this equation as:

$$\hat{\mathbf{y}} = \overbrace{\underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{d1} \\ 1 & x_{12} & \cdots & x_{d2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{dn} \end{bmatrix}}_{\text{dichotomous predictors}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}}^{\text{circular intercept}} + \underbrace{\begin{bmatrix} x_{(d+1)1} & \cdots & x_{p1} \\ x_{(d+1)2} & \cdots & x_{p2} \\ \vdots & \ddots & \vdots \\ x_{(d+1)n} & \cdots & x_{pn} \end{bmatrix}}_{\text{continuous predictors}} \overbrace{\begin{bmatrix} \beta_{(d+1)} \\ \beta_{(d+2)} \\ \vdots \\ \beta_{(c+d=p)} \end{bmatrix}}^{\text{linear slopes}}, \tag{3}$$

where the first part can be interpreted as the circular intercept of the model, taking values between 0 and 2π.

In general, the likelihood of the data for any wrapped distribution WD can be denoted as:

$$\mathrm{L}(\mathbf{y} \mid \mathbf{X},\ \boldsymbol{\beta},\ \rho) = \prod_{i=1}^{n} \mathrm{WD}(y_i \mid \mu_i = \hat{y}_i,\ \rho), \tag{4}$$

where $\mathbf{x}_{*,i}$ denotes the i^{th} row of the model matrix.

If we use the truncated WN distribution as the wrapped pdf, we obtain the likelihood of a dataset with WN distributed error terms:

$$\mathrm{L}_{\mathrm{WN}}(\mathbf{y} \mid \mathbf{X},\ \boldsymbol{\beta},\ \rho) = \frac{1}{(\sqrt{-4\pi \log \rho})^n} \prod_{i=1}^{n} \sum_{k=-2}^{2} \exp\left(\frac{-(r_n + 2\pi k)^2}{4 \log \rho} \right), \tag{5}$$

where k is usually referred to as the *wrapping coefficient*. Note that the circular concentration parameter ρ of the WN is related to the standard deviation σ of the (linear) Normal distribution as $\sigma = \sqrt{-2 \ln(\rho)}$.

If, in turn, we use the closed form of WC, we find the following likelihood:

$$\mathrm{L}_{\mathrm{WC}}(\mathbf{y} \mid \mathbf{X},\ \boldsymbol{\beta},\ \rho) = \frac{1}{(2\pi)^n} \prod_{i=1}^{n} \frac{1 - \rho^2}{1 + \rho^2 - 2\rho \cos(r_n)}. \tag{6}$$

2.2 Prior distributions

For the model parameters we define the following set of uninformative prior distributions, similar to the priors used by Ravindran (2002). Note that we use the same priors for both WN and WC.

We use the circular uniform distribution for both the intercept β_0 and the model parameters $\beta_1, \cdots, \beta_d$:

$$\pi(\beta_0, \cdots, \beta_d) = \left(\frac{1}{2\pi} \right)^{d+1}. \tag{7}$$

For the slopes $\beta_{d+1}, \cdots, \beta_p \in \mathbb{R}$, we write:

$$\begin{aligned} \pi(\beta_{d+1}, \cdots, \beta_p) &= \mathrm{MVN}(\boldsymbol{\mu},\ \Sigma) \\ &= \exp\left(-\frac{\beta_{d+1}^2 + \cdots + \beta_p^2}{2\sigma_\beta^2} \right), \end{aligned} \tag{8}$$

which is a multivariate normal density with mean vector $\boldsymbol{\mu} = 0$ and covariance matrix Σ being equal to a c-dimensional identity matrix multiplied by the prior variance σ_β^2. Note that we assume the prior variance to be equal for all continuous predictors in the model.

Finally, the circular concentration parameter ρ can take values between 0 and 1. We define the following prior density on this domain:

$$\pi(\rho) = \rho^{-0.5}(1 - \rho)^{-0.5}. \tag{9}$$

2.3 Posterior distribution

Having expressions for both the prior densities and the density of the data, we find the joint distribution being proportional to the multiplication of Equations 4, 7, 8 and 9:

$$\mathrm{J}(\boldsymbol{\beta},\ \rho \mid \mathbf{y}, \mathbf{X}) = \mathrm{L}(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \rho) \cdot \pi(\beta_0, \cdots, \beta_d) \cdot \pi(\beta_{d+1}, \cdots \beta_p) \cdot \pi(\rho), \tag{10}$$

where the likelihood L can either be the WN (Equation 5) or the WC likelihood (Equation 6). The joint distribution cannot be written in terms of a standard distribution, therefore it is not possible to directly sample from it. This problem can be overcome by using slice sampling techniques (Neal, 2003). A conceptual introduction to slice sampling is available in Appendix A1. We will explicitly show how slice sampling can be used to sample the joint distribution for both WN and WC in Appendix A2.

3 Simulation studies

To validate our methods, we perform a number of simulation studies. The results of these studies are presented in this section. The main studies cover an Ancova and a linear predictor (LP) setting. Some smaller studies are done to investigate how the sampler reacts to violations of assumptions. The division of this section will follow this distinction. Note that all simulations, unless stated otherwise, are performed with a fixed number of samples of 500 and a number of observations (N) of 50. We use two chains with different initial values. The total number of iterations for each chain is 7000, of which 3000 are regarded as burn-in. For the slice sampler, a step-size $w = 1$ is used (refer to Neal (2003)). Simulations with the WN sampler are performed with a wrapping coefficient k=1. Finally, we use a prior variance of 5. This value will be further discussed below. Convergence and autocorrelations are checked by means of trace and autocorrelation plots. Unless stated otherwise, we do not find convergence or autocorrelation problems.

3.1 Ancova setting

We start our simulation studies with an Ancova setting, that is, a regression model with one circular outcome (y), one dichotomous (x_1) and one linear (x_2) predictor:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon. \tag{11}$$

Note that model parameters β_0 and β_1 are circular, while β_2 is linear. The dichotomous variable x_1 is sampled from a Binomial distribution with success probability $p = 0.5$. The linear variable x_2 is sampled from a standard Normal distribution. The error ϵ is sampled from either a WN or WC distribution with mean $\mu = 0$ and circular concentration ρ.

For the model parameters, we use all combinations of the following values:

- β_0 = 0, 3, 6
- β_1 = 0, 3, 6
- β_2 = -5, -1, 0, 1, 5
- ρ = 0.5, 0.9

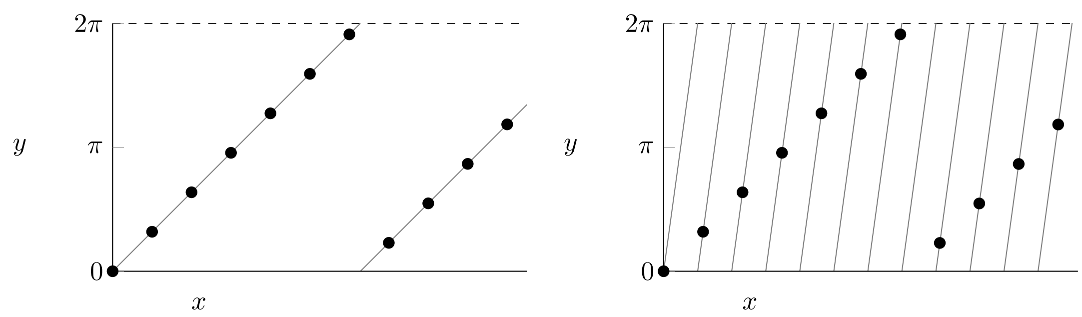
The intercept parameters can take values around the whole circle. The range of values for the slopes are restricted. Due to the nature of wrapping, steep slopes will result in data following a divergent pattern. For slopes steeper than (minus) five, the resulting data is unsuitable for regression. Therefore, we only look at slope values between minus and plus five. The same argument holds for the values of ρ. Values of the circular concentration lower than one half result in (approximately) uniformly distributed residuals. Therefore, only the cases $\rho = 0.5$ and $\rho = 0.9$ will be considered.

Using the values above for the model parameters, for each observation the circular outcome y is randomly sampled according to Equation 11. The resulting datasets with WN distributed residuals are analysed with the WN sampler. Datasets with WC distributed residuals are analysed with the WC sampler. In Table 1, we summarize the results of the ten designs with $\beta_0 = \beta_1 = 3$ for both the WN and the WC sampler. Note that results from the other designs are similar. Full results are available from the authors.

We first discuss the right-most column of Table 1, which concerns a convergence score. The first 100 samples of each design were visually checked for convergence by means of trace plots. We counted the number of samples that did converge, resulting in a convergence score between 0 and 100. For designs where $\rho = 0.9$ all of the first 100 samples converge. For the designs where $\rho = 0.5$, however, we observe a somewhat different pattern. Most of the samples within each design still converge, but in some of the samples the sampler starts to make *jumps*. This is due to the fact that the posterior distribution, besides the absolute maximum has multiple local maxima. These local maxima tend to present a higher portion of the density when the value of ρ decreases. That is, for lower values of ρ the sampler will be more likely to jump to another (local) maximum.

The meaning of these local maxima can be understood by looking at Figure 4. On the left we see a dataset with a circular outcome y and a linear predictor x, together with a (perfectly) fitting regression line. Note that this plot in fact depicts the *net* of a cylinder cut along the dashed line. If we look at Figure 4b, however, we see the same data together with a regression line that fits equally well. Put differently, due to circularity, more than one model might serve as a good fit to the data. This problem was observed, but not fully investigated in previous studies (Coles, 1998; Ferrari, 2009). A proposed workaround is the use of a relatively low value of the prior variance in the prior density of the slopes. This way, the sampler is *guided* towards the easier model. The value we use here (e.g., $\sigma_\beta = 5$) corresponds to the value that was used by Ravindran (2002).

From Table 1 we observe that despite the relatively low prior variance, for $\rho = 0.5$ the sampler cannot always be prevented from jumping to other local maxima, resulting in a low convergence score. The process of jumping is illustrated in Figure 5, where we plot the trace of the two chains (in red and blue) of β_2 for the design with true value $\beta_2 = 1$, analysed with the WC sampler for three different samples. The sample in Figure 5a converges well within the burn-in period (indicated in solid grey). Regarding the samples where the sampler does



(a) (b)

Figure 4. A dataset with a circular outcome y and a continuous predictor x together with two equally fitting regression lines.

jump, we distinguish between two cases. Figure 5b shows a sample with *minor* jumps. That is, after a couple of iterations the sampler jumps back to the absolute maximum (matching the true parameter value). We do not expect those jumps to influence the summary statistics. In Figure 5c, on the other hand, we see a sample with *severe* jumps. In this case it is likely that the summary statistics are affected. In this specific design, a total of 41 of the first 100 samples show severe jumps (resulting in a convergence score of 59). From Table 1, it seems that the WN sampler is less sensitive for having severe jumps.

A valid question is whether the problems with convergence we encounter in this study will pose a problem for researchers who want to apply our methods to empirical data. We argue that this does not necessarily need to be the case. First of all, as shown in Figure 5, it is clear from the trace plot when severe jumps are present. In other words, if a trace plot is consulted in addition to the summary statistics, the risk of making inferences based on a badly converged sample is minimal. Moreover, because we are dealing with a random process, it can simply help to rerun the analysis, optionally with different initial values.

In the remaining columns of Table 1, we report the summary statistics based on all samples. The reported absolute values of the biases in Table 1 of the mean (for β_0, β_1 and β_2) or the mode (for ρ, calculated as in Bickel and Frühwirth (2006)) are relative (in %), except for the case where β_2 equals zero. In that case, the absolute bias is reported. In terms of bias, we do not observe a clear difference in performance between the two sampler techniques for $\rho = 0.9$. At lower circular concentration ($\rho = 0.5$), the WN sampler seems to perform better in some of the designs. Note, however, that those cases of better performance correspond to a higher convergence score. In other words, the results of the WC sampler at $\rho = 0.5$ are disturbed by severe jumps of the sampler.

The interval widths (IW) reported are the widths of the 95% credible intervals (CI) for the intercepts and slope parameters (e.g. β_0, β_1 and β_2), and the Highest Posterior Density (HPD) interval for the circular concentration measure (ρ). The CI of circular parameters are calculated with respect to the point on the circle opposite to the posterior mean. The interval widths of ρ can be regarded equal for both sampling methods under all designs. For the designs under $\rho = 0.9$ the interval widths of the other parameters seem to be more narrow for the WC sampling method. The opposite seems to be the case at $\rho = 0.5$, however, those values again are hard to compare because of the difference in convergence between the two samplers. The coverage probabilities (CP) of all designs range between .92 and .98, which is good.

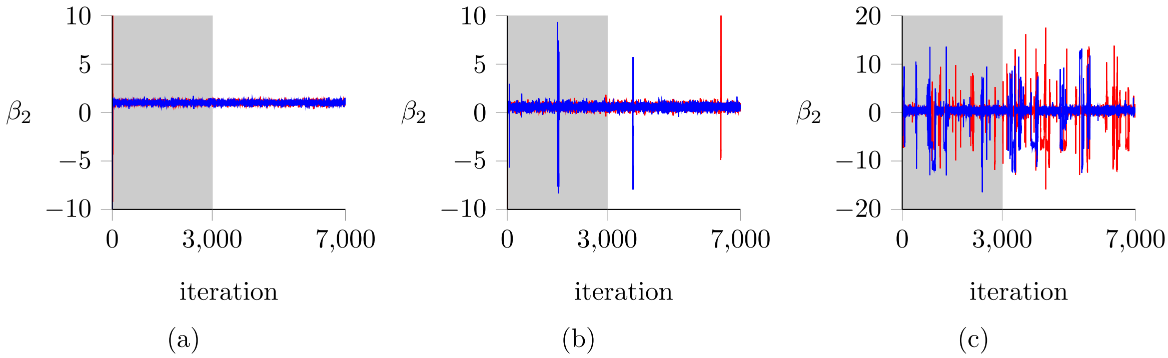


Figure 5. Trace plot of a sample with no jumps (left), some jumps (middle) and many jumps (right).

3.2 Linear predictor setting

The next part of our simulation studies concerns a linear regression setting in which a circular outcome (y) is predicted by two linear variables (x_1 and x_2), sampled from a standard bivariate normal distribution. Note that for now we assume the correlation between the two predictors to be zero. This results in a model equal to Equation 11, with the one difference that model parameters β_1 and β_2 now are both linear.

For the model parameters, we use all combinations of the following values:

- $\beta_0 = 3$
- $\beta_1 =$ -5, -1, 0, 1, 5
- $\beta_2 =$ -5, -1, 0, 1, 5
- $\rho =$ 0.5, 0.9

Using those values, the circular outcomes y are randomly sampled assuming WN/WC distributed error terms.

In Table 2, we summarize the results of the ten designs with $\beta_0 = 3$, $\beta_1 = 1$ for both the WN and the WC sampler in the same way as with the Ancova setting. Again, in terms of bias, the only evident differences between the two sampling methods correspond to cases in which relatively many severe jumps occur (as can be observed from the convergence score). With regard to interval widths we observe a pattern similar to the Ancova setting: for $\rho = 0.9$ the WC sampler seems to perform better for all model parameters but ρ (for which the performance is equal with both sampling methods), while for $\rho = 0.5$ the results are hard to compare due to convergence problems. Results regarding coverage and convergence are in line with the Ancova setting.

3.3 Sensitivity of the sampler

The remainder of our simulations concerns a few smaller studies set up to investigate how sensitive our sampling methods are to input data deviating from the ideal situation. In particular, we will look at sensitivity to sample size, to wrong assumptions about the distribution of the error terms, to multicollinearity for models with two continuous predictors and to unequal group sizes for models with a dichotomous predictor. We will use relative bias as a measure of *accuracy* of the parameter estimates. The *precision* is reflected in the standard deviation (sd), where an *in*crease in sd should be regarded as a *de*crease in precision.

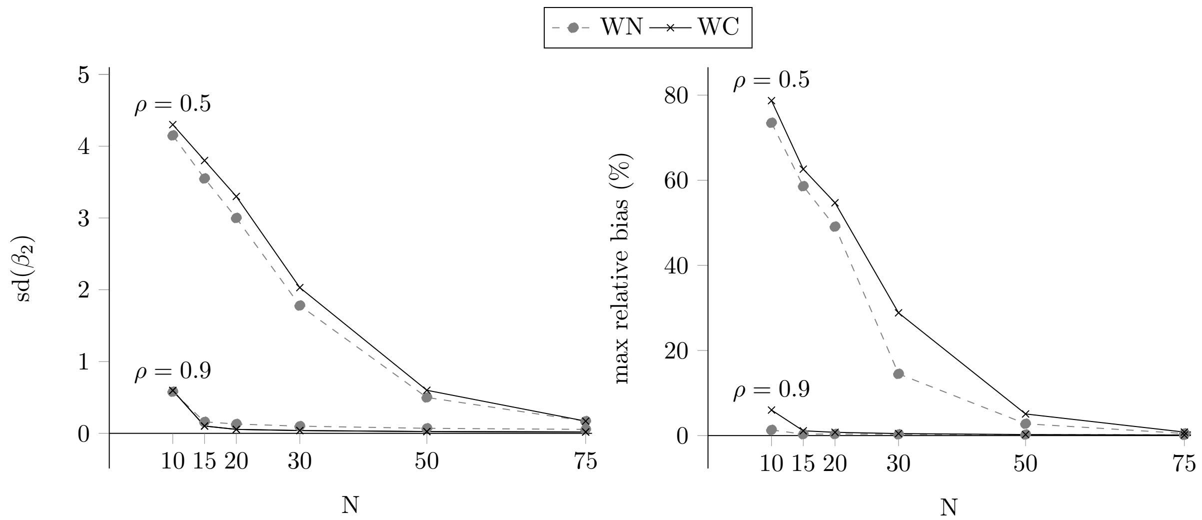
3.3.1 Sensitivity to sample size. We investigate sensitivity to sample size (N) for both the Ancova and LP setting. For the first type of model we use $\beta_0 = \beta_1 = 3$ and $\beta_2 = 1$, for the model with two continuous predictors we use $\beta_0 = 3$ and $\beta_1 = \beta_2 = 1$. Both the WN and WC models are analysed for $\rho = 0.5, 0.9$ with the corresponding sampling method. Finally, the sample size takes values of 10, 15, 20, 30, 50 and 75.

We observe an influence of sample size on both the precision and the accuracy of the parameter estimates. In Figure 6 the results are depicted for Ancova (top row) and LP models (bottom row) based on WN (dots) and WC (crosses) distributed errors. The two pairs of lines in each graph represent the two different values of ρ. In the left column the standard deviation of β_2[2] is plotted against the sample size. In the right column of Figure 6 we show the maximum observed relative bias (in %) of any of the model parameters (including ρ) against N. For $\rho = 0.5$, for all combinations of Ancova/LP and WN/WC, we observe a decrease in both the standard

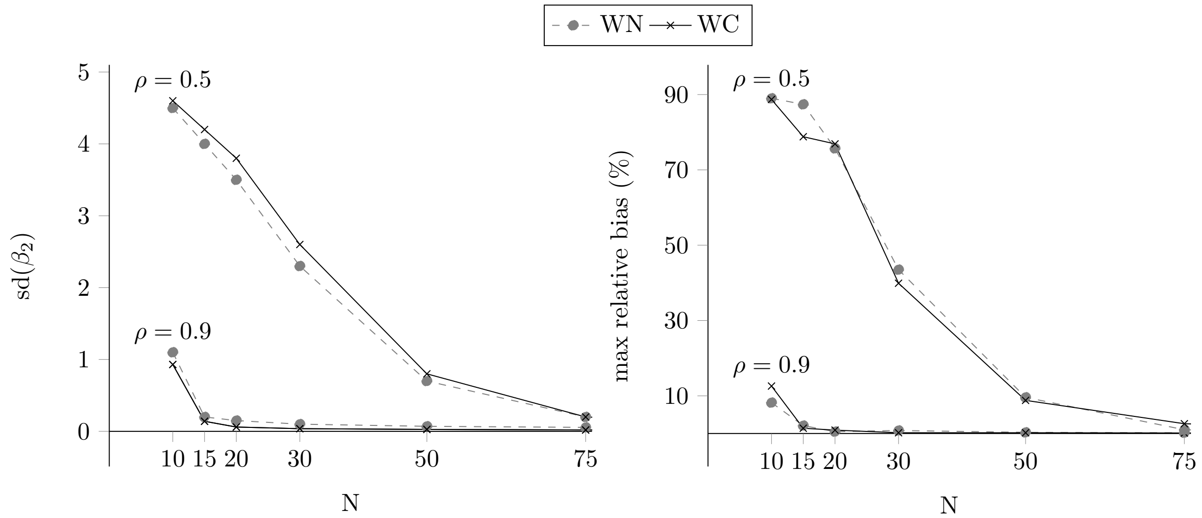
[2] Note that the pattern for the other model parameters was similar.

deviation and the relative bias with an increase in sample size. Standard deviations and biases are consequently higher than in the similar designs under $\rho = 0.9$. Note that for this higher circular concentration, both the standard deviation and the relative bias already tend to be constant for sample sizes of 15 and higher. Here, we have to remark that the results of $\rho = 0.5$ can be influenced by the severe jumps as described above.

Based on this study, we conclude that the sensitivity to sample size is highly dependent on the value of ρ. For a large circular concentration, both sampling methods already worked well (in terms of precision and accuracy) for both kind of models for sample sizes of 15 and higher. For the lower value of ρ, it took a larger sample size to obtain reliable results. It is hard to draw a thorough conclusion as the results, to some extent, might be influenced by poor convergence, but we think it is safe to say that we would recommend a minimum sample size of 30 under this condition. This recommendation is based on a model with two predictors. We expect the minimal sample size to be higher for more extensive models, but will not investigate this here. In this study we also do not address the possible relation between convergence and sample size.



(a) Sensitivity to sample size in the Ancova setting.



(b) Sensitivity to small sample size in the linear predictor setting.

Figure 6. Sensitivity to sample size as reflected in the standard deviation and the maximal relative bias.

3.3.2 Sensitivity to wrong assumption about the distribution of the residuals.

One of the most important assumptions of circular regression concerns the distribution of the error terms. Our methods assume either WN or WC distributed errors. To test the sensitivity of both samplers to a *wrong* assumption about the distribution of the residuals, we choose one of the Ancova ($\beta_0 = \beta_1 = 3$, $\beta_2 = 1$ and $\rho = 0.9$) and one of the LP settings ($\beta_0 = 3$, $\beta_1 = \beta_2 = 1$, $\rho = 0.9$) with either WN or WC distributed error terms, thus resulting in four datasets. We analyse these data with the wrong sampling method. Note that we only investigated the effect for $\rho = 0.9$. At this value of the circular concentration, we expect the effect to be larger than for lower values of ρ since WN and WC deviate the most (see Figure 3c).

The results of this simulation study can be found in Table 3. We observe no or little effect on the absolute value of the relative biases of the parameter estimates, except for ρ. The effect on accuracy of ρ is largest for the WN datasets analysed with the WC sampler, where we observe an underestimation in ρ of about 15 percent. Looking at Figure 3 this is as expected because the WN distribution for $\rho = 0.9$ looks like a WC distribution with lower concentration. Vice versa (i.e. WC data analysed with the WN method), the effect seems less severe (about 3%). Compared to the previous studies, the IW for the data with WN distributed residuals remain the same, while for the data with WC distributed residuals interval widths are wider. The coverage probabilities of the model parameters except ρ are lower than in the previous studies, but still acceptable. The CP of ρ are very low for all the datasets. All of the first 100 samples converged for all datasets.

In general, it can be concluded that a violation of the assumption about the distribution of the residuals mainly affects the parameter estimate of ρ. The summary statistics of the other parameters can be less precise in some cases, but this precision is still acceptable.

3.3.3 Sensitivity to multicollinearity.

To investigate the sensitivity to correlated (continuous) predictors we fit models (WN and WC) with two predictors with β_0=0, β_1=β_2=1 and ρ=0.9[3]. The simulated correlation between the predictors takes values of -0.9, -0.5, -0.1, 0, 0.1, 0.5 and 0.9. Note that the cases with zero correlation correspond to the previously analysed designs.

We observe that parameter estimates remain stable under the variation of correlation. However, multicollinearity has an effect on the precision of the estimates. The precision, seems to remain constant up to correlations of 0.5, but rapidly decreases for more strongly correlated predictors. In Figure 7, we plotted the standard deviation of β_1 against the correlation between x_1 and x_2 for the WC (crosses) and the WN (dots) models. Note that the standard deviation of β_2 took similar values. For both WN and WC, the standard deviation for highly correlated predictors is about four times as large as the standard deviation for uncorrelated predictors.

3.3.4 Sensitivity to unequal group size.

To investigate the sensitivity to unequal group size, we fit a model (WN and WC) with one dichotomous predictor for $\beta_0 = 0$, $\beta_1 = 1$ and $\rho = 0.5, 0.9$. The group size, indicated by the number of "ones" in the total of 50 observations, takes values of 5, 15, 25, 35 and 45.

Results are depicted in Figure 8, where dots indicate WN and crosses WC. The two pairs of lines are the results for the two different values of ρ. We observe an effect of group size on the precision of the parameter estimates. The most precise estimates seem to be obtained when group sizes are equal. Note that we did not find an effect of group size on the accuracy of the sampler.

[3] A similar study was performed under $\rho = 0.5$, but results were affected by poor convergence.

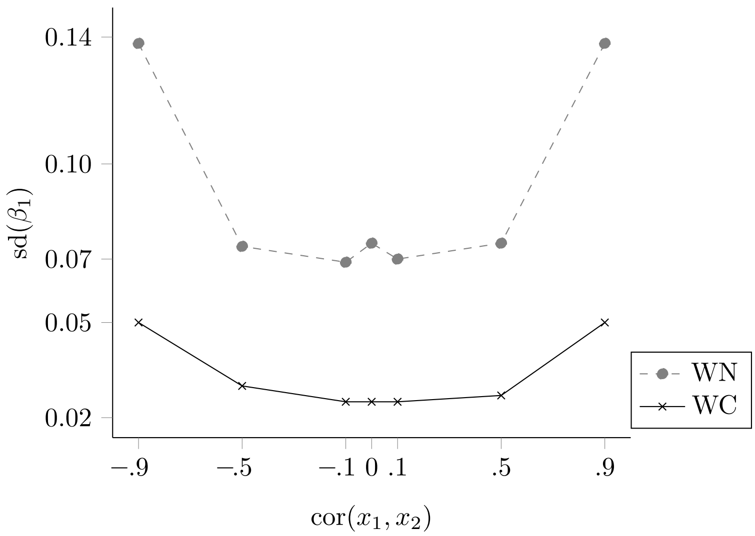


Figure 7. The effect of (multi)collinearity on the precision of the parameter estimates of the WN (blue) and WC (red) sampler for $\rho = 0.9$.

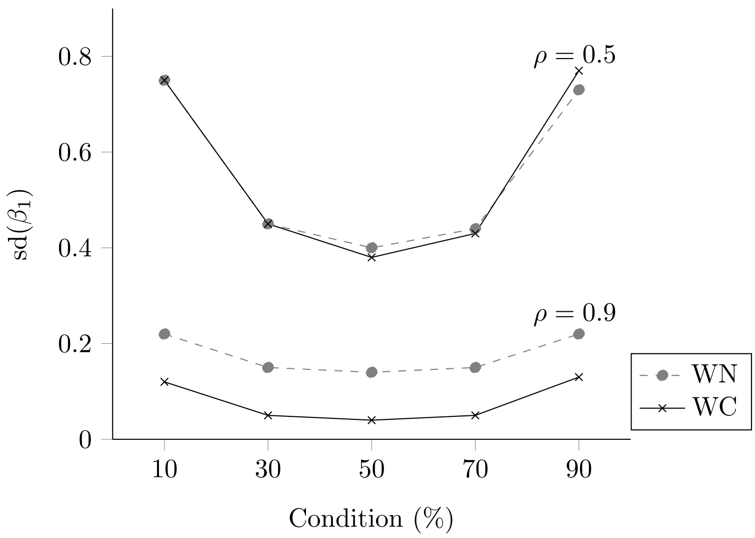


Figure 8. The effect of group size (on the x-axis) of a dichotomous variable on the standard deviation of its parameter estimate for the two sampling methods at three different values of ρ.

4 Application to an empirical dataset

In this section we demonstrate how our methods can be used in practice. To this end we apply our regression techniques to an empirical dataset from Gill and Hangartner (2010) on party preferences in the German Bundestag[4]. Our starting point will be the *full* model as it was presented in Gill and Hangartner.

4.1 Description of the data

The dataset of Gill and Hangartner contains information about the preferred ideological direction of five different parties (FDP, CDU/CSU, SPD, Greens and PDS) in the sixteen German elections between 1949 and 2005. The total number of observations (N) equals 60. Party preferences can

[4]Data were retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/15875, May 2nd 2017.

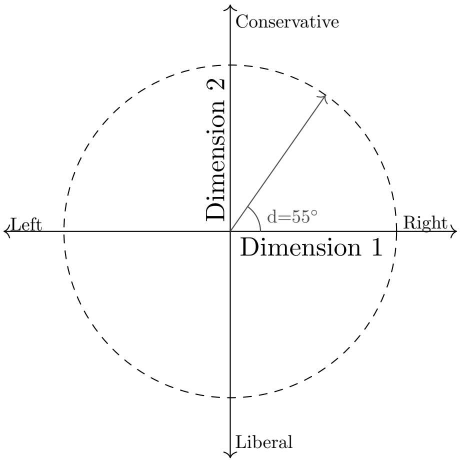


Figure 9. Two-dimensional ideological space as used by Gill and Hangartner (2010)

be referred to as a vector (d) in a two-dimensional space which is depicted in Figure 9. The first dimension (x-axis) captures attitude regarding economic issues ranging from the political left to right. The second dimension (y-axis) concerns the party's opinion on social/cultural matters, ranging from liberal (a low value) to conservative (high). It is the *direction* rather than the *length* of the vector (d) that is of particular interest since it gives information about the direction in which a party wants to shift policy. The vector of a hypothetical party is shown in grey in Figure 9 at a direction of 55°, meaning that this party's ideology would be somewhat on the right concerning economics and conservative with regard to social/cultural issues.

The preferred ideological direction serves as an outcome variable in our regression model. Besides the five party dummies, the dataset contains information on unemployment rate (Unemp), the proportion of out of wedlock births (Outofwed), the year since the first elections (Year), its square (YearSQ) and a dichotomous variable indicating German reunification (0 before 1990). Although the variable Year can be treated as continuous, it is special in the sense that it only takes discrete values. In Table 4 the circular mean of direction is given (overall and per party). In Table 5, we summarize the other variables in the dataset.

4.2 Full model

The model presented and analysed by Gill and Hangartner (2010) includes all variables. Their regression techniques are a combination of ML and Bayesian methods based on an intrinsic-like approach. The full model can be written as follows:

$$\begin{aligned}\text{Ideal direction} = \beta_0 + &\beta_1 \cdot \text{Unemp} + \beta_2 \cdot \text{Outofwed} + \beta_3 \cdot \text{Reunification} + \\ &\beta_4 \cdot \text{SPD} + \beta_5 \cdot \text{CDU/CSU} + \beta_6 \cdot \text{Greens} + \beta_7 \cdot \text{PDS} + \\ &\beta_8 \cdot \text{Year} + \beta_9 \cdot \text{YearSQ} + \epsilon.\end{aligned}$$

Note that the dummy for FDP was left out of the model, that is, it is the reference category. All data were manipulated as in Gill and Hangartner to make the results directly comparable. This means that we did not center or scale the variables. The model was fit with both WN and WC using 2 chains, each having a fairly high total of 20,000 iterations of which the first 15,000 were regarded as burn-in.

The autocorrelation of the samplers can be checked by means of a autocorrelation plot where the correlation between sampling results is plotted against the distance between the sampling results (lag). In Figure 10, the autocorrelation plots for two of the parameters (β_1 on the left and β_2 on the right) are shown for both the WN (top row) and WC sampler (bottom

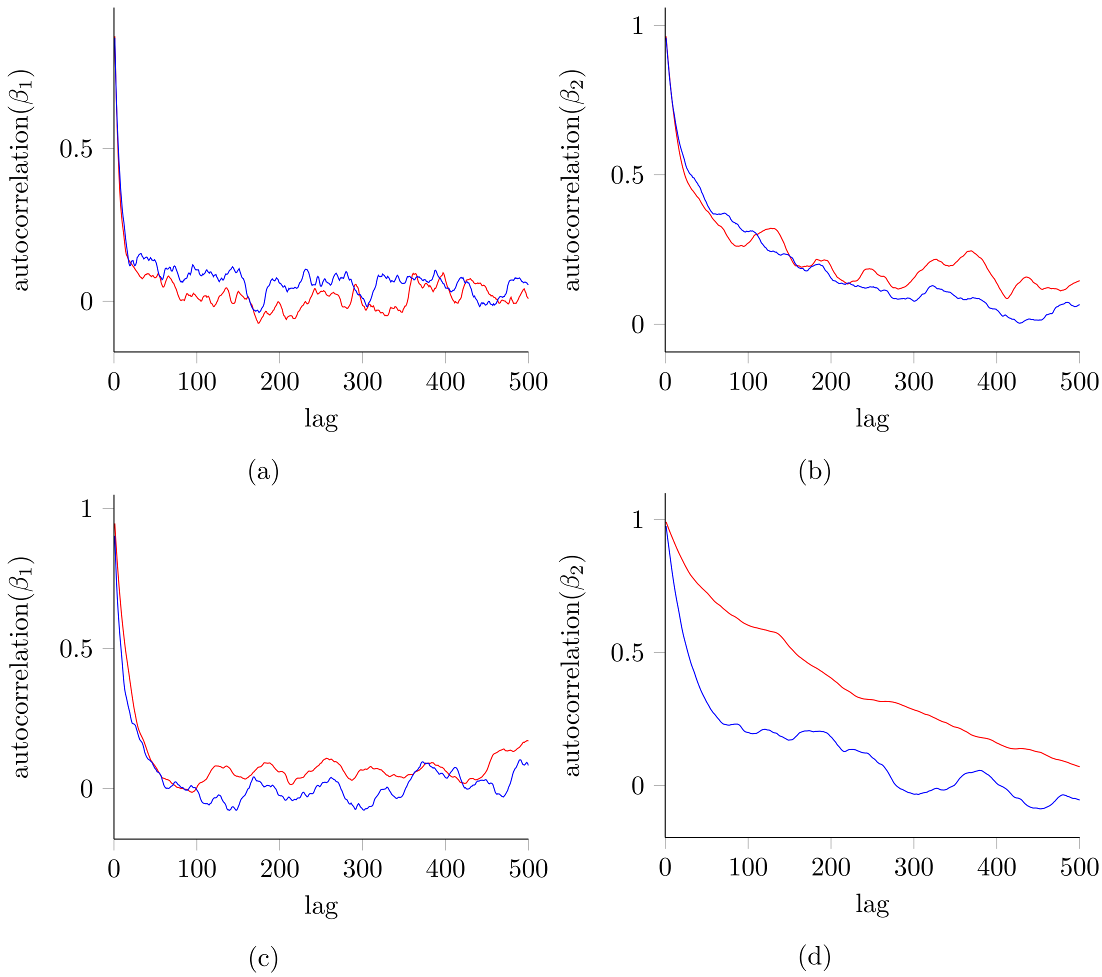


Figure 10. Autocorrelation plots for β_1 (left) and β_2 in the full model for both the WN (top row) and the WC sampler.

row). The autocorrelation for Unemp rapidly decreases to zero for both samplers, following the ideal pattern. For the autocorrelation of Outofwed, we still observe a decrease to zero, but at a slower rate, showing a somewhat more problematic pattern. In general, autocorrelations do not affect the parameter estimates as such, but can result in a underestimation of their precision. In some cases autocorrelation of the sampler can be reduced by centering the parameters. This was tried, but turned out not to be of help in our case. Another possibility is to fit a model with less parameters. It is possible to correct for autocorrelation by *thinning*.

In both samplers we experienced convergence problems. While for some of the parameters the sampler (WN or WC) converged, for many it did not. Problems ranged from chains that only partly overlapped to chains that did not overlap at all. In Figure 11 we show the trace plots for β_2 (left, no overlap between chains) and β_7 (right, some overlap) both resulting from the WC sampler.

Convergence problems make the sampling results unreliable and unsuitable for making inferences. The cause of the problems is not clear. One of the issues could be the relatively high number of predictors in the model compared to the sample size. This is a known issue for regression in general. We also observe a high level of multicollinearity between the continuous variables (ranging from .62 between Unemp and Year and .97 between Year and YearSQ), which makes it unreasonable to put them all in one model.

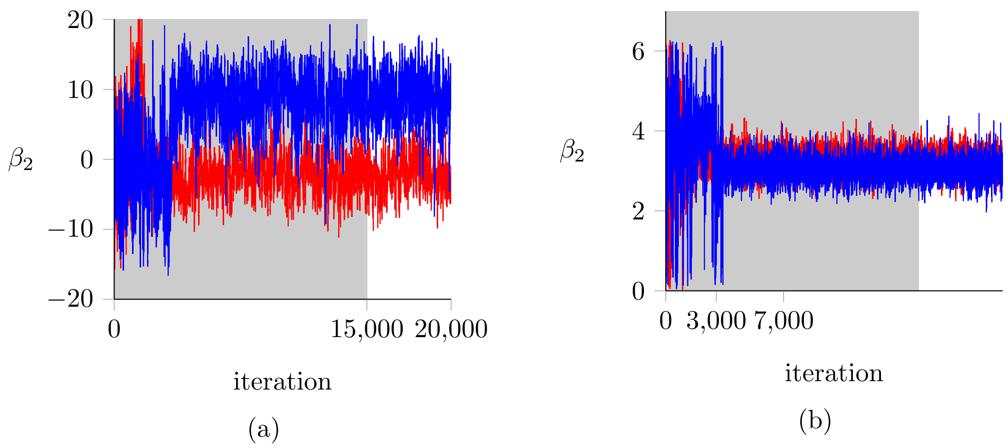


Figure 11. Trace plots of β_2 and β_7 in the full model for the WC sampler.

4.3 Univariate effects

To investigate, to some extent, what part of the problems is due to our method and what part can be attributed to the data, we decided to test all univariate effects. That is, six separate models were fit to the data: one model containing all party dummies and five models each having one of the other variables as predictor. All models were fitted with both WN and WC and a total of 7,000 iterations of which 3,000 were regarded as burn-in.

All models were checked for autocorrelation and convergence by means of autocorrelation/trace plots in the same way as shown above. We did not find problems with autocorrelation in any of the models. The models with YearSQ as a predictor, however, did not converge. For the WN sampler we also experienced convergence problems for the model with Year. In Table 6 the results are shown for models that did converge. Note that we do not report the intercepts here since our main interest is to check the effect of the predictor on preferred ideological direction. The lower and upper bound of the (circular) 95% CI are reported. We define a predictor to have an effect on the outcome variable, if this interval does *not* contain zero. Given this definition, we find an effect of the party dummies based on both methods and an effect of Year based on the WC method. Note that the parameter estimates for those variables that had an effect are comparable between WN and WC. The interval widths based on WC, however, seem to be somewhat narrower. On the other hand, we observe differences between WN/WC in the parameter estimates of the other variables. This is probably because there is no effect, making the parameter estimates a random deviation from zero. Our conclusions are not entirely in line with Gill and Hangartner (2010), who also report effects of Unemp, Outofwed, Year and YearSQ.

4.4 Final model

Based on our findings with the univariate models, we fitted a final restricted model including the party dummies. Although the WN model for Year did not converge, we still decided to keep it in the restricted model since an effect of Year was observed with the WC sampler. All other variables were left out of the model because we did not find univariate effects. The final restricted model can be written as:

$$\text{Ideal direction} = \beta_0 + \beta_1 \cdot \text{SPD} + \beta_2 \cdot \text{CDU/CSU} + \beta_3 \cdot \text{Greens} + \beta_4 \cdot \text{PDS} \ + \beta_5 \cdot \text{Year} + \epsilon. \quad (12)$$

This model again was fit with both WN and WC using 2 chains. Only a total of 7,000 iterations were needed of which the first 3,000 were regarded as burn-in. Note that we still regard the FDP party as the reference category.

We did not observe autocorrelation problems in the final model. The convergence in both samplers was good except for Year in WC, even when it was run with more iterations[5]. Therefore for WC, we decided to fit a model only including the party dummies (e.g. leaving out β_5 in Equation 12). Final results for both samplers can be found in Table 7.

We find an effect of the party dummies, which can be interpreted as follows. The intercept in the final model (β_0 in Equation 12) is 164.4° for WN and 171.3° for WC. The intercept can be interpreted as the averaged (over all elections) preferred ideological direction of the FDP, that is, the reference party (moderated by Year for the WN model). In Figure 9 this value would lie in the top left quadrant. The other parameters then can be viewed as the party's ideal direction relative to that of the FDP. For example, to find the preferred ideological direction of SDP for WC, we rotate another is 79.6° counter-clockwise relative to FDP. This makes a total angle of 252.9°, lying in the bottom left quadrant of Figure 9. Based on Table 7, we can conclude that all parties significantly differ in ideal direction from the FDP. Furthermore, if we calculate the average outcome for all parties, we find a good similarity with the reported means in Table 4. This is as expected. Our inferences with regard to the party differences are in line with Gill and Hangartner (2010). The exact estimates are hard to compare, as Gill and Harngartner only reported estimates based on the full model.

In the final model, we did not find an effect of the continuous variable Year. To show how such an effect could be interpreted, we look at the univariate effect of Year. In Table 6 a negative slope of -359.6 degrees per year was reported, while the intercept in this model equalled 204°. In this case, the intercept should be interpreted as the preferred ideological direction in 1949 (e.g. the first year in this dataset), averaged over all parties. For every year after 1949, then, the preferred ideological direction (averaged over all parties) is shifted 359.6° clockwise or 360-359.6=0.4° *counter*-clockwise. This results in a average direction of $204 + 56 \cdot 0.4 = 226.4°$ in 2005 (the last year in the dataset). Averaged over all parties, throughout the years the preferred ideological direction thus remained in the bottom left quadrant of Figure 9.

4.5 Conclusion

In conclusion, we saw that it was hard to fit to the full model from Gill and Hangartner (2010) using our methods due to autocorrelation and convergence problems. In part, this can be regarded as a general issue with our methods, but is should be noted that this dataset was not ideal (t.i., a high number of variables compared to the sample size and a high level of multicollinearity). Finally, we were able to fit a restricted model. We found that the resulting parameter estimates were easy to interpret and that our inferences were for an important part in line with Gill and Hangartner.

5 Discussion

In this paper we proposed a novel and promising alternative for circular regression using the wrapping approach. We developed models based on both the WN and WC distribution. Instead of using the data augmentation approach as adopted by Coles (1998) and Ravindran (2002), we based our techniques on the truncated version of the WN distribution as proposed by Kurz et al. (2014) and the closed form of the WC distribution (Mardia & Jupp, 1999). Where previous studies regarding circular regression based on the wrapping approach only focussed on mean estimation or models with one predictor (Ravindran, 2002; Ferrari, 2009), we increased the scope by making our methods applicable to models with an arbitrary number of continuous and dichotomous predictors.

[5]Note that this is different from our findings with the univariate models, where it was the WN sampler that did not converge for Year.

Previous methods were only tested in rather limited simulation settings only covering concentrated data (see f.e.: Ravindran (2002)). Under these *ideal* conditions our methods turn out to work equally well. However, in our simulations studies we also addressed less ideal and probably more realistic cases, thus investigating the practical boundaries of our methods in particular and the wrapping approach in general.

We showed that when continuous predictors are present, due to the nature of wrapping, it can be that multiple models fit the data equally well (see Figure 4). This results in a posterior distribution with multiple local maxima. In our simulations we guided our sampler in the right direction by choosing a relatively small value of the prior variance for slopes as was previously proposed by Coles (1998), citeAferrari2009wrapping and Ravindran (2002). However, for more dispersed data (e.g., smaller values of ρ), this turned out not to be enough to prevent all convergence problems. Especially the WC sampler seems to be sensitive for “jumping” to other local maxima. With respect to the convergence problems, we argue that it would be better to develop a sophisticated slice sampling scheme in which the absolute maximum of the posterior distribution is favoured over its local maxima. That is, the sampler should have a bias towards the *simplest* model. We are confident that this should solve problems with convergence.

We investigated some of the problems a researcher can experience when using the method in practice. That is, multicollinearity for models with multiple continuous predictors or unequal group sizes in models containing one or more dichotomous predictors. For both issues we saw that our methods still gave reliable parameter estimates. However, when the deviation in the data from the ideal situation became severe, a decrease in precision was observed. Note that this is in line with expectation; in linear regression a similar pattern can be observed. We concluded that the minimal sample size needed to obtain viable results depends on the circular concentration of the residuals. In concentrated data a sample size of 15 already could be sufficient. However, in general, we would recommend a minimum of 30. Note that these findings are based on models with two predictors. More severe problems may arise in more extensive models. These should be included in future simulation studies.

Both methods were applied to a dataset on preferred ideological direction. We experienced difficulties fitting a full model. However, it should be noted that this model could not serve as an ideal candidate for circular regression since it was based on a high number of (mutually correlated) variables. Based on this example, once more the importance of using trace plots as a diagnostic tool was illustrated. Finally, we were able to fit a restricted model to the dataset. We found that the resulting parameter estimates were easy to interpret.

Although we made some efforts in investigating the influence of wrongly distributed residuals on the sampling results, this could be further investigated in future research. Furthermore, the influence of the prior distributions could be an interesting topic for further investigation. Another interesting opportunity for future research would be to make our methods applicable to more complex regression models, such as multi-level data or time-series. In previous studies Bayesian analysis of complex regression models was based on the embedding approach (see f.e.: Wang and Gelfand (2013); Nuñez-Antonio and Gutiérrez-Peña (2014)) To our knowledge, however, multi-level Bayesian circular regression based on the wrapping approach is still non-existent.

In summary, this paper has introduced an adjusted and a more extensive method for Bayesian regression using the wrapping approach. We provided an extensive simulation study showing that results are reliable as long as the trace plots show convergence. We showed that our methods are flexible and can be applied to a wide range of data resulting in parameter estimates that are easy to interpret.

References

Abuzaid, A., & Allahham, N. (2015). Simple circular regression model assuming wrapped cauchy error. *Pakistan Journal of Statistics*, *31*(4), 385–398. doi: https://www.researchgate.net/publication/280383053

Bhattacharya, S., & SenGupta, A. (2009). Bayesian inference for circular distributions with unknown normalising constants. *Journal of statistical planning and inference*, *139*(12), 4179–4192.

Bickel, D. R., & Frühwirth, R. (2006). On a fast, robust estimator of the mode: comparisons to other robust estimators with applications. *Computational Statistics & Data Analysis*, *50*(12), 3500–3530. doi: http://doi.org/10.1016/j.csda.2005.07.011

Brunyé, T. T., Burte, H., Houck, L. A., & Taylor, H. A. (2015). The map in our head is not oriented north: evidence from a real-world environment. *PLOS ONE*, *10*(9), e0135803. doi: 10.1371/journal.pone.0135803

Coles, S. (1998). Inference for circular distributions and processes. *Statistics and Computing*, *8*(2), 105-113. doi: http://dx.doi.org/10.1023/A:1008930032595

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.

Ferrari, C. (2009). *The Wrapping Approach for Circular Data Bayesian Modeling* (Unpublished doctoral dissertation). University of Bologna.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data.* Cambridge: Cambridge University Press. doi: 10.1002/bimj.4710380307

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. doi: 10.1093/pan/mpq009

Herrera, P., Resa, Z., & Sordo, M. (2010). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. In *1st workshop on music recommendation and discovery (womrad), acm recsys, Barcelona, Spain.*

Jona-Lasinio, G., Gelfand, A., & Jona-Lasinio, M. (2013, January). Spatial analysis of wave direction data using wrapped Gaussian processes. *ArXiv e-prints.* doi: http://adsabs.harvard.edu/abs/2013arXiv1301.1446J

Kimpton, A., Corcoran, J., & Wickes, R. (2015). Public greenspaces and crime: an analysis of crime timing and public greenspace amenities. In *State of australian cities conference.*

Kurz, G., Gilitschenski, I., & Hanebeck, U. D. (2014). Efficient evaluation of the probability density function of a wrapped normal distribution. In *Sensor Data Fusion: Trends, Solutions, Applications (SDF), 2014* (p. 1-5). IEEE. doi: 10.1109/SDF.2014.6954713

Mardia, K. V., & Jupp, P. E. (1999). *Directional Statistics.* New York: Wiley. doi: 10.1002/9780470316979

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, *31*(3), 705-741. doi: http://www.jstor.org/stable/3448413

Nuñez-Antonio, G., & Gutiérrez-Peña, E. (2005a). A bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, *32*(10), 995-1001. doi: 10.1080/02664760500164886

Nuñez-Antonio, G., & Gutiérrez-Peña, E. (2005b). A Bayesian analysis of directional data using the von mises–fisher distribution. *Communications in Statisticsal Simulation and Computation*, *34*(4), 989–999.

Nuñez-Antonio, G., & Gutiérrez-Peña, E. (2014). A bayesian model for longitudinal circular data based on the projected normal distribution. *Computational Statistics & Data Analysis*, *71*, 506 - 519. doi: https://doi.org/10.1016/j.csda.2012.07.025

Nuñez-Antonio, G., Gutiérrez-Peña, E., & Escarela, G. (2011). A bayesian regression model for circular data based on the projected normal distribution. *Statistical Modelling*, *11*(3), 185–201.

Ravindran, P. (2002). *Bayesian Analysis of Circular Data Using Wrapped Distributions* (Unpublished doctoral dissertation). North Carolina State University.

Ravindran, P., & Ghosh, S. K. (2011). Bayesian analysis of circular data using wrapped distributions. *Journal of Statistical Theory and Practice*, *5*(4), 547-561. doi: http://dx.doi.org/10.1080/15598608.2011.10483731

Rentsch, S., & Rand, M. K. (2014, 10). Eye-hand coordination during visuomotor adaptation with different rotation angles. *PLOS ONE*, *9*(10), 1-15. doi: https://doi.org/10.1371/journal.pone.0109819

Von Mises, R. (1918). Uber die ganzzahligkeit der atomgewichte und verwandte fragen. *Physikalische Zeitschrift*, *19*, 490–500.

Wang, F., & Gelfand, A. E. (2013). Directional data analysis under the general projected normal distribution. *Statistical Methodology*, *10*(1), 113-127.

Table 1

Results of the simulation with one dichotomous and one continuous predictor for $\beta_0 = \beta_1 = 3.0$

True value of:			Abs(Bias)				IW				CP				
β_2	ρ	Sampler	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	Convergence
-5	0.5	WN	.15	.05	5.13	.38	1.41	1.79	2.46	.31	.96	.96	.95	.94	86
		WC	.02	.58	4.77	.70	1.28	1.62	2.48	.33	.97	.96	.95	.94	82
	0.9	WN	.16	.15	.11	.04	.41	.54	.28	.08	.95	.96	.95	.95	100
		WC	.08	.08	.03	.04	.15	.19	.10	.08	.95	.95	.95	.95	100
-1	0.5	WN	.05	.17	2.32	.79	1.32	1.70	1.62	.30	.96	.96	.96	.96	90
		WC	.03	.13	.46	.56	1.28	1.62	2.07	.32	.95	.94	.94	.98	83
	0.9	WN	.00	.08	.77	.07	.41	.54	.27	.08	.96	.95	.94	.94	100
		WC	.00	.10	.06	.42	.15	.20	.11	.08	.94	.95	.95	.95	100
0	0.5	WN	.71	.36	.00	1.28	1.40	1.78	1.88	.31	.93	.94	.94	.94	87
		WC	.17	.39	.00	3.05	1.37	1.72	2.39	.33	.93	.94	.94	.94	76
	0.9	WN	.20	.22	.00	.00	.41	.54	.28	.08	.93	.95	.94	.94	100
		WC	.01	.01	.00	.51	.15	.20	.10	.08	.95	.94	.95	.94	100
1	0.5	WN	.22	.32	2.60	.41	1.33	1.73	1.64	.30	.94	.94	.94	.95	91
		WC	.14	.25	6.26	3.2	1.38	1.74	2.48	.33	.94	.94	.94	.98	59
	0.9	WN	.05	.24	.16	.02	.41	.54	.27	.08	.95	.95	.93	.93	100
		WC	.01	.05	.07	.25	.15	.19	.10	.08	.95	.95	.95	.96	100
5	0.5	WN	.02	.33	3.27	1.52	1.34	1.74	2.09	.30	.95	.95	.96	.94	88
		WC	.04	.08	6.11	2.09	1.36	1.70	2.84	.33	.95	.96	.96	.92	79
	0.9	WN	.13	.27	.05	.01	.41	.54	.27	.08	.96	.95	.97	.96	100
		WC	.12	.11	.11	.42	.15	.19	.10	.08	.95	.96	.98	.94	100

Note. The bias shown is the absolute value of the relative bias (%), for $\beta_2 = 0$ the absolute value of the absolute bias is given.
Interval width (IW) concerns the 95% CI for β_0, β_1 and β_2, and HPD interval for ρ.
Convergence is the number of the first 100 samples that converged within the burn-in period of 3000.
CP is the coverage probability.

Table 2

Results for the simulation with two continuous predictors for $\beta_0 = 3.0, \beta_1 = 1.0$

True value of:			Abs(Bias)				IW				CP				
β_2	ρ	Sampler	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	Convergence
-5	0.5	WN	.71	8.73	7.44	.94	1.04	2.72	2.89	.31	.93	.95	.94	.95	90
		WC	.02	13.83	14.71	4.68	1.19	4.33	4.73	.35	.95	.95	.96	.87	69
	0.9	WN	.09	.02	.01	.34	.26	.28	.28	.08	.95	.95	.96	.94	100
		WC	.02	.10	.02	.11	.09	.10	.10	.08	.94	.96	.96	.95	100
-1	0.5	WN	.15	6.07	8.12	.60	1.04	2.51	2.52	.31	.95	.97	.96	.93	85
		WC	.02	9.39	13.23	1.40	.98	3.03	3.13	.34	.95	.95	.96	.91	78
	0.9	WN	.11	.37	.77	.01	.27	.28	.27	.08	.97	.95	.94	.95	100
		WC	.02	.16	.14	.30	.09	.10	.10	.08	.97	.95	.96	.96	100
0	0.5	WN	.65	7.56	.01	.19	1.10	2.77	2.94	.31	.97	.96	.96	.93	82
		WC	.15	14.67	.00	.91	1.00	3.09	3.12	.34	.93	.94	.96	.91	83
	0.9	WN	.16	.49	.00	.10	.27	.27	.27	.08	.96	.96	.95	.95	100
		WC	.05	.52	.01	.51	.10	.10	.10	.08	.92	.95	.94	.95	100
1	0.5	WN	.13	8.29	3.83	.61	1.03	2.60	2.66	.31	.97	.97	.96	.93	89
		WC	.13	9.89	10.12	2.83	1.07	3.40	3.47	.34	.96	.96	.96	.90	73
	0.9	WN	.13	.37	.17	.05	.27	.27	.27	.08	.95	.95	.95	.96	100
		WC	.00	.13	.18	.18	.09	.10	.10	.08	.94	.95	.95	.96	100
5	0.5	WN	.16	1.33	1.21	1.91	1.24	3.70	3.99	.32	.96	.97	.97	.92	86
		WC	.08	4.76	12.46	3.01	1.10	3.79	4.12	.34	.95	.95	.96	.89	74
	0.9	WN	.09	.05	02	.14	.26	.27	.28	.08	.95	.95	.93	.94	100
		WC	.09	.04	.01	.37	.10	.11	.11	.08	.96	.96	.95	.96	100

Note. The bias shown is the absolute value of the relative bias (%), for $\beta_2 = 0$ the absolute value of the absolute bias is given.
Interval width (IW) concerns the 95% CI for β_0, β_1 and β_2, and HPD interval for ρ.
Convergence is the number of the first 100 samples that converged within the burn-in period of 3000.
CP is the coverage probability.

Table 3

Regression results for the two types of models and the two types of residuals under the wrong assumption about the distribution of the error terms.

		Abs(Bias)			Bias	IW				CP				
Model	Residuals	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	Convergence
Ancova	WN	.03	.22	.07	-15.41	.41	.54	.27	.16	.90	.90	.89	.00	100
	WC	.08	.17	.11	- 3.57	.48	.64	.32	.10	.97	.98	.96	.47	100
LP	WN	.02	.05	.20	-15.23	.26	.27	.27	.16	.89	.90	.92	.00	100
	WC	.03	.13	.23	- 3.70	.31	.32	.32	.11	.97	.95	.96	.45	100

Note. The bias shown is the absolute value of the relative bias (%), except for ρ where the relative bias is reported.
Interval width (IW) concerns the 95% CI for β_0, β_1 and β_2, and HPD interval for ρ.
Convergence is the number of the first 100 samples that converged within the burn-in period of 3000.
CP is the coverage probability.

Table 4

Circular mean of preferred ideological direction in Gill and Hangartner (2010) (overall and per party).

	N	$\mu(^\circ)$	ρ
Overall	**60**	**221**	**.43**
FDP	16	164	.91
SPD	16	250	.87
CDU/CSU	16	359	.54
Greens	7	217	.99
PDS	5	225	.99

Table 5

Descriptive statistics of the predictor variables in Gill and Hangartner (2010).

	Type	Mean	SD	Min	Max
Reunification	Dichotomous	.42	.50	-	-
Unemp	Continuous	.07	.04	.01	.13
Outofwed	Continuous	.11	.06	.05	.23
Year	Continuous	32.18	17.37	.00	56.00
$\text{Year}^2/100$	Continuous	13.32	10.51	.00	31.36

Table 6

Effects on preferred ideological direction in all univariate models.

			Parameter estimate		95% CI	
Method	Model		Mean	SD	LB	UB
WN	M1: **Party**	**FDP**	164.3	.19	142.7	185.6
		SPD	85.4	.27	56.1	116.3
		CDU/CSU	188.5	.29	158.1	224.0
		Greens	52.1	.34	14.3	90.5
		PDS	60.2	.38	18.3	103.7
	M2: **Reunification**		17.8	.89	319.4	70.3
	M3: Unemp		-6.7	225.48	-456.8	416.9
	M4: Outofwed		56.6	197.98	-354.9	434.9
	M5: Year			*No convergence*		
	M6: YearSQ			*No convergence*		
WC	M1: **Party**	**Intercept**	171.3	.10	158.7	182.2
		SPD	79.6	.19	57.9	99.7
		CDU/CSU	191.9	.23	160.4	213.7
		Greens	45.8	.15	29.2	63.0
		PDS	53.3	.17	35.0	73.3
	M2: **Reunification**		10.2	.94	329.5	50.2
	M3: Unemp		74.2	213.35	-6.5	8.1
	M4: Outofwed		112.8	154.18	-202.6	413.1
	M5: Year		-359.6	0.55	-360.7	-358.5
	M6: YearSQ			*No convergence*		

Note. For dichotomous variables (boldface) the circular mean and standard deviation are reported, together with the lower bound (LB) and upper bound (UB) of the circular 95% CI. For the linear parameters the mean, the standard deviation and the lower and upper bound of the 95% CI are shown.

Table 7

Summary statistics of the regression parameters in the final model.

Method		Parameter estimate			Quantiles	
		Mean	Mode	SD	LB	UB
WN	**Intercept**	160.4		.29	129.5	195.0
	SPD	86.9		.26	56.6	116.1
	CDU/CSU	190.8		.29	159.7	226.8
	Greens	51.3		.36	11.3	91.1
	PDS	60.4		.39	16.4	103.7
	Year	.1		.41	- .8	.9
	ρ		.76		.70	.83
WC	**Intercept**	171.3		.10	158.7	182.2
	SPD	79.6		.19	57.9	99.7
	CDU/CSU	191.9		.23	160.4	213.7
	Greens	45.8		.15	29.2	63.0
	PDS	53.3		.17	35.0	73.3
	ρ		.75		.66	.82

Note. For dichotomous variables (boldface) the circular mean and standard deviation are reported, together with the lower bound (LB) and upper bound (UB) of the circular 95% CI. For the linear parameters the mean, the standard deviation and the lower and upper bound of the 95% CI are shown. For ρ we report its mode and the 95% HPD interval.

Appendix

Slice sampling

A1 Conceptual introduction

In this appendix we provide a short and conceptual introduction to slice sampling Neal (2003). This technique can be used to obtain random samples from all (multivariate) probability distributions of which a pdf is known that is proportional to the density of the distribution itself. This way, it is possible to sample from non-standard probability distributions.

For a random variable x having a pdf proportional to $f(x)$, the first few steps of obtaining a random sample in are depicted in Figure A1. First, a starting value x_0 is chosen and the density $f(x_0)$ at this point is calculated (Figure A1a). The next step is to uniformly sample an auxiliary variable d from the interval $[0, f(x_0)]$. We define a slice S containing all values of x for which $d < f(x)$ (Figure A1b). From this slice, a next value x_1 is uniformly sampled, which then will be used to obtain a new value of d resulting in a new slice S (Figure A1c).

In Figure A2, three samples of increasing size, sampled from a standard normal distribution by means of slice sampling, are depicted. On the top row the pairs (x_n, d_n) are plotted together with the probability density function. On the bottom row, the obtained random samples are depicted as histograms. We observe that if the process as described above is repeated sufficient many times, the resulting sample indeed resembles the pdf.

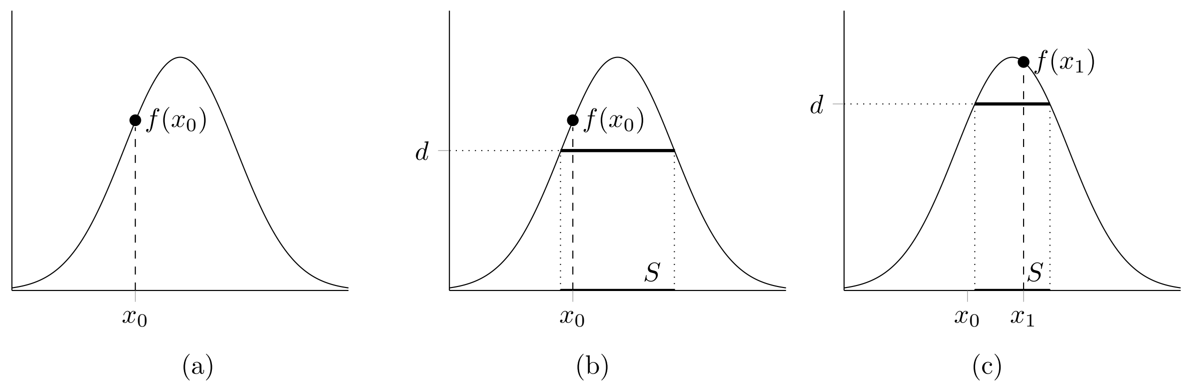


Figure A1. The first few steps in slice sampling a standard normal distribution.

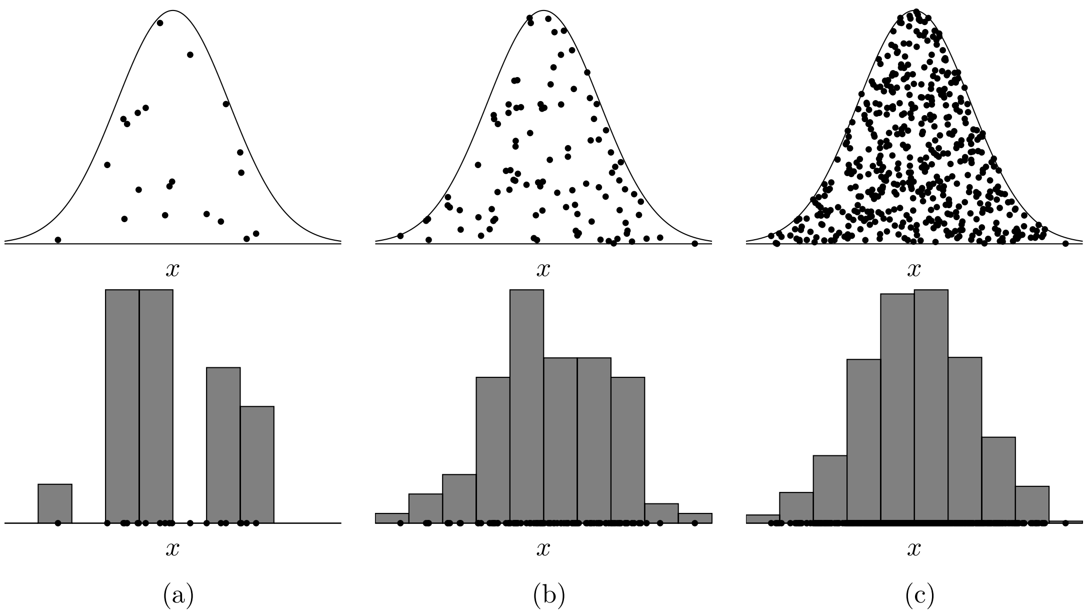


Figure A2. Results of sampling respectively 20, 100 and 500 points using slice sampling.

A2 Multivariate slice sampling scheme to sample from the joint distribution

In this appendix we will derive a multivariate slice sampling scheme for a model with one circular outcome, an intercept and one continuous predictor. The rules as described below are also applicable to more advanced models, where model parameters of extra dichotomous predictors can be sampled like intercepts and model parameters belonging to extra continuous predictors can be sampled like slopes.

Definition of starting state. The first step in the slice sampling scheme is to define a *starting state* $\mathbf{s}_0$, or, stated differently, to assign starting values to the model parameters:

$$\mathbf{s}_0 = \begin{bmatrix} \beta_{0_0} & \in & [0, 2\pi] \\ \beta_{1_0} & \in & [0, 2\pi] \\ \beta_{2_0} & \in & \mathbb{R} \\ \rho_0 & \in & [0, 1] \end{bmatrix}. \tag{13}$$

Updating the auxiliary variable. We define the joint density at the current state of the sampler $\mathbf{s}_{cur}$ (in the initial step $\mathbf{s}_{cur} = \mathbf{s}_0$) for wrapped Cauchy distributed errors:

$$\mathrm{J}_{\mathrm{WC}}(\mathbf{s}_{cur}) \propto \frac{\exp\left(-\frac{\beta_{1_{cur}}^2}{2\sigma_\beta^2}\right) \cdot (\rho_{cur})^{-0.5}(1-\rho_{cur})^{n-0.5}(1+\rho_{cur})^n}{\prod_{i=1}^n \{1 + \rho_{cur}^2 - 2\rho_{cur}\cos(r_{i_{cur}})\}}, \tag{14}$$

And for WN (note: $\sigma = \sqrt{-2\log(\rho)}$):

$$\mathrm{J}_{\mathrm{WN}}(\mathbf{s}_{cur}) \propto \prod_{i=1}^{n} \left\{ \sum_{k=-1}^{1} \exp\left(\frac{(r_{i_{cur}} + 2\pi k)^2}{4\log(\rho_{cur})}\right) \right\} \cdot \exp\left(-\frac{\beta_{1_{cur}}^2}{2\sigma_\beta^2}\right) \cdot (\rho_{cur})^{-0.5}(1-\rho_{cur})^{-0.5} \cdot (-\log \rho_{cur})^{-\frac{n}{2}}, \tag{15}$$

where:

$$r_{i_{cur}} = y_i - (\beta_{0_{cur}} + \beta_{1_{cur}} x_i) \mod 2\pi, \tag{16}$$

for both WC and WN.

Next, we sample a value z from a standard uniform distribution:

$$z \sim U(0,\ 1). \tag{17}$$

Using f we can update the auxiliary variable d as follows:

$$d = \begin{cases} z \cdot \mathrm{J}_{\mathrm{WN}}(\mathbf{s}_{cur}), & for\ \text{WN} \\ z \cdot \mathrm{J}_{\mathrm{WC}}(\mathbf{s}_{cur}), & for\ \text{WC} \end{cases}. \tag{18}$$

Updating intercepts. We draw a proposal $\beta_{0_{prop}}$ value for the intercept from its domain $[0, 2\pi]$:

$$\beta_{0_{prop}} \sim U(0,\ 2\pi). \tag{19}$$

We use the following acceptance rule to decide whether the proposal value is accepted. That is, we check if, given the current model parameters, the proposal value lies within the slice.

For WC:

$$d < \mathrm{J}_{\mathrm{WC}}(\beta_{0_{prop}} \mid \beta_{1_{cur}},\ \rho_{cur})$$
$$\frac{1}{z}\prod_{i=1}^{n}\{1+\rho_{cur}^2 - 2\rho_{cur}\cos(r_{i_{cur}})\} > \prod_{i=1}^{n}\{1+\rho_{cur}^2 - 2\rho_{cur}\cos(y_i - \beta_{0_{prop}} - \beta_{1_{cur}}x_i \mod 2\pi)\}. \tag{20}$$

For WN:

$$d < \mathrm{J}_{\mathrm{WN}}(\beta_{0_{prop}} \mid \beta_{1_{cur}},\ \rho_{cur})$$
$$z \cdot \prod_{i=1}^{n}\sum_{k=-1}^{1}\exp\left(\frac{(r_{i_{cur}}+2\pi k)^2}{4\log(\rho_{cur})}\right) < \prod_{i=1}^{n}\sum_{k=-1}^{1}\exp\left(\frac{(\{y_i - \beta_{0_{prop}} - \beta_{1_{cur}}x_i \mod 2\pi\} + 2\pi k)^2}{4\log(\rho_{cur})}\right). \tag{21}$$

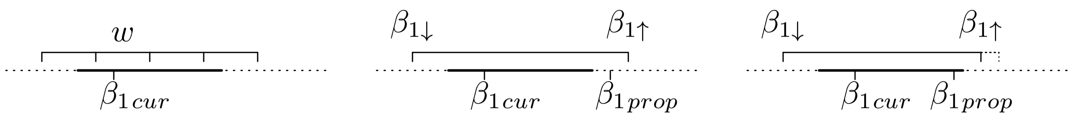
This process is repeated until a proposal value for the intercept is found that is within the slice. In that case the current state of the sampler is updated using this proposal value.

Updating slopes. The domain of a slope parameter, in this case β_1, is not bounded, therefore we cannot directly sample a proposal value from this domain. We use a *stepping out procedure* (Neal, 2003, p. 713) to find a bounded domain $[\beta_{1\downarrow},\ \beta_{1\uparrow}]$ that contains S. That is, an interval of fixed size w is placed randomly around the current value of the slope $\beta_{1_{cur}}$. Each side of the interval is then extended in steps of size w until its border lie outside the slice (see Figure A3a).

Next, we draw a proposal β_{1prop} value for the slope from $[\beta_{1\downarrow},\ \beta_{1\uparrow}]$ (see Figure A3b):

$$\beta_{1_{prop}} \sim U(\beta_{1\downarrow},\ \beta_{1_\uparrow}). \tag{22}$$

We use the following acceptance rule to decide whether the proposal value is accepted. That is, we check if, given the current model parameters, the proposal value lies within the slice.



(a) (b) (c)

Figure A3. Stepping out and shrinkage procedure.

For WC:

$$d < \mathrm{J}_{\mathrm{WC}}(\beta_{1_{prop}} \mid \beta_{0_{prop}},\ \rho_{cur})$$

$$z \cdot \frac{\exp\left(-\frac{\beta_{1_{cur}}^2}{2\sigma_\beta^2}\right)}{\prod_{i=1}^n \{1+\rho_{cur}^2 - 2\rho_{cur}\cos(r_{i_{cur}})\}} < \frac{\exp\left(-\frac{\beta_{1_{prop}}^2}{2\sigma_\beta^2}\right)}{\prod_{i=1}^n \{1+\rho_{cur}^2 - 2\rho_{cur}\cos(y_i - \beta_{0_{cur}} - \beta_{1_{prop}}x_i \mod 2\pi)\}}. \tag{23}$$

For WN:

$$d < \mathrm{J}_{\mathrm{WN}}(\beta_{1_{prop}} \mid \beta_{0_{prop}},\ \rho_{cur})$$

$$z \cdot \prod_{i=1}^n \left\{ \sum_{k=-1}^{1} \exp\left(\frac{(r_{i_{cur}} + 2\pi k)^2}{4\log(\rho_{cur})}\right)\right\} \cdot \exp\left(-\frac{\beta_{1_{cur}}^2}{2\sigma_\beta^2}\right) < \prod_{i=1}^n \left\{ \sum_{k=-1}^{1} \exp\left(\frac{(\{y_i - \beta_{0_{cur}} - \beta_{1_{prop}}x_i \mod 2\pi\} + 2\pi k)^2}{4\log(\rho_{cur})}\right)\right\} \cdot \exp\left(-\frac{\beta_{1_{prop}}^2}{2\sigma_\beta^2}\right). \tag{24}$$

If the proposal value lies outside the slice, its value is used to *shrink* the interval $[\beta_{1\downarrow},\ \beta_{1\uparrow}]$ (see Figure A3b). And the process is repeated. If, on the other hand, the proposal value for the slope is within the slice (see Figure A3c), the current state of the sampler is updated using this proposal value.

Updating the circular concentration. We draw a proposal ρ_{prop} value for the circular concentration from its domain $[0, 1]$:

$$\rho_{prop} \sim U(0,\ 1). \tag{25}$$

We use the following acceptance rule to decide whether the proposal value is accepted. That is, we check if, given the current model parameters, the proposal value lies within the slice.

For WC:

$$d < \mathrm{J}_{\mathrm{WC}}(\rho_{prop} \mid \beta_{0_{prop}},\ \beta_{1_{cur}})$$

$$z \cdot \frac{(\rho_{cur})^{-0.5}(1-\rho_{cur})^{n-0.5}(1+\rho_{cur})^n}{\prod_{i=1}^n 1+\rho_{cur}^2 - 2\rho_{cur}\cos(r_{i_{cur}})} < \frac{(\rho_{prop})^{-0.5}(1-\rho_{prop})^{n-0.5}(1+\rho_{prop})^n}{\prod_{i=1}^n 1+\rho_{prop}^2 - 2\rho_{prop}\cos(r_{i_{cur}})}. \tag{26}$$

For WN:

$$d < \mathrm{J}_{\mathrm{WN}}(\rho_{prop} \mid \beta_{0_{prop}},\ \beta_{1_{cur}})$$

$$z \cdot \prod_{i=1}^n \left(\sum_{k=-1}^{1} \exp\left(\frac{(r_{i_{cur}} + 2\pi k)^2}{4\log(\rho_{cur})}\right)\right) \cdot (\rho_{cur})^{-0.5}(1-\rho_{cur})^{-0.5} \cdot (-\log\rho_{cur})^{-\frac{n}{2}} < \prod_{i=1}^n \left(\sum_{k=-1}^{1} \exp\left(\frac{(r_{i_{cur}} + 2\pi k)^2}{4\log(\rho_{prop})}\right)\right) \cdot (\rho_{prop})^{-0.5}(1-\rho_{prop})^{-0.5} \cdot (-\log\rho_{prop})^{-\frac{n}{2}}. \tag{27}$$

This process is repeated until a proposal value for the circular concentration is found that is within the slice. In that case the current state of the sampler is updated using this proposal value.