Monotonic Effects: A Principled Approach for Including Ordinal Predictors in Regression

Models

Paul-Christian Bürkner[1] & Emmanuel Charpentier[2]

[1] Department of Psychology, University of Münster, Germany

[2] Assistance publique - Hôpitaux de Paris, France

Author Note

Correspondence concerning this article should be addressed to Paul-Christian Bürkner,

Department of Psychology, University of Münster, Fliednerstrasse 21, 48149 Münster,

Germany. E-mail: paul.buerkner@gmail.com

10   Abstract

11   Ordinal predictors are commonly used in regression models. Yet, they are often incorrectly

12   treated as either nominal or metric thus under- or overestimating the contained information.

13   This is understandable insofar as generally applicable solutions or corresponding statistical

14   software are still underdeveloped. We propose a new way of parameterizing regression

15   coefficients of ordinal predictors, which we call monotonic effects. The reparameterization is

16   done in terms of a scale parameter b taking care of the direction and size of the effect and a

17   simplex parameter ζ modeling the normalized differences between categories. This ensures

18   that predictions are monotonotically increasing or decreasing, while changes between

19   adjacent categories may vary across categories. This formulation nicely generalizes to both

20   interaction terms as well as multilevel structures. Monotonic effects may not only be applied

21   to ordinal predictors, but also to other discrete variables for which a monotonic relationship

22   is plausible. This includes variables representing count data or discrete points in time.

23   Fitting motonotonic effects in a fully Bayesian framework is straightforward with the R

24   package brms, which also allows to incorporate prior information and to check the

25   assumption of monotonicity.

26   *Keywords:* Regression, Isotonic, Ordinal variables, Bayesian statistics, brms, Stan, R

Monotonic Effects: A Principled Approach for Including Ordinal Predictors in Regression Models

1 Introduction

Over the last few decades, much statistical research has been devoted to handling ordinal response variables in regression models starting with the seminal paper of McCullagh (1980; see also Agresti, 2010; Bürkner and Vuorre, 2018; Liu & Agresti, 2005; Tutz, 2011 for an overview). In Psychology, for instance, this kind of data is omnipresent in the form of Likert scale items, which are often treated as being continuous for convenience without ever testing this assumption (Liddell & Kruschke, 2017). With researchers realizing the importance of correctly modeling ordinal responses, the related models – often simply called *ordinal models* – are now increasingly applied in scientific practice. In the statistical language R (R Core Team, 2018), for instance, several packages are available to fit ordinal models, among others "ordinal" (Christensen, 2018), "VGAM" (Yee, Stoklosa, Huggins, & others, 2015), or "brms" (Bürkner, 2017b, 2017a) to name the perhaps most general ones.

Ordinal *predictors* seem to have received less attention in statistical research. There are only two lines of research known to the authors of the present paper. One is a penalized regression approach specifically designed for ordinal predictors (Gertheiss, 2014; Gertheiss & Tutz, 2009; Gertheiss, Oehrlein, & others, 2011b) and the other are categorcial types of isotonic regression (Barlow, Bremner, Brunk, & Bartholomew, 1972; Robertson, Wright, & Dykstra, 1988).

We begin by explaining the former approach. The main idea of the method of Gertheiss and Tutz (2009) is to penalize large differences between adjacent categories. This reflects the expectation that if a variable is ordinal, changes may happen somewhat smoothly and larger differences should thus be unlikely. This approach allows for a very flexible handling of ordinal predictors in a way closely related to regression splines (Gertheiss & Tutz, 2009). The direction of the changes remains unspecified and may vary across the range of the ordinal variable. Depending on the research question and variables under study, this

54 assumption might be somewhat too flexible. More specifically, we do often expect the

55 changes between adjacent categories to be *monotonic*, that is consistently negative or

56 positive across the full range of the ordinal variable. At the same time, the size of the

57 changes may still vary across categories by a substantial amount as ordinality does not

58 necessarily contain information about distance between categories.

59   The research on *isotonic regression* deals with regression models subject to order

60 contraints (Barlow et al., 1972; Robertson et al., 1988). For instance, the effect of a drug is

61 assumed to be monotonically[1] increasing with an increasing dose – an assumption that we

62 often want to "hard-code" into our models. Depending on the research question and nature

63 of the variable on which we want to impose a monotonicity constraint, different techniques

64 may be favorable. If the variable is essentially continuous, such as the dose of a drug, we can

65 use parametric functions which are known to be monotonic (e.g., the log or logistic functions

66 in simple cases) or use semi-parameteric approaches such as monotonic splines (Kelly & Rice,

67 1990; C.-I. C. Lee, 1996; Leitenstorfer & Tutz, 2007; Pya & Wood, 2015). If the variable

68 under study is categorical, the monotonicity assumption reduces to an ordering constraint on

69 the group means with respective to the response variable. From the perspective of classical

70 *frequentist* statistics, the latter case has been studied extensively in Barlow et al. (1972) and

71 Robertson et al. (1988; see also Best and Chakravarti, 1990; R. L. Dykstra & Robertson,

72 1982; C.-I. C. Lee, 1981; Wu, Woodroofe, & Mentz, 2001). For the purpose of studying

73 ordinal predictors, we are primarily interested in the categorical type of isotonic regression.

74   The idea and scope of the two approaches discuseed above is somewhat different.

75 While isotonic regression only restricts the *direction* of the changes, the penalized regression

76 method restricts the *size* of the changes leaving the direction untouched (although we may

77 also introduce order constraints in the latter according to Gertheiss et al., 2011b). None of

78 these two approaches is more reasonable per se and we should not necessarily think of them

---

[1] The term "isotonic" is mostly used synonymously to "monotonic" in the mathematical-statistical literature. We prefer the latter as we believe it to be understandable by a wider audience outside of mathematics.

79 as competing in how ordinal predictors should ideally be modeled. Rather, they make use of

80 different sets of assumptions which both reflect important aspects of ordinal variables. As we

81 will see later on, we may even combine them naturally within the same framework.

82   In the present paper, we will introduce a monotonicity imposing parameterization of

83 ordinal predictors, which we call *monotonic effects*. They represent a way to generalize the

84 assumption of linearity to ordinal predictors. As explained in detail in the next section, the

85 estimated parameters have an intuitive meaning and are thus easy to interprete and

86 communicate. In simple cases, they also turn out to be equivalent to the results of

87 categorical isotonic regression.

88   The structure of this paper is a follows. In Section 2, we will introduce monotonic

89 effects as well as their mathematical foundation in detail. We continue by explaining a

90 software implementation of monotonic effects in the R package "brms" (Bürkner, 2017b,

91 2017a), which supports a wide and growing range of Bayesian regression models. In Section

92 4, a case study dealing with measures of chronic widespread pain (Cieza et al., 2004;

93 Gertheiss, Hogger, Oberhauser, & Tutz, 2011a) will be discussed, in which we make extensive

94 use of monotonic effects. We end with a conclusion in Section 5. Some short mathematical

95 proofs about the properties of monotonic effects are presented in the Appendix.

2   Monotonic Effects

97   A predictor which we want to model as monotonic must be integer valued. The

98 integers may represent, for instance, count data, discrete points in time, or categories of an

99 ordinal variable. Since the latter is possibly the most relevant use case in psychology and

100 related disciplines, we speak of predictor categories in the following to refer to the values of a

101 monotonic predictor. As opposed to a continuous predictor, predictor categories are not

102 assumend to be equidistant with respect to their effect on the response variable. Instead, the

103 distance between adjacent predictor categories is estimated from the data and may vary

104 across categories. This is realized by parameterizing as follows: One parameter, b, takes care

of the direction and size of the effect similar to an ordinary regression parameter, while an

additional parameter vector, $\boldsymbol{\zeta}$, estimates the normalized distances between consecutive

predictor categories. For a single monotonic predictor, $\boldsymbol{x}$, the corresponding predictor term

η_n of observation n looks as follows:

$$\eta_n = b \sum_{i=1}^{x_n} \zeta_i \tag{1}$$

For notational convenience, we define $\sum_{i=1}^{0} \zeta_i = 0$. The parameter b can take on any

real value, while $\boldsymbol{\zeta}$ is a simplex, which means that is it satisfies $\zeta_i \in [0, 1]$ and $\sum_{i=1}^{C} \zeta_i = 1$

with C being the number of categories (highest possible integer). If the monotonic effect is

used in a linear model and the lowest category of $\boldsymbol{x}$ is 0, b can be interpreted as the expected

difference between the highest and the lowest category of $\boldsymbol{x}$, while ζ_i describes the expected

difference between the categories i and $i - 1$ in the form of a proportion of the overall

difference b. Thus, the above parameterization emits an intuitive intepretation while

guaranteeing the monotonicity of the effect (see Appendix A for a formal proof). For

notational convenience we define $\mathrm{mo}(x, \boldsymbol{\zeta}) = \sum_{i=1}^{x} \zeta_i$ and call $\mathrm{mo}()$ the *monotonic transform.*

As visualized in Figure 1, we can understand monotonic effects as implying a piecewise linear

curve of which all components have the same sign. In a simple linear model, monotonic

effects are equivalent to categorical isotonic regression (see Appendix A).

Interaction terms including a monotonic predictor $\boldsymbol{x}$ can be canonically written as

$$\eta_n = b \, z_n \, \mathrm{mo}(x_n, \boldsymbol{\zeta}_x) \tag{2}$$

where $\boldsymbol{z}$ is another predictor. If $\boldsymbol{z}$ is monotonic as well, then z_n is simply replaced by

$\mathrm{mo}(z_n, \boldsymbol{\zeta}_z)$. One modeling choice to be made is whether different terms including $\boldsymbol{x}$ should

have the same or different simplex parameters associated with $\boldsymbol{x}$. For example. a predictor

term consisting of the main effects and two-way interaction between a monotonic predictor $\boldsymbol{x}$

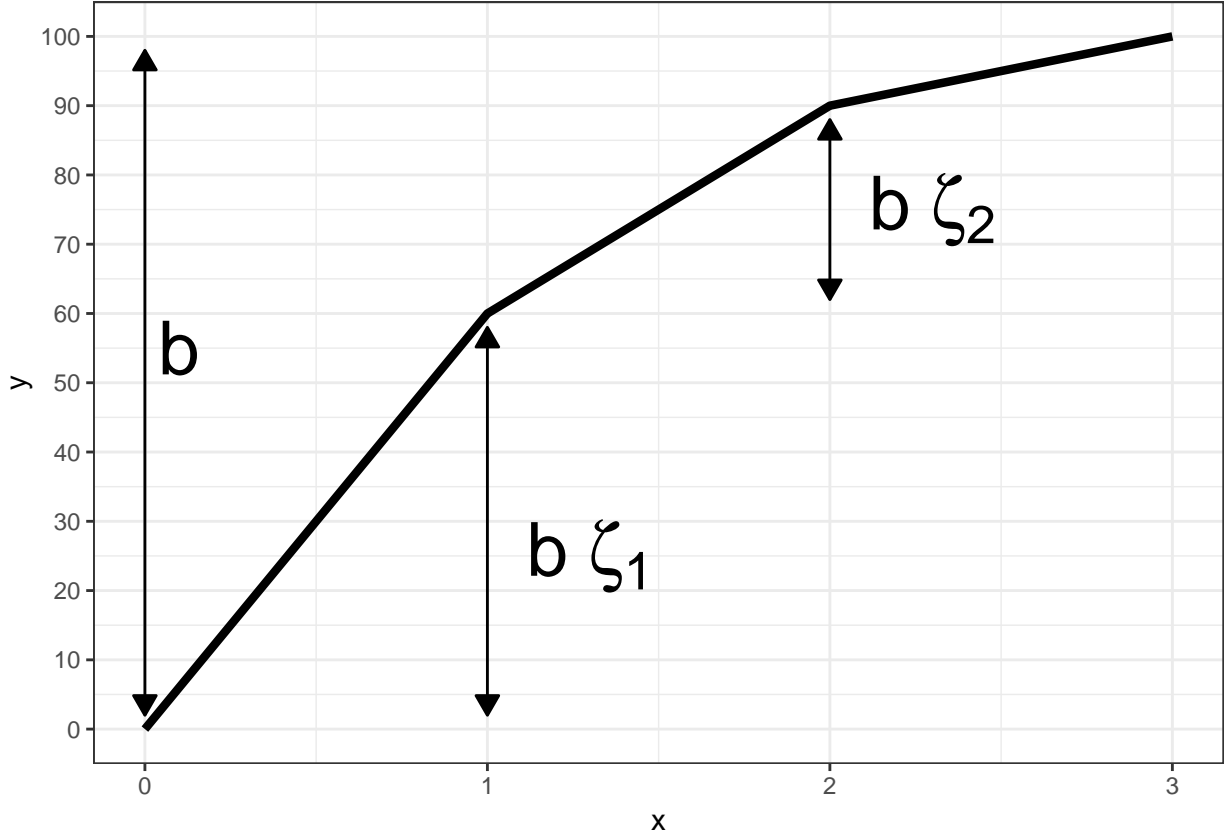and an arbitrary predictor $\boldsymbol{z}$ could be formulated as

Figure 1. Visualization of a monotonic effect with four categories. Parameters were set to $b = 100$ and $\boldsymbol{\zeta} = (0.5, 0.3, 0.2)$.

$$\eta_n = b_1 \, z_n + b_2 \, \mathrm{mo}(x_n, \boldsymbol{\zeta}_{xb_2}) + b_3 \, z_n \, \mathrm{mo}(x_n, \boldsymbol{\zeta}_{xb_3}), \tag{3}$$

127   where $\boldsymbol{\zeta}_{xb_2}$ and $\boldsymbol{\zeta}_{xb_3}$ are two independent simplex parameters. Under this formulation,

128    $\boldsymbol{x}$ may not necessarily be conditionally monotonic for all values of $\boldsymbol{z}$ (see Appendix A for a

129   counter example). Rather the monotoniciy being modeled depends on the chosen

130   parameteriziation. For instance, if the predictor $\boldsymbol{z}$ is dummy coded as 0 and 1 representing

131   the two categories of a dichotomous variable, the formulation above models the effect of $\boldsymbol{x}$ to

132   be monotonic for category 0 as well as for the *change* between category 1 and 0. Conversely,

133   when using cell mean coding rather than dummy coding for $\boldsymbol{z}$, the model assumes a different

134   monotonic effect of $\boldsymbol{x}$ for both categories of $\boldsymbol{z}$. In the latter case, $\boldsymbol{x}$ is conditionally

135   monotonic on $\boldsymbol{z}$. If we fix all simplex parameters corresponding to the same monotonic

136 variable $\boldsymbol{x}$ to the same value, conditionally monotoncity is achieved in general (proof

137 provided in Appendix A):

138 **Proposition 2.1.** Let $\boldsymbol{\eta}$ be an arbitrary linear predictor term containing the monotonic

139 predictor $\boldsymbol{x}$ with the corresponding simplex parameter $\boldsymbol{\zeta}$ being the same for all terms

140 including $\boldsymbol{x}$. Then $\boldsymbol{\eta}$ is monotonic in $\boldsymbol{x}$ conditionally on all possible combinations of all other

141 predictor variables.

142   While fixing all simplex parameters associated with $\boldsymbol{x}$ to the same vector quarantees

143 conditional monotonicity, it may be not flexible enough for many common situations. For

144 instance, if one wanted to model different monotonic effects for two groups, it would imply

145 the shape ($\boldsymbol{\zeta}$) of the predictions to be the same across groups with just their total range (b)

146 to be different. As explained in Section 3, in brms we make use of both parameterizations

147 (varying and constant $\boldsymbol{\zeta}$) at different places in the package.

148 2.1   Monotonic effects in a Bayesian framework

149   The present paper describes monotonic effects as embedded in a fully Bayesian

150 framework. We consider every statistical models a *Bayesian* model if it quantifies the

151 uncertainty in all observed and unobserved variables by means of probabilities. This is often

152 expressed in terms of Bayes' Theorem, which states that the posterior distribution $p(\theta|y)$ of

153 the model parameters θ given the data y can be expressed in terms of the product of

154 likelihood $p(y|\theta)$ and prior distribution $p(\theta)$ as well as a normalizing constant $p(y)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

155   A thorugh introduction to Bayesian statistics is outside the scope of the present paper.

156 Instead, we refer to well established text books such as McElreath (2016), Kruschke (2014),

157 and A. Gelman et al. (2013).

158   With respect to monotonic effects, a fully Bayesian framework has two main

159 implications. First, such a framework allows to incorporate monotonic effects in a large class

160 of regression models without the need to develop any problem specific estimators. Second, it

161 implies that we can think of prior distributions for b and $\boldsymbol{\zeta}$. Such prior distributions enable

162 us to incorporate information, which does not come directly from data in terms of the

163 likelihood contribution, such as expert knowledge or findings from previous studies.

164   Priors for b can be derived based on the a-priori expectation regarding the differences

165 between highest and lowest category, which we call *maximal difference* in the following. Any

166 family of prior distributions typically applied to regression coefficients can be applied on b,

167 as well. As a weakly-informative prior for b, we can understand any location shift

168 distribution – such as a normal of student-t distribution – centered around zero and with a

169 scale parameter large enough to allow for large but plausible maximal differences, while

170 penalizing unplausiblly large maximal differences. This scale will necessarily depend on the

171 scale of the response distribution and also on the range of the monotonic predictor.

172 Alternatively, one may use an improper flat prior that treats all real values as being equally

173 likely a-priori in the hope that the data alone is sufficient to identify b. We will come back to

174 this in the discussion of our case study.

175   Setting a prior on the simplex parameter $\boldsymbol{\zeta}$ requires a different approach. The

176 canonical prior of a simplex parameter is the Dirchlet distribution, a multivariate

177 generalization of the beta distribution (Frigyik, Kapila, & Gupta, 2010). It is non-zero for all

178 valid simplexes (i.e., for $\boldsymbol{\zeta}$ with $\zeta_i \in [0, 1]$ and $\sum_{i=1}^{C} \zeta_i = 1$) and zero otherwise. The Dirichlet

179 prior has a single parameter vector $\boldsymbol{\alpha}$ of the same length as $\boldsymbol{\zeta}$. Its density is defined as

$$f(\boldsymbol{\zeta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{C} \zeta_i^{\alpha_i - 1}, \tag{5}$$

180   where $B(\boldsymbol{\alpha})$ is a normalizing constant. As the a-priori expectation of ζ_i is given by

181 $w_i = \mathbb{E}(\zeta_i) = \alpha_i/\alpha_0$, with $\alpha_0 = \sum_{i=1}^{C} \alpha_i$, higher values of α_i in comparison to the sum over $\boldsymbol{\alpha}$

182 imply higher a-priori values of ζ_i. Moreover, the higher the sum over $\boldsymbol{\alpha}$, the higher the

183 certainty in each of the proportions w_i.

184   In the absense of any problem specific information, a reasonable default prior on $\boldsymbol{\zeta}$

would surely be one that assumed all differences between adjacent categories to be the same on average while being considerably uncertain about this expectation. Such a prior would imply, on average, a linear trend but with enough uncertainty to allow for all other possible monotonic trends as well. The Dirichlet prior with a constant $\boldsymbol{\alpha} = 1$ puts equal probability on all valid simplex and can thus be understood as the multivariate generalization of the uniform prior on simplexes. Since we have $w_i = 1/(C - 1)$, this prior centers $\boldsymbol{\zeta}$ around a linear trend with large uncertainty and thus appears to be a good default prior in the absense of any problem specific information.

2.2 Penalizing larger changes between categories

In a Bayesian framework, larger differences between adjacent categories can naturally be penalized by means of priors on b and $\boldsymbol{\zeta}$. If we expect the total effect b to be small, we can use a zero-centered prior on b with comparatively small tails. For instance, if we expect b to be between -10 and 10 with probability 95% as well as higher probability for values closer to zero, we can use a Normal$(0, 5)$ prior. The straightforward logic behind this prior is that the normal distribution has approximately 95% probability between -2 and 2 standard deviations around its mean.

When it comes to the shape of the monotonic effect, we have to take a closer look at the prior on $\boldsymbol{\zeta}$. As discussed above, a constant vector $\boldsymbol{\alpha}$ of the Dirichlet prior on $\boldsymbol{\zeta}$ implies a linear trend in expectation. In other words, for constant $\boldsymbol{\alpha}$, the prior means of all changes ζ_i between adjacent categories are the same. The higher the sum over $\boldsymbol{\alpha}$, the higher the certainty in that expectation. Thus, if we expect a linear trend with some certainty, we assign all elements of $\boldsymbol{\alpha}$ to the same value a. To get an intuition about what is a reasonable value for a, we may use the standard deviation of the elements ζ_i, which can be computed as

$$\mathrm{SD}(\zeta_i) = \sqrt{\frac{\alpha_i(\alpha_0 - \alpha_i)}{(\alpha_0^2(\alpha_0 + 1)}}. \tag{6}$$

Although the standard deviation is an imperfect measure of variability for the Dirichlet

209 distribution as the latter is not symmetric in general, we still believe the former to be helpful
210 in better understanding the implications of ones chosen priors. For the default of $a = 1$ and
211 a total of $C = 5$ categories, we get a rather large standard deviation of $\mathrm{SD}(\zeta_i) = 0.19$. If we
212 set, for example, $a = 5$, we get $\mathrm{SD}(\zeta_i) = 0.09$ and thus much higher certainty in changes of
213 equal size.

214 Of course, the process of increasing $\boldsymbol{\alpha}$ on average works equally well even if we do not
215 expect all changes to be the same a-priori. For instance, if $C = 5$ and we expect a 3-times
216 larger change between the first two categories than between all the other categories with
217 some certainty, we may set $\boldsymbol{\alpha} = (9, 3, 3, 3)$. As a result, we get $w_1 = 1/2$ and $w_i = 1/6$ else.
218 As standard deviations, we get $\mathrm{SD}(\zeta_1) = $ `r sd_dirichlet(c(9, 3, 3, 3))[1]` and
219 $\mathrm{SD}(\zeta_i) = $ `r sd_dirichlet(c(9, 3, 3, 3))[2]` else.

220 Alternatively, and perhaps favorably, we can directly plot the marginals of the Dirichlet
221 distribution. These marginal priors are known to be beta distributions with shape
222 parameters $s_1 = \alpha_i$ and $s_2 = \alpha_0 - \alpha_i$. For $\boldsymbol{\alpha} = (9, 3, 3, 3)$, the marginal distributions of ζ are
223 exemplified in Figure 2. All of the above approaches to better understand the Dirichlet prior
224 have in common that they ignore the dependencies between elements of $\boldsymbol{\zeta}$. More precisely,
225 elements of $\boldsymbol{\zeta}$ are always negatively correlated as an increase in one element needs to be
226 reflected in a decrease in the other elements to satisfy the sum-to-one constraint. A possible
227 solution would be to plot the multivariate density of the Dirichlet prior, but this will become
228 increasingly hard for higher dimensional $\boldsymbol{\zeta}$ (i.e., for variables with more than three
229 categories) and so we do not illustrate this approach in the present paper.

3   Implementation in brms

231 The brms package (Bürkner, 2017b, 2017a) provides an interface to fit Bayesian
232 generalized (non-)linear multilevel models using Stan (Carpenter et al., 2017; Stan
233 Development Team, 2017), which is a C++ package for performing full Bayesian inference
234 (see also http://mc-stan.org/). It supports a wide range of distributions, allowing users to fit
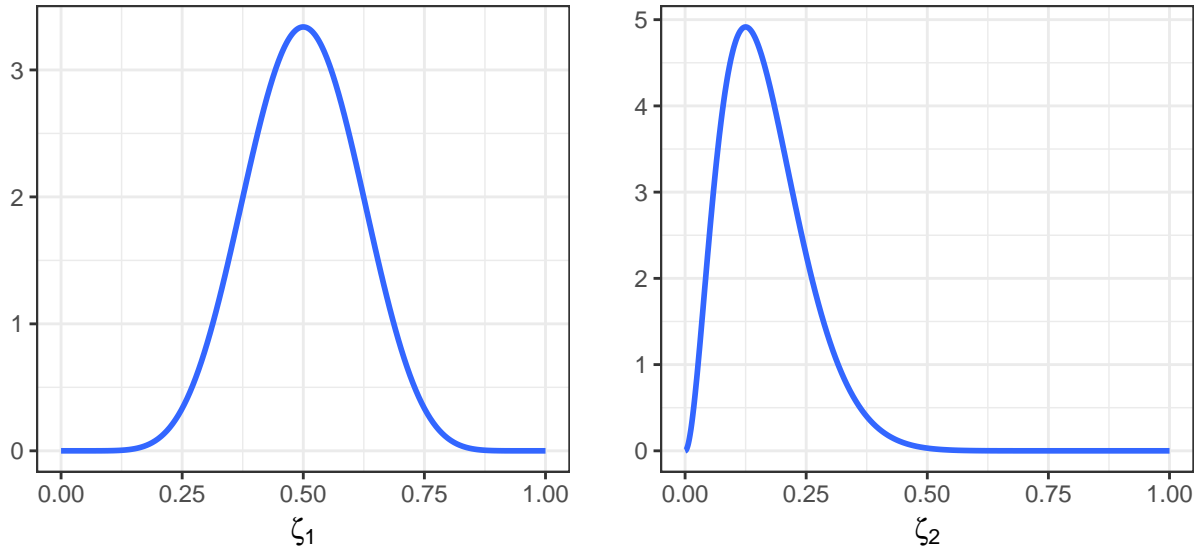
Figure 2. Densities of marginal priors of ζ_1 and ζ_2 for $\boldsymbol{\alpha} = (9, 3, 3, 3)$. The marginal priors of ζ_3 and ζ_4 are in this case identical to the one of ζ_2.

235   – among others – linear, count data, survival, response times, ordinal, zero-inflated, and even

236   self-defined mixture models all in a multilevel context.

237   In brms, monotonic effects are fully integrated into the formula syntax, which builds on

238   and extends standard R formula syntax as well as the multilevel formula syntax initially

239   created for the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Monotonic

240   predictors can be used like any other predictor variable and, with respect to the formula

241   syntax, behave like a numeric predictor. Suppose the response variable `y` is predicted by a

242   monotonic variable `x` and a non-monotonic variable `z` (i.e., a continuous or categorical

243   variable). Then the corresponding model formula is

```
y ~ mo(x) + z
```

244   Modeling both main effects and interaction of `x` and `z` can be achieved by

```
y ~ mo(x) * z
```

245   Depending on whether `z` is a continuous or categorical variable, this will imply a

246   different predictor term, which is fully determined by and thus consistent with the basic R

247 formula syntax. If `z` is monotonic as well, then `z` is simply replaced by `mo(z)`. Please note

248 that for models including interactions with monotonic variables, brms will use *different*

249 simplex parameters for different terms of the same monotonic variable (e.g., for the main

250 effect of `x` and the interaction of `x` and `z`). This results is much greater modeling flexibility

251 as explained in the former section.

252   An especially well developed feature of brms is its multilevel formula syntax allowing

253 to model, for instance, hierachically nested data structures such as multiple observations per

254 person in a longitudinal study. Suppose we wanted to fit a monotonic effect *per* person in a

255 multilevel model, then we could specify this as follows:

```
y ~ mo(x) + (mo(x) | person)
```

256   The `mo(x)` term outside the brackets denotes the *average* monotonic effect across

257 persons, while the `(mo(x) | person)` term indicates that the *difference* between the

258 individual monotonic effects per person and the average effect should be modeled as well (for

259 more details on the brms formula syntax see Bürkner (2017a)). For this parameterization to

260 make sense in combination with monotonic effects, we treat the shape (i.e., the simplex

261 parameter ζ) as constant across persons and only vary the size and direction of the effect

262 (i.e, b) as varying across persons. This restricts the flexibility of the model but results in

263 much more stable estimates and less convergence problems in particular if the number of

264 observations per person (or more generally, per level of the grouping factor) is small.

265   4   Case study: Measures of chronic widespread pain

266   To illustrate the application of monotonic effects in practice, we will reanalyse data

267 used to validate measures of chronic widespread pain (CWP) from patients' point of view

268 (Cieza et al., 2004; Gertheiss et al., 2011a). There is not universally accepted definition of

269 CWP, but "it may be characterized by pain involving several regions of the body, which

270 causes problems in functioning, psychological distress, poor quality of sleep or difficulties in

daily life" (Gertheiss et al., 2011a, p. 378). The applied CWP measures are coming from the

international classification of functioning (ICF; Organization, 2001) and are rated by clinical

staff not by patients themselves. Thus, it is important to validate which and to what degree

CWP measures actually relate to subjective physical health in order to better understand

their implications for patients' life.

For each of the 420 patients, the present data contains information on 67 CWP

measures as well as a subjective measure of physical health based on the SF-36 questionnaire

(Ware & Sherbourne, 1992). The data is freely available in the R package "ordPens"

(Gertheiss, 2015) and is explained in detail in Gertheiss et al. (2011a) and Cieza et al.

(2004). In the data set, the variable of subjective physcial health is called `phcs` while the

CWP measures are named according to their offical ICF coding (see Gertheiss et al., 2011a

for explanation).

In our first model, we will predict the subjective physical health (variable `phcs`) only

by the impairments in "walking" (variable `d450`) and "moving around" (variable `d455`),

which were both measured on a five point scale between 0 ("no problem") and 4 ("complete

problem"). Both of these variables were strong predictors of `phcs` in the analysis of

Gertheiss et al. (2011a). The category labels of these variables suggest that their

relationship with `phcs` will be monotonic. More specifically, we expect the subjective

phyiscal health to decrease with an increase in impairments in "walking" or "moving around"

or basically any other everyday functioning.

For the present example – and for most other data sets we have seen so far – the

default priors of brms on monotonic effects work well in terms of sampling efficiency and

convergence. However, for illustrationary purposes, we still manually specify our own priors

for each model if even priors are similar to the default ones. Based on knowledge about the

outcome scale, it is unlikely that any WCP measure across its full range will influence the

subjective physical health by more than 20 points. We code this expectation as a

Normal(0, 10) prior on the size parameters b. That way, $|b|$ will only exceed 10 and 20

298 outcome points with probabilities of roughly 32% and 5%, respectively. With regard to the

299 shape of the effects, we have no particular prior expectations and thus assume a uniform

300 Dirichlet prior as explained in Section 2.1, which is also the default in brms. When

301 specifying the Dirichlet prior for "walking", we have to take into account that the highest

302 category 4 ("complete problem") is actually not present in the data set (we will see how to

303 solve the problem of missing extreme categories later on). Thus, the corresponding prior

304 requires a vector of reduced size. In brms, we can specify the above priors by means of the

305 following code:

```r
library(brms)
prior1 <- prior(normal(0, 10), class = "b") +
  prior(dirichlet(1, 1, 1), class = "simo", coef = "mod4501") +
  prior(dirichlet(1, 1, 1, 1), class = "simo", coef = "mod4551")
```

306 We use class `simo` to refer to the simplex parameters of monotonic effects. The required

307 coefficient names `"mod4501"` and `"mod4551"` are constructed as `mo<variable><index>`,

308 where `<index> = 1` unless a single regression term contains multiple simplexes – which only

309 happens for interactions of monotonic effects. Finally, we fit the model in brms via

```r
fit1 <- brm(phcs ~ mo(d450) + mo(d455), data = cwp, prior = prior1)
```

310 As illustrated in Figure 3, both predictors show a strong negative relationship to

311 subjective physical health. Moreover, these relationships are clearly (at least visually)

312 non-linear. For impairments in walking, for instance, changes in the outcome are strongest

313 between the first two categories implying that the most subjective physical health is lost as

314 soon as any problems in walking occur. This impression is confirmed by the summary

315 estimates of the simplex parameters (see Table 1).

316 We can also show this non-linearity using model comparison. First, we fit a linear

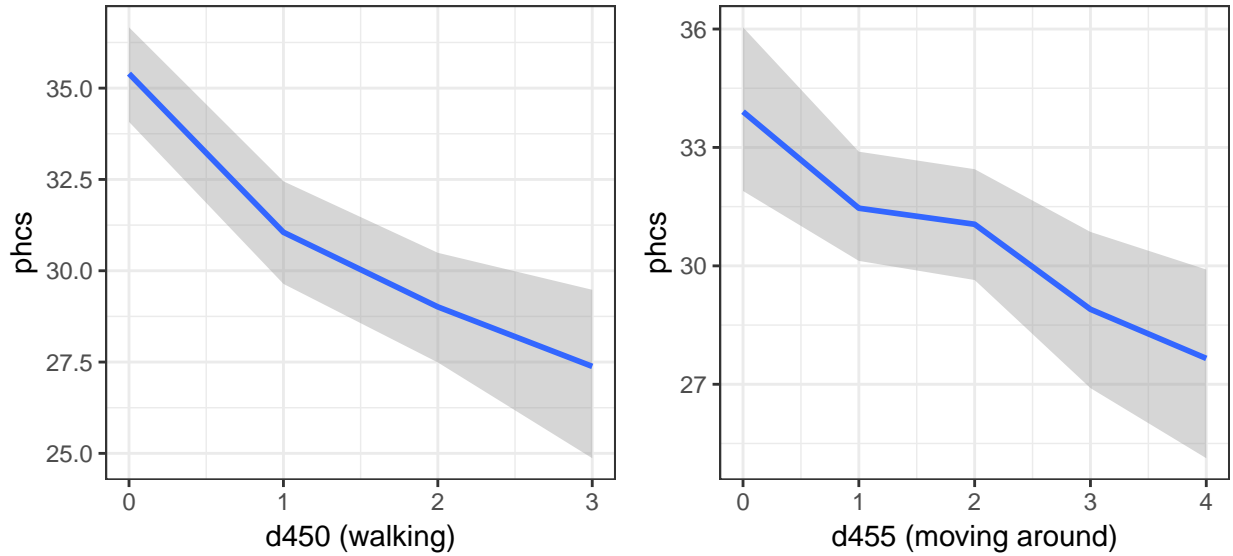317 model with the same predictors via

Figure 3. Effects of impairments in walking and moving around on subjective physical health as estimated by model `fit1`.

```r
prior2 <- prior(normal(0, 2.5), class = "b")
fit2 <- brm(phcs ~ d450 + d455, data = cwp, prior = prior2)
```

318   and then compute model weights, for instance, by means of the WAIC (Watanabe,

319   2010):

```r
model_weights(fit1, fit2, weights = "waic")
```

320   This yields a weight of 92% for the monotonic model, again providing evidence that a

321   linear model may be too restrictive to adequately describe the relationship between

322   impairments in walking or moving around and subjective physical health.

323   In the original study of Gertheiss et al. (2011a) on this data set, the purpose was to

324   select a subset of the total of 67 CWP measures that are related to subjective physcial health

325   in a relevant manner. For the purpose of the present case study, we also aim at a form of

326   variable selection but with a somewhat different focus. As most CWP measures show small

327   to medium correlations with at least a few other CWP measures, we expect only few of them

328   to have a considerable non-zero effect on subjective physcial health after controlling for all

Table 1

Summary of estimated simplexes for impairments in walking and moving around.

	walking			moving around		
	Estimate	l-95% CI	u-95% CI	Estimate	l-95% CI	u-95% CI
simo[1]	0.54	0.33	0.78	0.39	0.14	0.66
simo[2]	0.26	0.05	0.49	0.07	0.00	0.23
simo[3]	0.20	0.01	0.43	0.35	0.06	0.64
simo[4]				0.19	0.01	0.48

other measures. For simplicity, we are not including "environmental factor" variables into our analysis as they were measured on a different scale (from -4 "complete barrier to 4 'complete facilitator") as all the other variables (from 0 "no problem" to 4 "complete problem").

This leaves us with a total of 51 predictors, which is still quite a lot to estimate for a data set containing 420 observations, in particular because of rather high intercorrelations of predictors. For this reason, we impose some regularization on the the size parameters b by applying the *regularized horseshoe* prior (Carvalho, Polson, & Scott, 2009; Piironen & Vehtari, 2016; Piironen, Vehtari, & others, 2017). This prior has very fat tails and an infinite spike at zero which results in close-to-zero coefficients to be shrunken to zero, while greater coefficients located in the tails of the prior remain largely unchanged. Thus, the horseshoe prior can be used to guide variable selection (Piironen et al., 2017). There are a lot of options to tune the horseshoe prior, but for the purpose of the present case study, we will only use the `par_ratio` argument. With this argument, we can formalize our prior expectation about the number of non-zero effects, which we will set to 10%, that is roughly 5 of the total 51 predictors. The smaller `par_ratio`, the stronger the shrinkage towards zero. In brms, we can specify this prior as follows:

```
prior3 <- prior(horseshoe(par_ratio = 0.1), class = "b")
```

345 Before we actually fit the model, we add an artifical row to our data set which contains

346 the maximal value (4) for each of the predictor variables and a missing value (NA) for the

347 subjective physical health measure phcs. This ensures that all size parameters are on the

348 same scale even if the maximal category was not actually present in the data set, as we had

349 seen above for impairments in walking. The model including 51 CWP measures as

350 monotonic predictors is then set up via

```
fit3 <- brm(phcs | mi() ~ ..., data = cwp, prior = prior3)
```

351 The mi() term on the left-hand side of the formula ensures that the newly added row

352 with a missing value in phcs is actually included in the model, as otherwise it would just

353 have been removed during the data preparation step. The right-hand side abbreviated above

354 as .... actually contains separate mo() terms for the 51 included CWP measures, which we

355 did not write out due to the length of that expression.

356 As illustrated in Figure 4, only few predictors have an effect that deviates from zero in

357 a relevant manner after applying regularization of the horseshoe prior. Most notably, these

358 are impairments in "walking" (d450) and "moving around (d455), but also in 'community

359 life" (d910) and "sensation of pain" (b280), which all seem to have a negative effect on

360 subjective physical heatlh after controlling for all other predictors. This does not necessarily

361 imply that other predictors have no additional predictive value. To better better understand

362 the latter, different variable selection techniques – for instance those explained in Gertheiss

363 et al. (2011a) – may be favorable.

364 Similar to what we did before, we can compare fit of the monotonic model to a

365 corresponding linear model on which we apply the horseshoe prior as well. Performing model

366 comparison by means of the WAIC yields a weight of 94% for the monotonic model, which

367 thus seems to perform better than its linear counterpart even though most of the predictors

368 actually have an effect very close to zero (see Figure 4). Intuitively, one may expect that

369 monotonic effects tend to overfit the data in such a case as they have much more parameters

370 than linear effects. However, this is not what actually happens. If the size parameter b is

371 close to zero, there is not much to learn about the corresponding simplex parameter ζ, which

372 will thus have a posterior distribution close to its prior. Still, this uncertainty will not lead to

373 overfitting as changes in ζ do not influence predictions as long as b is small. In other words, a

374 monotonic predictor with a close to zero effect naturally reduces to a simple linear predictor.

5   Conclusion

376   In the present paper, we introduced a principled approach to including ordinal

377 predictors in regression models, which we called *monotonic effects.* In simple cases, they

378 coincide with estimates provided by isotonic regression while allowing to penalize larger

379 changes between adjacent categories via prior distributions. Thus, monotonic effects

380 naturally combine important ideas of existing methods for modeling ordinal predictors.

381 Moreover, monotonic effects nicely integrate into the framework of generalized linear

382 regression and can even be used together with multilevel structures. They are fully supported

383 in the brms R package, which fits Bayesian regression models using Stan and provides an

384 intuitive user interface based on widely known R formula syntax. To date, ordinal predictors

385 are still mostly treated as either nominal or metric thus under- or overestimating the

386 contained information. Monotonic effects avoid these problems but still allow for an intuitive

387 interpretation of the estimated parameters. In summary, we hope that monotonic effects can

388 solve some longstanding problems in the treatment of ordinal predictors.

389   Monotonic effects have been implemented in brms for about two years at the time of

390 writing this paper, which allowed us (and users of brms) to get a reasonable amount of

391 experience with their behavior. From what we have seen in our own data sets and what

392 users reported, sampling efficiency and convergence were good and rarely much worse that

393 when using a purely categorical or linear approach. This is notable insofar, as elements of a

simplex tend to be negatively correlated, sometimes rather strongly, thus making MCMC sampling more difficult. The fact that this has not been a major issue – at least from our experience – may be largely due the advanced Hamiltonian Monte-Carlo samplers implemented in Stan, which are designed to work well even for highly inter-correlated posteriors (Betancourt, Byrne, Livingstone, & Girolami, 2014; Hoffman & Gelman, 2014).

For simple cases such as regression models with only a single monotonic effect and normally distributed errors, maximum likelihood estimators can be developed as well (Barlow et al., 1972; Robertson et al., 1988). However, we believe them to be of limited practical applicability since the supported models were necessarily of far less complexity as compared to what can be fitted right away in a fully Bayesian framework. Moreover, computing uncertainty estimates for simplex parameters in a frequentist framework is naturally difficult as for finite data, the distribution of their estimators are not necessarily sufficiently normal. This is true in particular for elements of the simplex that come close to the naturally lower or upper boundaries (0 or 1) of the simplex. Thus, any confidence interval constructed based on approximate standard errors will likely be inappropriate in many cases. For these reasons, we chose not to investigate the frequentist properties of monotonic effects any further in the present paper.

Although our primary focus was the use of monotonic effects for modelling strictly ordinal predictors, we want to point out that monotonic effects may be applied to other kinds of discrete variables, as well. Such variables may represent, for instance, count data or discrete points in time. As an example for the former, we can think of participants solving a sequence of figural analogy tasks with the value of interest being the number of tasks solved correctly. This count variable could then be used as predictor of a general intelligence score. It is plausible to assume the number of correctly solved items to be monotonically related to general intelligence and so the applications of a monotonic effect appears reaonsable. As an example for the latter, we could think of a longidutional study with few measurement points. If the outcome was a skill gradually aquired over time, we would expect time to be

monotonically related to it. Of course, time may also be modeled as continuous, but for very

few time points, using a monotonic effect may be a more reliable solution without strong

assumptions outside of monotonicity.

⁴²⁴   **References**

⁴²⁵   Agresti, A. (2010). *Analysis of ordinal categorical data.* Chichester: John Wiley &

⁴²⁶ Sons. doi:10.1002/9780470594001

⁴²⁷   Barlow, R. E., Bremner, J. M., Brunk, H. D., & Bartholomew, D. J. (1972). *Statistical*

⁴²⁸ *inference under order restrictions: The theory and application of isotonic regression.* John

⁴²⁹ Wiley & Sons.

⁴³⁰   Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects

⁴³¹ models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

⁴³²   Best, M. J., & Chakravarti, N. (1990). Active set algorithms for isotonic regression; a

⁴³³ unifying framework. *Mathematical Programming*, *47*(1-3), 425–439.

⁴³⁴   Betancourt, M., Byrne, S., Livingstone, S., & Girolami, M. (2014). The geometric

⁴³⁵ foundations of hamiltonian monte carlo. *arXiv Preprint arXiv:1410.5110.*

⁴³⁶   Bürkner, P.-C. (2017a). Advanced Bayesian multilevel modeling with the R package

⁴³⁷ brms. *arXiv Preprint*, 1–15. Retrieved from https://arxiv.org/abs/1705.11123

⁴³⁸   Bürkner, P.-C. (2017b). brms: An R package for Bayesian multilevel models using

⁴³⁹ Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

⁴⁴⁰   Bürkner, P.-C., & Vuorre, M. (2018). Ordinal regression models in psychological

⁴⁴¹ research: A tutorial.

⁴⁴²   Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . .

⁴⁴³ Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical*

⁴⁴⁴ *Software.*

⁴⁴⁵   Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the

⁴⁴⁶ horseshoe. In *Artificial intelligence and statistics* (pp. 73–80).

⁴⁴⁷   Christensen, R. H. B. (2018). ordinal—regression models for ordinal data.

⁴⁴⁸   Cieza, A., Stucki, G., Weigl, M., Kullmann, L., Stoll, T., Kamen, L., . . . Walsh, N.

⁴⁴⁹ (2004). ICF core sets for chronic widespread pain. *Journal of Rehabilitation Medicine*, *36*(0),

63–68.

Dykstra, R. L., & Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, 708–716.

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). Introduction to the Dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton.* Retrieved from https://www2.ee.washington.edu/techsite/papers/documents/UWEETR-2010-0006.pdf

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition.* Boca Raton: Chapman and Hall/CRC.

Gertheiss, J. (2014). ANOVA for factors with ordered levels. *Journal of Agricultural, Biological, and Environmental Statistics*, *19*(2), 258–277.

Gertheiss, J. (2015). *ordPens: Selection and/or smoothing of ordinal predictors.* Retrieved from https://CRAN.R-project.org/package=ordPens

Gertheiss, J., & Tutz, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review*, *77*(3), 345–365.

Gertheiss, J., Hogger, S., Oberhauser, C., & Tutz, G. (2011a). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*(3), 377–395.

Gertheiss, J., Oehrlein, F., & others. (2011b). Testing linearity and relevance of ordinal predictors. *Electronic Journal of Statistics*, *5*, 1935–1959.

Hoffman, M., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian monte carlo. *The Journal of Machine Learning Research*, *15*(1), 1593–1623.

Kelly, C., & Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 1071–1085.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial Introduction with R*

[477] (2nd Edition.). Burlington, MA: Academic Press.

[478] Lee, C.-I. C. (1981). The quadratic loss of isotonic regression under normality. *The*
[479] *Annals of Statistics*, *9*(3), 686–688.

[480] Lee, C.-I. C. (1996). On estimation for monotone dose—response curves. *Journal of*
[481] *the American Statistical Association*, *91*(435), 1110–1119.

[482] Leitenstorfer, F., & Tutz, G. (2007). Generalized monotonic regression based on
[483] b-splines with an application to air pollution data. *Biostatistics*, *8*(3), 654–673.

[484] Liddell, T., & Kruschke, J. K. (2017). Analyzing ordinal data with metric models:
[485] What could possibly go wrong? *Open Science Framework*. doi:10.17605/OSF.IO/9H3ET

[486] Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview
[487] and a survey of recent developments. *Test*, *14*(1), 1–73.

[488] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal*
[489] *Statistical Society. Series B (Methodological)*, 109–142.

[490] McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R*
[491] *and Stan*. CRC Press.

[492] Organization, W. H. (2001). *International classification of functioning disability and*
[493] *health: ICF*. Geneva: World Health Organization.

[494] Piironen, J., & Vehtari, A. (2016). On the hyperprior choice for the global shrinkage
[495] parameter in the horseshoe prior. *arXiv Preprint arXiv:1610.05559*.

[496] Piironen, J., Vehtari, A., & others. (2017). Sparsity information and regularization in
[497] the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051.

[498] Pya, N., & Wood, S. N. (2015). Shape constrained additive models. *Statistics and*
[499] *Computing*, *25*(3), 543–559.

[500] R Core Team. (2018). *R: A Language and Environment for Statistical Computing*.
[501] Vienna, Austria: R Foundation for Statistical Computing. Retrieved from
[502] https://www.R-project.org/

[503] Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical*

504   *inference.* John Wiley & Sons.

505   Stan Development Team. (2017). *Stan modeling language: User's guide and reference*

506   *manual.* Retrieved from http://mc-stan.org/manual.html

507   Tutz, G. (2011). *Regression for categorical data* (Vol. 34). Cambridge University Press.

508   Ware, J. E., & Sherbourne, C. D. (1992). The mos 36-item short-form health survey

509   (sf-36): I. conceptual framework and item selection. *Medical Care*, 473–483.

510   Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely

511   applicable information criterion in singular learning theory. *Journal of Machine Learning*

512   *Research*, *11*(Dec), 3571–3594.

513   Wu, W. B., Woodroofe, M., & Mentz, G. (2001). Isotonic regression: Another look at

514   the changepoint problem. *Biometrika*, *88*(3), 793–804.

515   Yee, T. W., Stoklosa, J., Huggins, R. M., & others. (2015). The VGAM package for

516   capture–recapture data using the conditional likelihood. *J. Statist. Soft*, *65*(5), 1–33.

517   **Appendix**

518   **Appendix A: Mathematical Proofs**

519   *Proof. (Monotonicity)* For all values x between 0 and $C - 1$, we have

$$\eta(x+1) - \eta(x) = b\sum_{i=1}^{x+1}\zeta_i - b\sum_{i=1}^{x}\zeta_i = b\zeta_{x+1}. \tag{7}$$

520   Since $\zeta_{x+1} > 0$, the linear predictor $\eta(x)$ is monotonically increasing if $b \geq 0$ and

521   monotonically decreasing if $b \leq 0$.   □

522   *Proof. (Equivalence to categorical isotonic regression)* Consider a simple linear model of a

523   continuous response y regressed on a categorical predictor x with categories $j \in \{0, ..., C\}$.

524   Further, let μ_j be the group mean of category j with respect to the response variable. Then

525   the model for observation n can be written as

$$y_n = \mu_{x_n} + e_n, \tag{8}$$

526   where e_n are errors of the regression. In categorical isotonic regression, we estimate

527   $\mu = (\mu_0, ..., \mu_C)$ under the order-constraint $\mu_0 \leq \mu_1 \leq ... \leq \mu_C$ or $\mu_0 \geq \mu_1 \geq ... \geq \mu_C$. Using

528   a monotonic effect, we write:

$$y_n = b_0 + b_1 \sum_{i=1}^{x_n}\zeta_i + e_n. \tag{9}$$

529   Hence, we can identify μ_0 with b_0 and μ_j with $b_0 + b_1\sum_{i=1}^{j}\zeta_i$ for $j > 0$. This

530   identification is bijective within the set of order-constraint μ.   □

531   *Proof. (Proposition 2.1)* Under the stated assumptions, we can without, loss of generality,

532   write the linear predictor $\eta = \eta(x)$ as

$$\eta(x) = b_0 + \sum_{i=1}^{K} b_i \, \mathrm{mo}(x, \zeta) = b_0 + \left(\sum_{i=1}^{K} b_i\right) \mathrm{mo}(x, \zeta). \tag{10}$$

533   Since all other predictors have been fixed to some constants, their contribution to η can be

534   absorbed by the intercept b_0 and the regression coefficients b_1 to b_K which are all related to

535 x. If we define $b_x = \sum_{i=1}^{K} b_i$ we see that $\eta(x)$ is monotonic in x with the sign of the effet

536 determined by the sign of b_x.   ☐

537 *Proof. (Counter example to conditional monotonicity for varying simplex parameters)*

538 Consider the situation shown in Figure 5, where quite clearly, the effect of $\boldsymbol{x}$ is monotonic for

539 group a, but non-monotonic for group b. Suppose further that we named the grouping

540 variable $\boldsymbol{z}$ and aplied dummy coding such that $a = 0$ and $b = 1$. Using different simplex

541 parameters for the main effect of $\boldsymbol{x}$ and the interaction effects between $\boldsymbol{x}$ and $\boldsymbol{z}$, the linear

542 predictor reads as follows:

$$\eta(x, z) = b_1 \, z + b_2 \, \text{mo}(x, \boldsymbol{\zeta}_{xb_2}) + b_3 \, z \, \text{mo}(x, \boldsymbol{\zeta}_{xb_3}) \tag{11}$$

543   For group a this results in $\eta(x, 0) = b_2 \, \text{mo}(x, \boldsymbol{\zeta}_{xb_2})$ so that $b_2 = 100$ as well as

544 $\boldsymbol{\zeta}_{xb_2} = (0.8, 0.2)$ are completely defined by the curve of group a. For group b, we have

$$\eta(x, 1) = b_1 + b_2 \, \text{mo}(x, \boldsymbol{\zeta}_{xb_2}) + b_3 \, \text{mo}(x, \boldsymbol{\zeta}_{xb_3}). \tag{12}$$

545   As the curve of group b starts at the origin, we have $b_1 = 0$. Due to the chosen

546 parameterization of $\boldsymbol{z}$, the term $b_3 \, \text{mo}(x, \boldsymbol{\zeta}_{xb_3})$ models the *difference* between in the effect of

547 $\boldsymbol{x}$ between the two groups, which visualized as a dashed line in Figure 5 and is clearly

548 monotonic. Consequently, we have $b_3 = 60$ and $\boldsymbol{\zeta}_{xb_3} = (\frac{1}{6}, \frac{5}{6})$. Although the assumptions of

549 the monotonic effects are fully met, the effect of $\boldsymbol{x}$ in group b is non-monotonic. Thus, $\boldsymbol{x}$ is

550 not conditionally monotonic given $\boldsymbol{z}$.

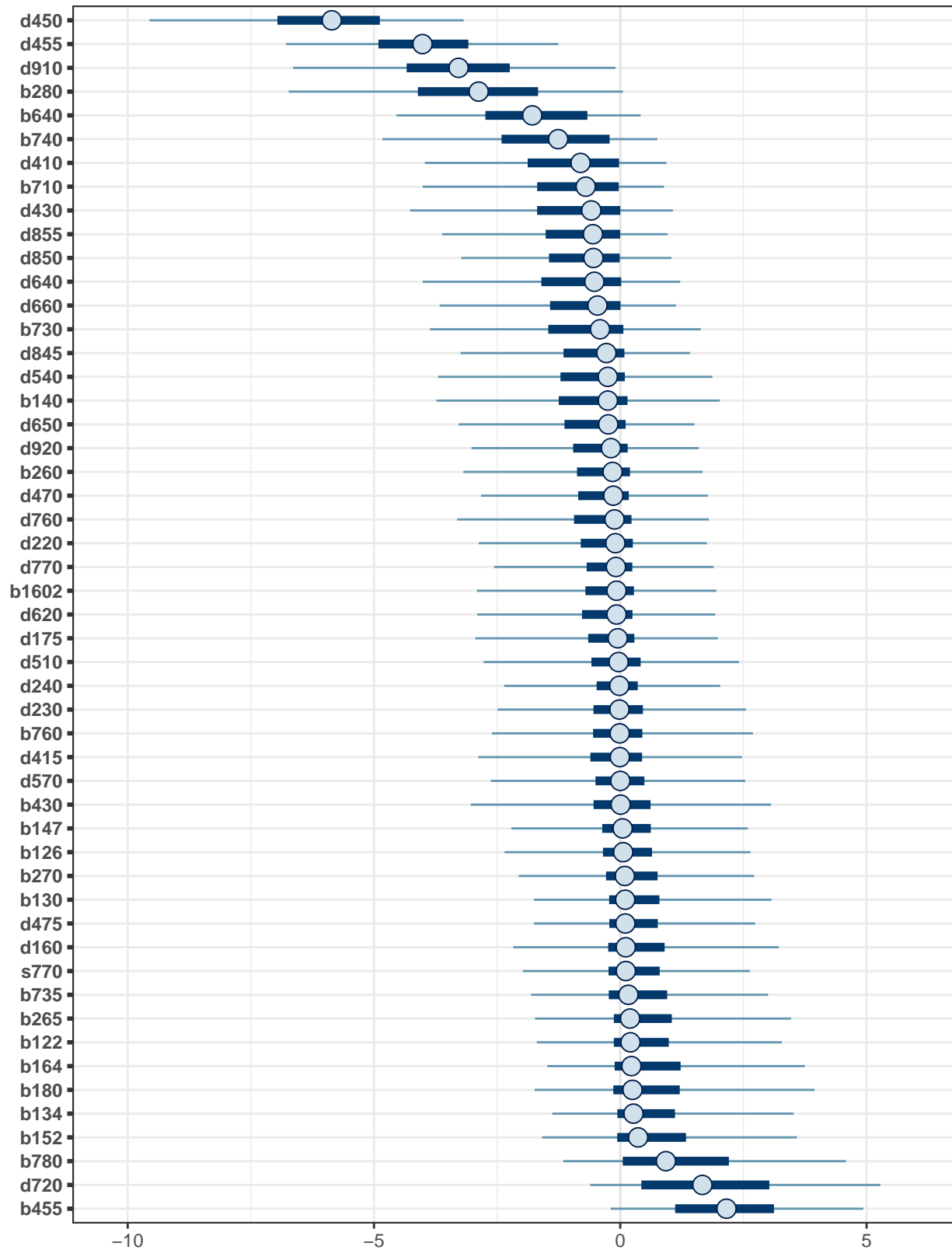551   ☐

Figure 4. Size parameters of 51 CWP measures as estimated by model `fit3`. See Gertheiss et al. (2011a) for details on the variable names.
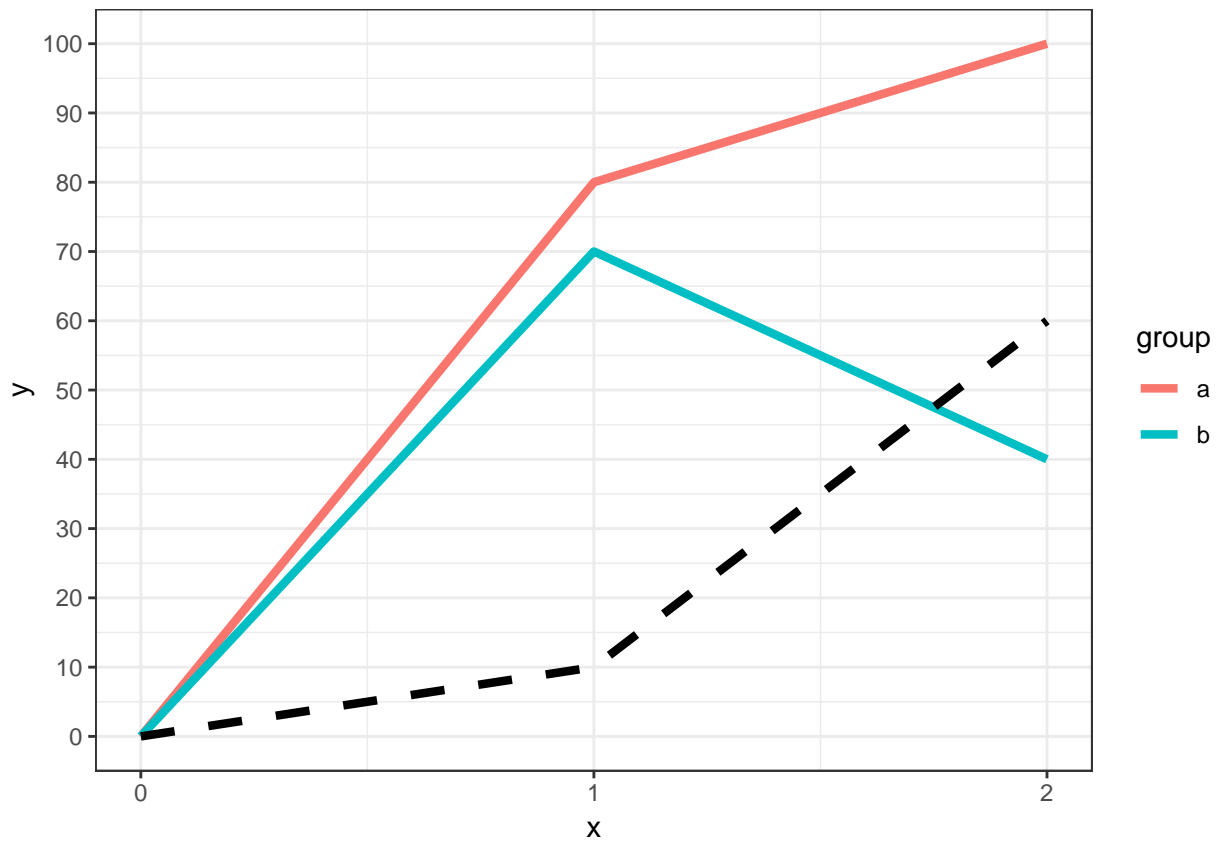
Figure 5. Counter example to the conditional monotonicity for varying simplex parameters. The dashed line shows the difference between the groups a and b as a function of $\boldsymbol{x}$.