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

Abstract

Ordinal predictors are commonly used in regression models. Yet, they are often incorrectly treated as either nominal or metric thus under- or overestimating the contained information. This is understandable insofar as generally applicable solutions or corresponding statistical software are still underdeveloped. We propose a new way of parameterizing regression coefficients of ordinal predictors, which we call monotonic effects. The reparameterization is done in terms of a scale parameter b taking care of the direction and size of the effect and a simplex parameter ζ modeling the normalized differences between categories. This ensures that predictions are monotonically increasing or decreasing, while changes between adjacent categories may vary across categories. This formulation nicely generalizes to both interaction terms as well as multilevel structures. Monotonic effects may not only be applied to ordinal predictors, but also to other discrete variables for which a monotonic relationship is plausible. This includes variables representing count data or discrete points in time. Fitting monotonic effects in a fully Bayesian framework is straightforward with the R package brms, which also allows to incorporate prior information and to check the assumption of monotonicity.

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

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 and Agresti, 2005; Tutz, 2011

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

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, 2017, 2018) 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 et al., 2011b; Gertheiss & Tutz, 2009) and the other are categorical 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 assumption might be somewhat too flexible. More specifically, we do often expect the changes between adjacent categories to be *monotonic*, that is consistently negative or positive across the full range of the ordinal variable. At the same time, the size of the changes may still vary across categories by a substantial amount as ordinality does not necessarily contain information about distance between categories.

The research on *isotonic regression* deals with regression models subject to order constraints (Barlow et al., 1972; Robertson et al., 1988). For instance, in some contexts, the effect of a drug may be assumed to be monotonically[1] increasing with an increasing dose – an assumption that we often want to "hard-code" into our models. Depending on the research question and nature of the variable on which we want to impose a monotonicity constraint, different techniques may be favorable. If the variable is essentially continuous, such as the dose of a drug, we can use parametric functions which are known to be monotonic (e.g., the log or logistic functions in simple cases) or use semi-parametric approaches such as monotonic splines (Kelly & Rice, 1990; Lee, 1996; Leitenstorfer & Tutz, 2007; Pya & Wood, 2015). If the variable under study is categorical, the monotonicity assumption reduces to an ordering constraint on the group means with respective to the response variable. From the perspective of classical *frequentist* statistics, the latter case has been studied extensively in Barlow et al. (1972) and Robertson et al. (1988; see also Best and Chakravarti, 1990; Dykstra and Robertson, 1982; Lee, 1981; Wu, Woodroofe, and Mentz, 2001). For the purpose of studying ordinal predictors, we are primarily interested in the categorical type of isotonic regression.

---

[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.

⁵² The idea and scope of the two approaches discussed above is somewhat different.
⁵³ While isotonic regression only restricts the *direction* of the changes, the penalized regression
⁵⁴ method restricts the *size* of the changes leaving the direction untouched (although we may
⁵⁵ also introduce order constraints in the latter according to Gertheiss et al., 2011b). None of
⁵⁶ these two approaches is more reasonable *per se* and we should not necessarily think of them
⁵⁷ as competing in how ordinal predictors should ideally be modeled. Rather, they make use of
⁵⁸ different sets of assumptions which both reflect important aspects of ordinal variables. As
⁵⁹ we will see later on, we may even combine them naturally within the same framework.

⁶⁰ In the present paper, we will introduce a monotonicity imposing parameterization
⁶¹ of ordinal predictors, which we call *monotonic effects*. They represent a way to generalize
⁶² the assumption of linearity to ordinal predictors. As explained in detail in the next section,
⁶³ the estimated parameters have an intuitive meaning and are thus easy to interpret and
⁶⁴ communicate. In simple cases, they also turn out to be equivalent to the results of categorical
⁶⁵ isotonic regression.

⁶⁶ The structure of this paper is a follows. In Section 2, we will introduce monotonic
⁶⁷ effects as well as their mathematical foundation in detail. We continue by explaining a
⁶⁸ software implementation of monotonic effects in the R package "brms" (Bürkner, 2017,
⁶⁹ 2018), which supports a wide and growing range of Bayesian regression models. In Section 4,
⁷⁰ a case study dealing with measures of chronic widespread pain (Cieza et al., 2004; Gertheiss
⁷¹ et al., 2011a) will be discussed, in which we make extensive use of monotonic effects. We
⁷² end with a discussion in Section 5. Some short mathematical proofs about the properties of
⁷³ monotonic effects are presented in the Appendix.

Monotonic Effects

⁷⁵ A predictor which we want to model as monotonic must have discrete values in an
⁷⁶ ordered set and be coded as integers. The integers may represent, for instance, count data,
⁷⁷ discrete points in time, or categories of an ordinal variable. Since the latter is possibly
⁷⁸ the most relevant use case in psychology and related disciplines, we speak of predictor
⁷⁹ categories in the following to refer to the values of a monotonic predictor. As opposed to a
⁸⁰ continuous predictor, predictor categories are not assumed to be equidistant with respect
⁸¹ to their effect on the response variable. Instead, the distance between adjacent predictor
⁸² categories is estimated from the data and may vary across categories. This is realized by
⁸³ parameterizing as follows: One parameter, b, takes care of the direction and size of the
⁸⁴ effect similar to an ordinary regression parameter, while an additional parameter vector, ζ,
⁸⁵ estimates the normalized distances between consecutive predictor categories. For a single
⁸⁶ monotonic predictor, x, with a lowest possible value of zero[2], the corresponding predictor
⁸⁷ term η_n of observation n looks as follows:

---

[2]Note that this convention differs of the one customarily used in statistical software, where indices of vectors, matrices, etc. usually start at one. However, starting at zero eases the notation of monotonic effects and so we adopt this approach in the present paper.
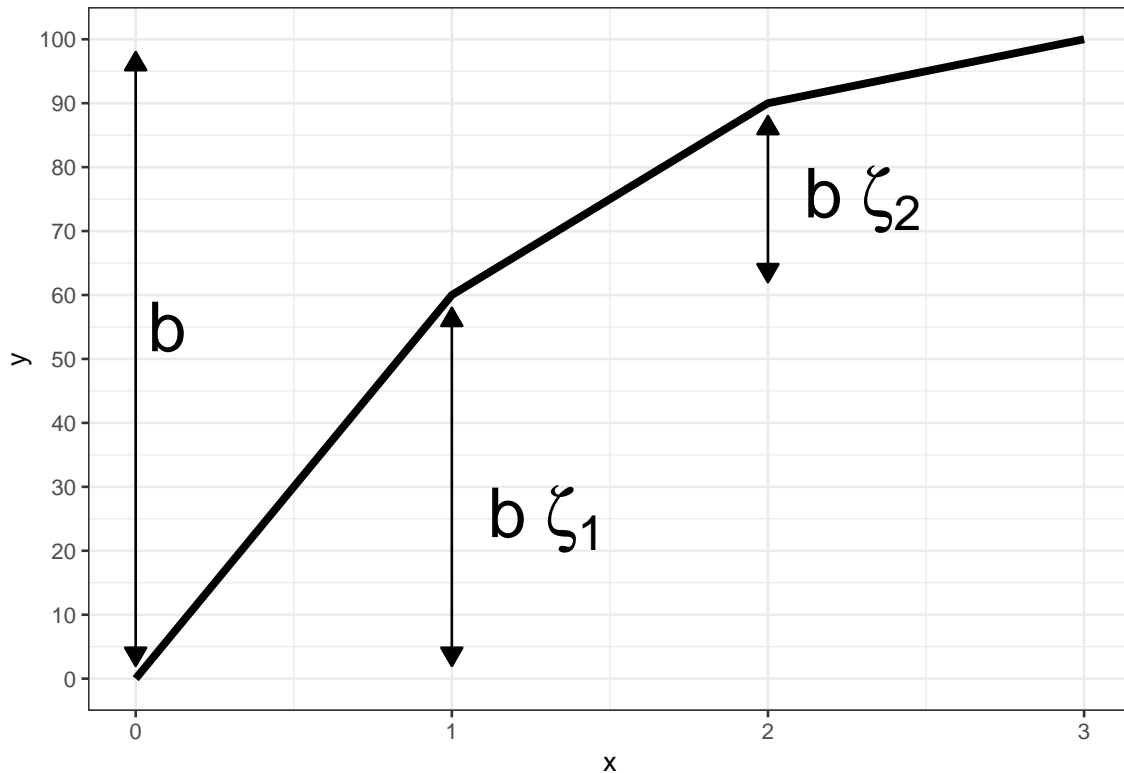
Figure 1. Visualization of a monotonic effect with four categories. Parameters were set to $b = 100$ and $\boldsymbol{\zeta} = (0.6, 0.3, 0.1)$.

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

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 $\sum_{i=1}^{D} \zeta_i = 1$ with D being the highest possible integer value that $\boldsymbol{x}$ can take on. Since we start counting at zero, D is equal to the number of differences between two adjacent categories and also equal to the total number of categories minus one. If the monotonic effect is used in a linear model, 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 has an intuitive interpretation while guaranteeing the monotonicity of the effect (see Appendix A for a formal proof). For notational convenience we define $\text{mo}(x, \boldsymbol{\zeta}) = \sum_{i=1}^{x} \zeta_i$ and call $\text{mo}()$ 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

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

where $\boldsymbol{\zeta}_{xb_2}$ and $\boldsymbol{\zeta}_{xb_3}$ are two independent simplex parameters. Under this formulation, $\boldsymbol{x}$ may not necessarily be conditionally monotonic for all values of $\boldsymbol{z}$ (see Appendix A for a counter example). Rather the monotonicity being modeled depends on the chosen parameterization.

For instance, if the predictor $\boldsymbol{z}$ is dummy coded as 0 and 1 representing the two categories of a dichotomous variable, the formulation above models the effect of $\boldsymbol{x}$ to be monotonic for category 0 as well as for the *change* between category 1 and 0. Conversely, when using cell mean coding rather than dummy coding for $\boldsymbol{z}$, the model assumes a different monotonic effect of $\boldsymbol{x}$ for both categories of $\boldsymbol{z}$. In the latter case, $\boldsymbol{x}$ is conditionally monotonic on $\boldsymbol{z}$. If we fix all simplex parameters corresponding to the same monotonic variable $\boldsymbol{x}$ to the same value, conditionally monotonicity is achieved in general (proof provided in Appendix A):

Proposition 1. *Let $\boldsymbol{\eta}$ be an arbitrary linear predictor term containing the monotonic predictor $\boldsymbol{x}$ with the corresponding simplex parameter $\boldsymbol{\zeta}$ being the same across all terms including $\boldsymbol{x}$. Then $\boldsymbol{\eta}$ is monotonic in $\boldsymbol{x}$ conditionally on all possible combinations of all other predictor variables.*

While fixing all simplex parameters associated with $\boldsymbol{x}$ to the same vector guarantees conditional monotonicity, it may be not flexible enough for many common situations. For instance, if one wanted to model different monotonic effects for two groups, it would imply the shape ($\boldsymbol{\zeta}$) of the predictions to be the same across groups with just their total range (b) to be different. As explained in Section 3, in brms we make use of both parameterizations (varying and constant $\boldsymbol{\zeta}$) at different places in the package.

Monotonic effects in a Bayesian framework

The present paper describes monotonic effects as embedded in a fully Bayesian framework. We consider every statistical models a *Bayesian* model if it quantifies the uncertainty in all observed and unobserved variables (conventionally denoted as data and parameters respectively) by means of probabilities. This is often expressed in terms of Bayes' Theorem, which states that the posterior distribution $p(\theta|y)$ of the model parameters θ given the data y can be expressed in terms of the product of likelihood $p(y|\theta)$ and prior distribution $p(\theta)$ as well as a normalizing constant $p(y)$:

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

138 A thorough introduction to Bayesian statistics is outside the scope of the present
139 paper. Instead, we refer to well established text books such as McElreath (2016), Kruschke
140 (2014), and Gelman et al. (2013).

141 With respect to monotonic effects, a fully Bayesian framework has two main impli-
142 cations. First, such a framework allows to incorporate monotonic effects in a large class
143 of regression models without the need to develop any problem-specific estimators. Second,
144 it implies that we can think of prior distributions for b and $\boldsymbol{\zeta}$. Such prior distributions
145 enable us to incorporate information, which does not come directly from data in terms of
146 the likelihood contribution, such as expert knowledge or findings from previous studies.

147 Priors for b can be derived based on the *a priori* expectation regarding the differences
148 between highest and lowest category, which we call *maximal difference* in the following.
149 Any family of prior distributions typically applied to regression coefficients can be applied
150 on b, as well. As a weakly-informative prior for b, we can understand any location shift
151 distribution – such as a normal of student-t distribution – centered around zero and with
152 a scale parameter large enough to allow for large but plausible maximal differences, while
153 penalizing implausibly large maximal differences. This scale will necessarily depend on
154 the scale of the response distribution and also on the range of the monotonic predictor.
155 Alternatively, one may use an improper flat prior that treats all real values as being equally
156 likely *a priori* in the hope that the data alone is sufficient to identify b. We will come back
157 to this in the discussion of our case study.

158 Setting a prior on the simplex parameter $\boldsymbol{\zeta}$ requires a different approach. The canonical
159 prior of a simplex parameter is the Dirichlet distribution, a multivariate generalization of
160 the beta distribution (Frigyik, Kapila, & Gupta, 2010). It is non-zero for all valid simplexes
161 (i.e., for $\boldsymbol{\zeta}$ with $\zeta_i \in [0,1]$ and $\sum_{i=1}^{D} \zeta_i = 1$) and zero otherwise. The Dirichlet prior has a
162 single parameter vector $\boldsymbol{\alpha}$ of the same length as $\boldsymbol{\zeta}$. Its density is defined as

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

163 where $B(\boldsymbol{\alpha})$ is a normalizing constant (Balakrishnan, 2014). As the *a-priori* expec-
164 tation of ζ_i is given by $w_i = \mathbb{E}(\zeta_i) = \alpha_i/\alpha_0$, with $\alpha_0 = \sum_{i=1}^{D} \alpha_i$, higher values of α_i in
165 comparison to the sum over $\boldsymbol{\alpha}$ imply higher *a priori* values of ζ_i. Moreover, the higher the
166 sum over $\boldsymbol{\alpha}$, the higher the certainty in each of the proportions w_i.

167 In the absense of any problem specific information, a reasonable default prior on $\boldsymbol{\zeta}$
168 would surely be one that assumed all differences between adjacent categories to be the same
169 on average while being considerably uncertain about this expectation. Such a prior would
170 imply, on average, a linear trend but with enough uncertainty to allow for all other possible
171 monotonic trends as well. The Dirichlet prior with a constant $\boldsymbol{\alpha} = 1$ puts equal probability
172 on all valid simplex and can thus be understood as the multivariate generalization of the

173 uniform prior on simplexes. Since we have $w_i = 1/D$, this prior centers $\boldsymbol{\zeta}$ around a linear
174 trend with large uncertainty and thus appears to be a good default prior in the absense of
175 any problem specific information.

Penalizing larger changes between categories

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

184   When it comes to the shape of the monotonic effect, we have to take a closer look at
185 the prior on $\boldsymbol{\zeta}$. As discussed above, a constant vector $\boldsymbol{\alpha}$ of the Dirichlet prior on $\boldsymbol{\zeta}$ implies a
186 linear trend in expectation. In other words, for constant $\boldsymbol{\alpha}$, the prior means of all changes
187 ζ_i between adjacent categories are the same. The higher the sum over $\boldsymbol{\alpha}$, the higher the
188 certainty in that expectation. Thus, if we expect a linear trend with some certainty, we
189 assign all elements of $\boldsymbol{\alpha}$ to the same value a. To get an intuition about what is a reasonable
190 value for a, we may use the standard deviation of the elements ζ_i, which can be computed
191 as (see Balakrishnan, 2014):

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

192   Although the standard deviation is an imperfect measure of variability for the Dirichlet
193 distribution as the latter is not symmetric in general, we still believe the former to be helpful
194 in better understanding the implications of ones chosen priors. For the default of $a = 1$ and
195 a total of $D = 5$ categories, we get a rather large standard deviation of $\mathrm{SD}(\zeta_i) = 0.19$. If we
196 set, for example, $a = 5$, we get $\mathrm{SD}(\zeta_i) = 0.09$ and thus much higher certainty in changes of
197 equal size.

198   Of course, the process of increasing $\boldsymbol{\alpha}$ on average works equally well even if we do not
199 expect all changes to be the same *a priori*. For instance, if $D = 5$ and we expect a 3-times
200 larger change between the first two categories than between all the other categories with
201 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.
202 As standard deviations, we get $\mathrm{SD}(\zeta_1) = 0.11$ and $\mathrm{SD}(\zeta_i) = 0.09$ else.

203   Alternatively, and perhaps favorably, we can directly plot the marginals of the Dirichlet
204 distribution. These marginal priors are known to be beta distributions with shape parameters
205 $s_1 = \alpha_i$ and $s_2 = \alpha_0 - \alpha_i$ (Balakrishnan, 2014). For $\boldsymbol{\alpha} = (9, 3, 3, 3)$, the marginal distributions
206 of ζ are exemplified in Figure 2. All of the above approaches to better understand the
207 Dirichlet prior have in common that they ignore the dependencies between elements of $\boldsymbol{\zeta}$.
208 More precisely, elements of $\boldsymbol{\zeta}$ are always negatively correlated as an increase in one element
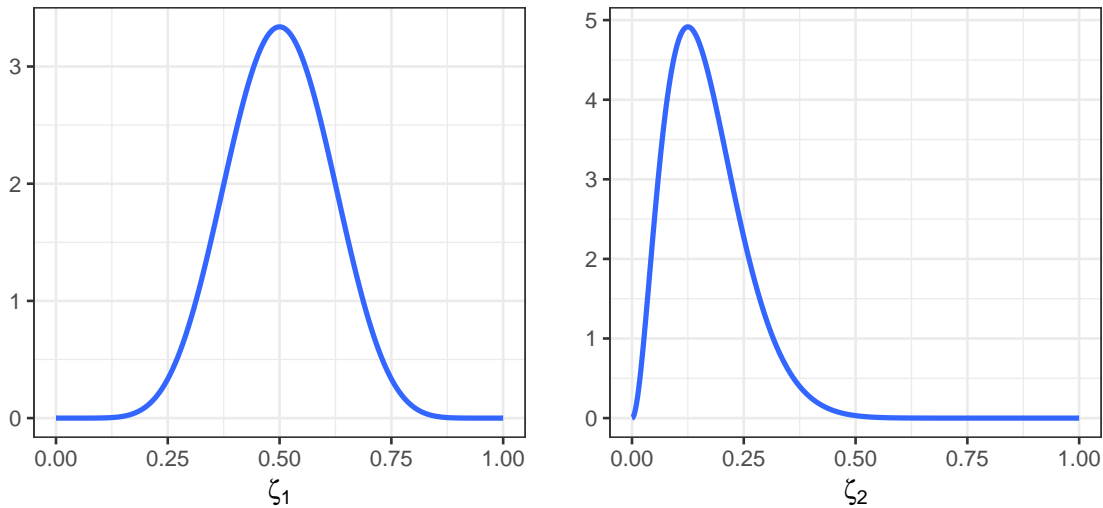209 needs to be reflected in a decrease in the other elements to satisfy the sum-to-one constraint.

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.

A possible solution would be to plot the multivariate density of the Dirichlet prior, but this will become increasingly hard for higher dimensional $\boldsymbol{\zeta}$ (i.e., for variables with more than three categories) and so we do not illustrate this approach in the present paper.

Implementation in brms

The brms package (Bürkner, 2017, 2018) provides an interface to fit Bayesian generalized (non-)linear (multilevel) regression models using Stan (Carpenter et al., 2017; Stan Development Team, 2017), which is a C++ package for performing full Bayesian inference (see also http://mc-stan.org/). It supports a wide range of distributions, allowing users to fit – among others – linear, count data, survival, response times, ordinal, zero-inflated, and even self-defined mixture models all in a multilevel context.

In brms, monotonic effects are fully integrated into the formula syntax, which builds on and extends standard R formula syntax as well as the multilevel formula syntax initially created for the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Monotonic predictors can be used like any other predictor variable and, with respect to the formula syntax, behave like a numeric predictor. Suppose the response variable `y` is predicted by a monotonic variable `x` and a non-monotonic variable `z` (i.e., a continuous or categorical variable). Then the corresponding model formula is

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

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

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

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

²²⁹ different predictor term, which is fully determined by and thus consistent with the basic
²³⁰ R formula syntax. If `z` is monotonic as well, then `z` is simply replaced by `mo(z)`. Please
²³¹ note that for models including interactions with monotonic variables, brms will use *different*
²³² simplex parameters for different terms of the same monotonic variable (e.g., for the main
²³³ effect of `x` and the interaction of `x` and `z`). This results is much greater modeling flexibility
²³⁴ as explained in the former section.

²³⁵ Variable which should be modeled as monotonic may either be integer valued or an
²³⁶ ordered factor. In the latter case, the ordered factor will be transformed to an integer
²³⁷ variable with the lowest factor level being identified with zero as described above.

²³⁸ An especially well developed feature of brms is its multilevel formula syntax allowing
²³⁹ to model, for instance, hierarchically nested data structures such as multiple observations
²⁴⁰ per person in a longitudinal study. Suppose we wanted to fit a monotonic effect *per* person
²⁴¹ in a multilevel model, then we could specify this as follows:

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

²⁴² The `mo(x)` term outside the brackets denotes the *average* monotonic effect across per-
²⁴³ sons, while the `(mo(x) | person)` term indicates that the *difference* between the individual
²⁴⁴ monotonic effects per person and the average effect should be modeled as well (for more
²⁴⁵ details on the brms formula syntax see Bürkner (2018)). For this parameterization to make
²⁴⁶ sense in combination with monotonic effects, we treat the shape (i.e., the simplex parameter
²⁴⁷ ζ) as constant across persons and only vary the size and direction of the effect (i.e, b) as
²⁴⁸ varying across persons. This restricts the flexibility of the model but results in much more
²⁴⁹ stable estimates and less convergence problems in particular if the number of observations
²⁵⁰ per person (or more generally, per level of the grouping factor) is small.

²⁵¹ ### Case study: Measures of chronic widespread pain

²⁵² To illustrate the application of monotonic effects in practice, we will reanalyze data
²⁵³ used to validate measures of chronic widespread pain (CWP) from patients' point of view
²⁵⁴ (Cieza et al., 2004; Gertheiss et al., 2011a). There is not universally accepted definition of
²⁵⁵ CWP, but "it may be characterized by pain involving several regions of the body, which
²⁵⁶ causes problems in functioning, psychological distress, poor quality of sleep or difficulties in
²⁵⁷ daily life" (Gertheiss et al., 2011a, p. 378). The applied CWP measures are coming from
²⁵⁸ the international classification of functioning (ICF; Organization, 2001) and are rated by
²⁵⁹ clinical staff not by patients themselves. Thus, it is important to validate which and to
²⁶⁰ what degree CWP measures actually relate to subjective physical health in order to better
²⁶¹ understand their implications for patients' life.

²⁶² For each of the 420 patients, the present data contains information on 67 CWP
²⁶³ measures as well as a subjective measure of physical health based on the SF-36 questionnaire
²⁶⁴ (Ware & Sherbourne, 1992). The data is freely available in the R package "ordPens"
²⁶⁵ (Gertheiss, 2015) and is explained in detail in Gertheiss et al. (2011a) and Cieza et al.
²⁶⁶ (2004). In the data set, the variable of subjective physical health is called `phcs` while the

CWP measures are named according to their official 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 physical 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 illustrative 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 outcome points with probabilities of roughly 32% and 5%, respectively. With regard to the shape of the effects, we have no particular prior expectations and thus assume a uniform Dirichlet prior as explained in Section 2.1, which is also the default in brms. When specifying the Dirichlet prior for "walking", we have to take into account that the highest category 4 ("complete problem") is actually not present in the data set (we will see how to solve the problem of missing extreme categories later on). Thus, the corresponding prior requires a vector of reduced size. In brms, we can specify the above priors by means of the 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")
```

We use class `simo` to refer to the simplex parameters of monotonic effects. The required coefficient names `"mod4501"` and `"mod4551"` are constructed as `mo<variable><index>`, where `<index> = 1` unless a single regression term contains multiple simplexes – which only 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)
```

As illustrated in Figure 3, both predictors show a strong negative relationship to subjective physical health. Moreover, these relationships are clearly (at least visually) non-linear. For impairments in walking, for instance, changes in the outcome are strongest between the first two categories implying that the most subjective physical health is lost as soon as any problems in walking occur. This impression is confirmed by the summary estimates of the simplex parameters (see Table 1).

We can also show this non-linearity using model comparison. First, we fit a linear model with the same predictors via
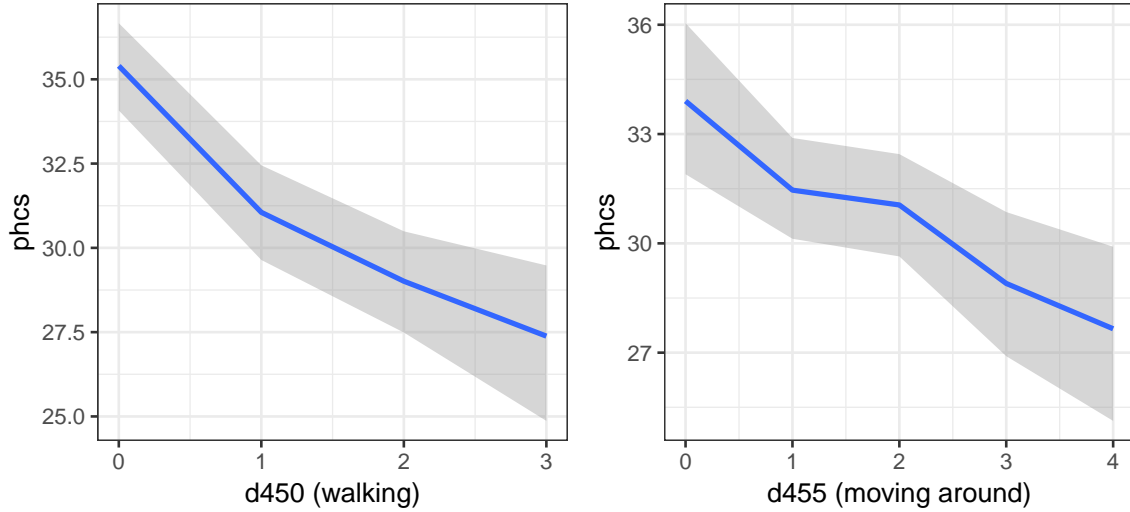
Figure 3. Effects of impairments in walking and moving around on subjective physical health as estimated by model `fit1`.

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

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

303   and then compute model weights, for instance, by means of the stacking of posterior-
304 predictive distribution (Yao, Vehtari, Simpson, & Gelman, 2017)[3]:

```
model_weights(fit1, fit2, weights = "loo2")
```

305   This yields a weight of 94% for the monotonic model, again providing evidence that
306 a linear model may be too restrictive to adequately describe the relationship between
307 impairments in walking or moving around and subjective physical health.

308   In the original study of Gertheiss et al. (2011a) on this data set, the purpose was
309 to select a subset of the total of 67 CWP measures that are related to subjective physical

---

[3]In a Bayesian framework, models may be compared by various means for instance Bayes factors (Kass & Raftery, 1995), (approximate) cross-validation methods (Vehtari, Gelman, & Gabry, 2017), information criteria (Vehtari et al., 2017; Watanabe, 2010) or stacking of posterior-predictive distributions (Yao et al., 2017). A discussion of the pros and cons of these various approaches is outside the scope of the present paper.

health in a relevant manner. For the purpose of the present case study, we also aim at a form of variable selection but with a somewhat different focus. As most CWP measures show small to medium correlations with at least a few other CWP measures, we expect only few of them to have a considerable non-zero effect on subjective physical health after controlling for all 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 inter-correlations 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")
```

Before we actually fit the model, we add an artificial row to our data set which contains the maximal value (4) for each of the predictor variables and a missing value (`NA`) for the subjective physical health measure `phcs`. This ensures that all size parameters are on the same scale even if the maximal category was not actually present in the data set, as we had seen above for impairments in walking. The model including 51 CWP measures as monotonic predictors is then set up via

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

The `mi()` term on the left-hand side of the formula ensures that the newly added row with a missing value in `phcs` is actually included in the model, as otherwise it would just have been removed during the data preparation step. The right-hand side abbreviated above as `...`. actually contains separate `mo()` terms for the 51 included CWP measures, which we did not write out due to the length of that expression.

As illustrated in Figure 4, only few predictors have an effect that deviates from zero in a relevant manner after applying regularization of the horseshoe prior. Most notably, these are impairments in "walking" (`d450`) and "moving around (`d455`), but also in 'community life" (`d910`) and "sensation of pain" (`b280`), which all seem to have a negative effect on subjective physical health after controlling for all other predictors. This does not necessarily imply that other predictors have no additional predictive value. To better better understand the latter, different variable selection techniques – for instance those explained in Gertheiss

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

350   Similar to what we did before, we can compare fit of the monotonic model to a
351 corresponding linear model on which we apply the horseshoe prior as well. Performing model
352 comparison by means of the stacking of posterior-predictive distributions (Yao et al., 2017)
353 yields a weight of 90% for the monotonic model, which thus seems to perform better than
354 its linear counterpart even though most of the predictors actually have an effect very close
355 to zero (see Figure 4). Intuitively, one may expect that monotonic effects tend to overfit
356 the data in such a case as they have much more parameters than linear effects. However,
357 this is not what actually happens. If the size parameter b is close to zero, there is not much
358 to learn about the corresponding simplex parameter ζ, which will thus have a posterior
359 distribution close to its prior. Still, this uncertainty will not lead to overfitting as changes in
360 ζ do not influence predictions as long as b is small. In other words, a monotonic predictor
361 with a close to zero effect naturally reduces to a simple linear predictor.

362   **Discussion**

363   In the present paper, we introduced a principled approach to including ordinal predic-
364 tors in regression models, which we called *monotonic effects*. In simple cases, they coincide
365 with estimates provided by isotonic regression while allowing to penalize larger changes be-
366 tween adjacent categories via prior distributions. Thus, monotonic effects naturally combine
367 important ideas of existing methods for modeling ordinal predictors. Moreover, monotonic
368 effects nicely integrate into the framework of generalized linear regression and can even be
369 used together with multilevel structures. They are fully supported in the brms R package,
370 which fits Bayesian regression models using Stan and provides an intuitive user interface
371 based on widely known R formula syntax. To date, ordinal predictors are still mostly treated
372 as either nominal or metric thus under- or overstating the contained information. Monotonic
373 effects avoid these problems but still allow for an intuitive interpretation of the estimated
374 parameters. In summary, we hope that monotonic effects can solve some longstanding
375 problems in the treatment of ordinal predictors.

376   As illustrated in the case of interactions of monotonic and categorical predictors, the
377 use of monotonic predictors introduce some subtleties in the parameterization of a model,
378 which may highlight structural information. For instance, one may choose to model the effect
379 of the monotonic predictor to be monotonic for each category, or rather choose *differences*
380 between categeories to be monotonic. Both options may be reasonable depending on the
381 research question and a-priori information available.

382   Monotonic effects have been implemented in brms for about two years at the time
383 of writing this paper, which allowed us (and users of brms) to get a reasonable amount of
384 experience with their behavior. From what we have seen in our own data sets and what users
385 reported, sampling efficiency and convergence were good and rarely much worse that when
386 using a purely categorical or linear approach. This is notable insofar, as elements of a simplex
387 tend to be negatively correlated, sometimes rather strongly, thus making MCMC sampling
388 more difficult. The fact that this has not been a major issue – at least from our experience –
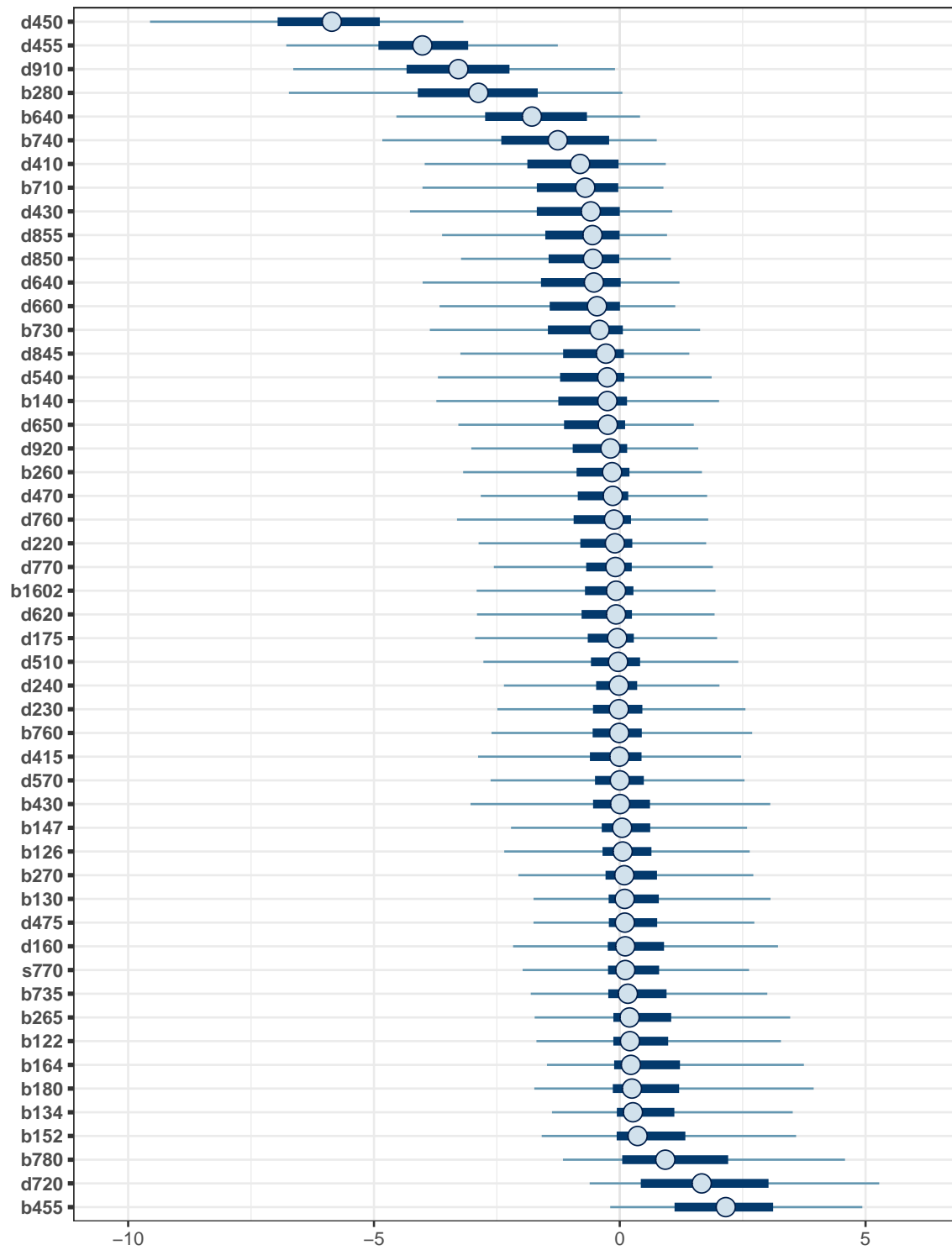389 may be largely due the advanced Hamiltonian Monte-Carlo samplers implemented in Stan,

Figure 4. Size parameters of 51 CWP measures as estimated by model `fit3`. See Gertheiss et al. (2011a) for details on the variable names.

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.

We want to take this opportunity to briefly discusss the default approach to ordinal predictors (coded as ordered factors) in R, which is to compute orthogonal polynomials on their integer representations to model linear, qudratic, cubic, etc. terms of the predictors (Chambers, Hastie, & others, 1992). We believe this approach to be suboptimal for various reasons. First, it assumes the levels of the ordinal variable to be equidistant, which is clearly an oversimplication. Second, it does not ensure monotonicty. Third, resulting parameter estimates may be hard to interprete and fourth, penalizing complexity is much less straightforward than for, say, monotonic effects.

Although our primary focus was the use of monotonic effects for modeling 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 reasonable. As an example for the latter, we could think of a longitudinal study with few measurement points. If the outcome was a skill gradually acquired 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.

Although we focused on the discrte case in the present paper, the idea of monotonic effects also generalizes to continuous data. In this case, the sum in the definition of monotonic effects becomes an integral and ζ a non-negative function to be integrated over. A similar idea is used in I-splines (i.e., intergral splines) whose basis functions represent integrals over the non-negative basis functions of another spline (Ramsay, 1988). In future research, it may thus be worthwhile to study continuous versions of monotonic effects and to relate them to existing methods (e.g., Ramsay, 1988; Pya & Wood, 2015) that ensure monotonicty in the continuous case.

References

Agresti, A. (2010). *Analysis of ordinal categorical data.* Chichester: John Wiley & Sons. doi:10.1002/9780470594001

Balakrishnan, N. (2014). Continuous multivariate distributions. *Wiley StatsRef: Statistics Reference Online.*

Barlow, R. E., Bremner, J. M., Brunk, H. D., & Bartholomew, D. J. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression.* John Wiley & Sons.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Best, M. J., & Chakravarti, N. (1990). Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, *47*(1-3), 425–439.

Betancourt, M., Byrne, S., Livingstone, S., & Girolami, M. (2014). The geometric foundations of hamiltonian monte carlo. *arXiv Preprint arXiv:1410.5110.*

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 1–15.

Bürkner, P.-C., & Vuorre, M. (2018). Ordinal regression models in psychological research: A tutorial.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software.*

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial intelligence and statistics* (pp. 73–80).

Chambers, J. M., Hastie, T. J., & others. (1992). *Statistical models in s* (Vol. 251). Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA.

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

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

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

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.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

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* (2nd Edition.). Burlington, MA: Academic Press.

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

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

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

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

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

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

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

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

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

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

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

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441.

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

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference.* John Wiley & Sons.

Stan Development Team. (2017). *Stan modeling language: User's guide and reference manual.* Retrieved from http://mc-stan.org/manual.html

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

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Ware, J. E., & Sherbourne, C. D. (1992). The mos 36-item short-form health survey (sf-36): I. Conceptual framework and item selection. *Medical Care*, 473–483.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(Dec), 3571–3594.

Wu, W. B., Woodroofe, M., & Mentz, G. (2001). Isotonic regression: Another look at the changepoint problem. *Biometrika*, *88*(3), 793–804.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2017). Using stacking to average bayesian predictive distributions. *arXiv Preprint arXiv:1704.02030.*

Yee, T. W., Stoklosa, J., Huggins, R. M., & others. (2015). The VGAM package for capture–recapture data using the conditional likelihood. *J. Statist. Soft*, *65*(5), 1–33.

539   **Appendix**

540   **Appendix A: Mathematical Proofs**

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

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

542   Since $\zeta_{x+1} > 0$, the linear predictor $\eta(x)$ is monotonically increasing if $b \geq 0$ and monotoni-
543   cally decreasing if $b \leq 0$. $\square$

544   *Proof. (Equivalence to categorical isotonic regression)* Consider a simple linear model of a
545   continuous response y regressed on a categorical predictor x with categories $j \in \{0, ..., D\}$.
546   Further, let μ_j be the group mean of category j with respect to the response variable. Then
547   the model for observation n can be written as

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

548   where e_n are errors of the regression. In categorical isotonic regression, we estimate
549     $\mu = (\mu_0, ..., \mu_C)$ under the order-constraint $\mu_0 \leq \mu_1 \leq ... \leq \mu_C$ or $\mu_0 \geq \mu_1 \geq ... \geq \mu_C$. Using
550   a monotonic effect, we write:

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

551   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
552   identification is bijective within the set of order-constraint μ. $\square$

553   *Proof. (Proposition 2.1)* Under the stated assumptions, we can without, loss of generality,
554   write the linear predictor $\eta = \eta(x)$ as

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

555   Since all other predictors have been fixed to some constants, their contribution to η can be
556   absorbed by the intercept b_0 and the regression coefficients b_1 to b_K which are all related to
557   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 effect
558   determined by the sign of b_x. $\square$

559   *Proof. (Counter example to conditional monotonicity for varying simplex parameters)* Con-
560   sider the situation shown in Figure 5, where quite clearly, the effect of $\boldsymbol{x}$ is monotonic for
561   group a, but non-monotonic for group b. Suppose further that we named the grouping
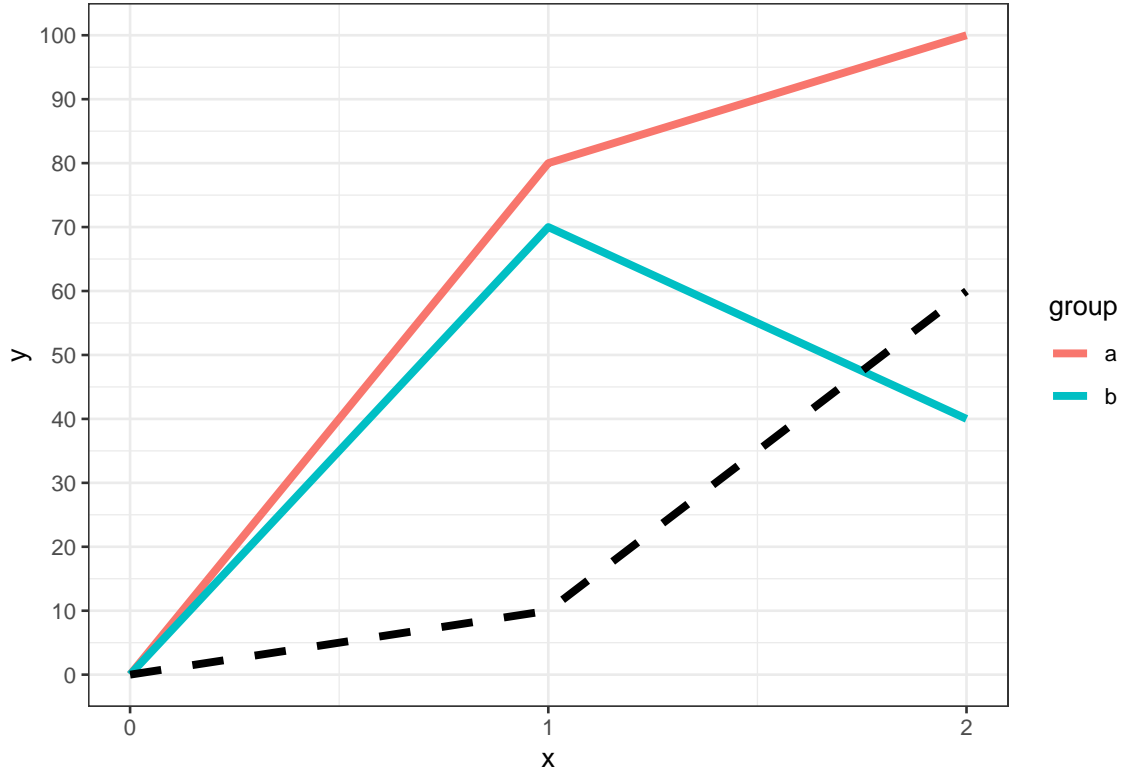562   variable $\boldsymbol{z}$ and applied dummy coding such that $a = 0$ and $b = 1$. Using different simplex

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 $\boldsymbol{x}$.

parameters for the main effect of $\boldsymbol{x}$ and the interaction effects between $\boldsymbol{x}$ and $\boldsymbol{z}$, the linear predictor reads as follows:

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

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

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

As the curve of group b starts at the origin, we have $b_1 = 0$. Due to the chosen parameterization of $\boldsymbol{z}$, the term $b_3\, \text{mo}(x, \boldsymbol{\zeta}_{xb_3})$ models the *difference* between in the effect of $\boldsymbol{x}$ between the two groups, which visualized as a dashed line in Figure 5 and is clearly monotonic. Consequently, we have $b_3 = 60$ and $\boldsymbol{\zeta}_{xb_3} = (\frac{1}{6}, \frac{5}{6})$. Although the assumptions of the monotonic effects are fully met, the effect of $\boldsymbol{x}$ in group b is non-monotonic. Thus, $\boldsymbol{x}$ is not conditionally monotonic given $\boldsymbol{z}$.

$\square$