Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Bürkner, Paul Christian; Charpentier, Emmanuel

Modelling monotonic effects of ordinal predictors in Bayesian regression models

Published in:
BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY

DOI:
10.1111/bmsp.12195

E-pub ahead of print: 01/01/2020

Document Version
Peer reviewed version

Please cite the original version:
Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY*. https://doi.org/10.1111/bmsp.12195

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Modeling Monotonic Effects of Ordinal Predictors in Bayesian Regression Models

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

[1] Department of Computer Science, Aalto University, Finland
[2] Assistance publique - Hôpitaux de Paris, France

Abstract

Ordinal predictors are commonly used in regression models. They are often incorrectly treated as either nominal or metric, thus under- or overestimating the contained information. Such practices may lead to worse inference and predictions compared to methods which are specifically designed for this purpose. We propose a new method for modeling ordinal predictors that applies in situations in which it is reasonable to assume their effects to be monotonic. The parameterization of such monotonic effects is realized in terms of a scale parameter b representing the direction and size of the effect and a simplex parameter $\boldsymbol{\zeta}$ modeling the normalized differences between categories. This ensures that predictions increase or decrease monotonically, while changes between adjacent categories may vary across categories. This formulation generalizes to 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. In simulation studies, we show that the model is well calibrated and, in case of monotonicity, has similar or even better predictive performance than other approaches designed to handle ordinal predictors. Using Stan, we developed a Bayesian estimation method for monotonic effects, which allows to incorporate prior information and to check the assumption of monotonicity. We have implemented this method in the R package brms, so that fitting monotonic effects in a fully Bayesian framework is now straightforward.

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

Correspondence concerning this article should be addressed to Paul-Christian Bürkner, Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland. E-mail: paul.buerkner@gmail.com

Over the last few decades, a substantial amount of 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 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 continuous out of 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; Bürkner, 2018) to name the perhaps most general ones. Ordinal *predictors* seem to have received less attention in statistical research. In R, for instance, the standard treatment of ordinal predictors is still to compute orthogonal polynomials on their integer representations to model linear, qudratic, cubic, etc. terms of the predictors (Chambers & Hastie, 1992). We believe this approach to be suboptimal for various reasons, most notably because it assumes the ordinal categories to be equidistant, which is clearly an oversimplication, and because it yields parameter estimates we would consider hard to interpret.

The literature on ordinal predictors may be divided into three partially inter-conntected lines of research. The first is based on penalized regression/spline approaches specifically designed for ordinal predictors (Alvarez, Bailey, & Katz, 2011; Gertheiss, 2014; Gertheiss & Oehrlein, 2011; Gertheiss & Tutz, 2009; Gu, 2013), the second are categorical types of isotonic regression (Barlow, Bremner, Brunk, & Bartholomew, 1972; Robertson, Wright, & Dykstra, 1988), and the third are ordinal latent variable models (Jöreskog, 1994; Winship & Mare, 1984). We begin by explaining the penalized regression approach. The main idea of the method proposed by Gertheiss and Tutz (2009) is to penalize large differences between adjacent categories. This is done by imposing a penality on the squared differences between the means of adjacent catgories, that is, on $(\eta(x) - \eta(x-1))^2$, where x denotes values of the ordinal predictor and $\eta(x)$ denotes the predicted mean at category x. The penalty reflects the expectation that, if a predictor is ordinal, changes may happen smoothly and larger differences should thus be unlikely. This approach allows for a principled and flexible handling of ordinal predictors in a way closely related to regression splines (Gertheiss & Tutz, 2009; Gu, 2013). It also has a Bayesian interpretation in terms of priors on the category means (Gertheiss & Tutz, 2009). In the original version of this approach (used in Gertheiss & Tutz, 2009; Gertheiss, 2014; Gertheiss & Oehrlein, 2011), the direction of the changes remains unspecified and may vary across the range of the ordinal variable.

In many practical settings, 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 (e.g., Barlow et al., 1972). For instance, the subjective well-being may be monotonically related to measures of physical or psychological health, which we would typically assess via Likert scales and hence in an ordinal manner. If we have theoretical reasons to expect a monotonic relationship, we may want to incorporate this assumption into our model to improve accuracy of the parameter estimates and predictions, but of course also to test whether this assumption was justified in the first place. Even when monotonicity

is justified, the size of the changes may still vary across ordinal categories by a substantial amount as ordinality does not contain information about the distance between categories.

The major line of statistical research which concerns itself with regression models subject to order constraints (i.e., monotonicity) is known as *isotonic regression*[1] (Barlow et al., 1972; Robertson et al., 1988). Depending on the research question and nature of the variable on which we want to impose a monotonicity constraint, different techniques may be more favorable. If the variable is essentially continuous, such as time intervals or 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 (Gu, 2013; He & Shi, 1998; Helwig, 2017; Kelly & Rice, 1990; Lee, 1996; Leitenstorfer & Tutz, 2006, 2007; Pya & Wood, 2015; Ramsay, 1988; Wang & Small, 2015). If the variable under study is categorical, the monotonicity assumption reduces to an ordering constraint on the predicted category means. Using frequentist approaches, 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; Rufibach, 2010; Wu, Woodroofe, and Mentz, 2001). Bayesian approaches to order constraint category means and testing of these contraints have been develeoped as well (e.g., Klugkist & Mulder, 2008; Danaher, Roy, Chen, Mumford, & Schisterman, 2012; Mulder & Raftery, 2019). For the purpose of studying ordinal predictors, we are primarily interested in the categorical type of isotonic regression although continuous types may provide useful predictions also for categorical predictors if they have sufficient number of categories (e.g., see Helwig, 2017). Building on the penalized regression of Gertheiss and Tutz (2009) and Gu (2013), Helwig (2017) proposed to impose order constraints on the category means so that the implied relationship between response and ordinal predictor is monotonic. Combining the two approaches can lead to improved predictions compared to penalized or isotonic regression alone, provided that the true relationship is monotonic (Helwig, 2017).

The above described approaches to modeling ordinal predictors, especially those which induce some regularization, have good theoretical and practical properties when it comes to predictive accuracy (e.g., Gertheiss & Tutz, 2009; Gu, 2013; Helwig, 2017). Further, the parameter estimates are easy to interpret as they simply consist of the (regularized) response means per ordinal predictor category. As such, they are conceptually closer to how categorical predictors, rather than continuous predictors, are handled in regression models. In contrast, in the present paper, we introduce a new monotonicity imposing parameterization for ordinal predictor terms which behaves much like a continuous predictor term. However, we do not make the assumption of equidistance of the predictor values, which is clearly unwarrented for ordinal variables. The proposed parameterization is designed to fit naturally into generalized linear modeling frameworks and their extentions. As such, it can be seamlessly combined with other types of predictor terms to model parameters of arbitrary response distributions, and may even be used within interactions or multilevel structures. To make this approach easy to rememeber, we simply call it *monotonic effects*, by which, of course, we do not want to imply that this is the only possible way to impose monotonicity. As explained in detail in the next section, the estimated parameters have an

[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.

intuitive meaning and are thus easy to interpret and communicate. In contrast to existing approaches, we work in a fully Bayesian framework for model specification and estimation, which increases the complexity of models in which monotonic effects can be incorporated and also allows to specify prior distributions on the corresponding parameters. The latter may not only be used to incorporate additional subject matter knowledge into the model that would otherwise remain unused, but also to regularize the model's predictions and make it robust against overfitting even in the absense of such specific knowledge.

The method proposed in the present paper, as well as other approaches discussed above, model ordinal predictors as *manifest* variables, that is, do not explicitly consider potential measurement error in these predictors. In contrast, in latent variable models, it is common to model ordinal variables as indicators of an underlying *latent* continuous variable, from which the observed ordinal variable originated via categorization (e.g., Winship & Mare, 1984; Finney & DiStefano, 2006; Jöreskog, 1994; Lei, 2009). Such models then estimate the relationship between this latent variable and the (manifest or latent) response variable. This way, latent ordinal models are able to take measurement error into account and provide estimates of how the relationship between variables would have been if we had been able to directly observe the underlying true continuous construct. Importantly, this also implies a monotonic relationship between the manifest ordinal predictor and the response variable. A latent approach may be a reasonable modeling choice (a) if there is substantial measurement error and/or we are interested in the (hypothetical) latent relationships between variables. In contrast, a manifest approach maybe a reasonable choice (a) if the variables are measured very precisely or simply known by design (e.g., discrete points in time in a longituditinal study), (b) if the main focus lies on making predictions for new response values, and/or (c) if the observable manifest relationships between variables are simply those which are of interest. It is beyond the scope of this paper to make a general point about manifest vs. latent approaches for ordinal variables. However, we want to point out that our proposed method treats ordinal predictors as manifest variables, as is the case for a lot of other prominent approaches (e.g., Klugkist & Mulder, 2008; Gertheiss & Tutz, 2009; Gu, 2013).

The structure of this paper is as follows. In Section 2, we will introduce monotonic effects as well as their mathematical foundation in detail. We continue by explaining a software implementation of monotonic effects in the R package "brms" (Bürkner, 2017; Bürkner, 2018) in Section 3, which supports a wide and growing range of Bayesian regression models. In Section 4, we perform a simulation study to investigate parameter recovery of monotonic effects and compare their performance to other approaches proposed in the literature. In Section 5, a case study dealing with measures of chronic widespread pain (Cieza et al., 2004; Gertheiss et al., 2011) will be discussed, in which we make extensive use of monotonic effects. We end with a discussion in Section 6. Mathematical proofs about the properties of monotonic effects as well as further simulation results are presented in the Appendix.

1 Monotonic Effects

We will develop monotonic effects in the context of a distributional regression framework (Bürkner, 2018) in which the response y is distributed according to distribution D with P distributional parameters $\psi_1, ..., \psi_P$. We write

$$y_n \sim D(\psi_{1n}, \psi_{2n}, \ldots, \psi_{Pn})$$

to stress the dependency on the nth observation. The domain of each parameter ψ_p depends on the distribution D. For instance, the mean parameter of a normal distribution might take on all real values while the probability parameter of a binomial distribution can only take on values in the interval $[0, 1]$. Each ψ_p ($1 \leq p \leq P$; with individual elements ψ_{pn}) may be predicted by a vector of predictor variables $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)$ where each variable $\boldsymbol{x}_k$ is itself a vector of length N. To regress ψ_p on $\boldsymbol{X}$, we formulate ψ_p in terms of a generalized linear model (GLM). To reduce the notational burden, we will drop the index p in the following as the GLM is formulated in the same way for all distributional parameters. We write $\psi = g(\eta)$, where g is the response function (i.e., inverse link function) and $\eta \in \mathbb{R}^N$ is a linear predictor term. Its nth element, η_n, may be written as

$$\eta_n = \sum_{j=0}^{J} b_j f_j(X_n). \tag{1}$$

In Equation (1), X_n denotes the vector $(x_{1n}, \ldots, x_{Kn})$ of predictor values of the nth observation, f_j are (possibly non-linear) transformations of the predictor variables and b_j are the regression coefficients. Typically, $f_0 = 1$ is a constant function to include an intercept into the model. The notation above is a slightly non-standard formulation of a GLM (in fact we could speak of a generalized additive model in this context; Hastie, 2017). We use this notation in order to naturally generalize the framework to monotonic effects as explained in the following.

A predictor variable which we want to model as monotonic must have discrete values in an ordered set, which are coded as integers. The integer value may represent, for instance, count data, discrete points in time, or categories of an ordinal variable. Since the latter is possibly the most relevant use case in psychology and related disciplines, in the following, we are going to concentrate on this example of an application for a monotonic predictor. We are going to refer to the values of such a variable as *predictor categories*. As opposed to the values of a continuous predictor, predictor categories should not be assumed equidistant with respect to their effect on the response variable. Instead, the distance between adjacent predictor categories is estimated from the data and may vary across categories.

Suppose we have an ordinal predictor $\boldsymbol{x}$ which we want to model as having a monotonic effect. Ordinal variables contain no information about the distance between adjacent categories. Thus, without loss of generality, we can code the categories of $\boldsymbol{x}$ so that the lowest possible category is zero[2] and the largest is D. 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. For any value $x \in \{0, \ldots, D\}$ that $\boldsymbol{x}$ can take on, we define

[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 simplifies the notation of monotonic effects and so we adopt this approach in the present paper.

$$\text{mo: } \{0, \ldots, D\} \to [0, D], \quad x \to \text{mo}(x, \boldsymbol{\zeta}) = D \sum_{i=1}^{x} \zeta_i \tag{2}$$

and call it the *monotonic transform*. For notational convenience, we set $\sum_{i=1}^{0} \zeta_i = 0$. The vector $\boldsymbol{\zeta}$ is defined as a simplex, which means that is it satisfies $\zeta_i \in [0, 1]$ and $\sum_{i=1}^{D} \zeta_i = 1$. By definition, the elements of $\boldsymbol{\zeta}$ represent the normalized distances between consecutive predictor categories. As we can identify any set of $D + 1$ ordinal categories with $\{0, \ldots, D\}$, the monotonic transform is invariant under ordinality preserving transformations of $\boldsymbol{x}$.

The additive increment of x_n (i.e., the nth value of $\boldsymbol{x}$) to η_n can be written as:

$$b\,\text{mo}(x_n, \boldsymbol{\zeta}) = b\,D \sum_{i=1}^{x_n} \zeta_i \tag{3}$$

where b can take on any real value. In the above parameterization, b represents the size and the sign of the effect similar to an ordinary regression coefficient. That is, we do not have to specify the sign of the monotonic effect a-priori but let the model find out itself if effect is positive or negative, just as we do it for coefficients in ordinary regression models. To explicitely bring monotonic effects into our GLM framework from (1) we can set $b_j = b$ and $f_j = \text{mo}(., \boldsymbol{\zeta})$. However, monotonic effects cannot be included in standard GLMs because the transformations f_j are not fully known a-priori but contain the parameter $\boldsymbol{\zeta}$, which needs to be estimated along with all other model parameters. As such, monotonic effects have some similarities with regression splines, a fact we will come back to later on.

If the monotonic effect is used in a linear model, b can be interpreted as the expected average difference between two adjacent categories 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 between lowest and highest categories. Thus, this parameterization has an intuitive interpretation while guaranteeing the monotonicity of the effect (see Proof A1 in Appendix A). 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 Proof A2 in Appendix A). A conceptual advantage of monotonic effects over isotonic regression – or other approaches working directly on the category means – is that the former emits a single regression coefficient, b, which can directly be post-processed further. An example where this is useful are path models, in which regression coefficients are multiplied along the paths of interest, and monotonic effects can naturally be incorporated in such models.

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

$$b\ \text{mo}(x_n, \boldsymbol{\zeta})\, f(X_n) \tag{4}$$

where $f(.)$ is an arbitrary function on the set of predictor variables $\boldsymbol{X}$, and may of course include further monotonic effects. In more complex predictor terms, monotonic effects of $\boldsymbol{x}$ may also appear multiple times. As such, one modeling choice to be made is whether different simplex parameters related to $\boldsymbol{x}$ should be the same or allowed to have different

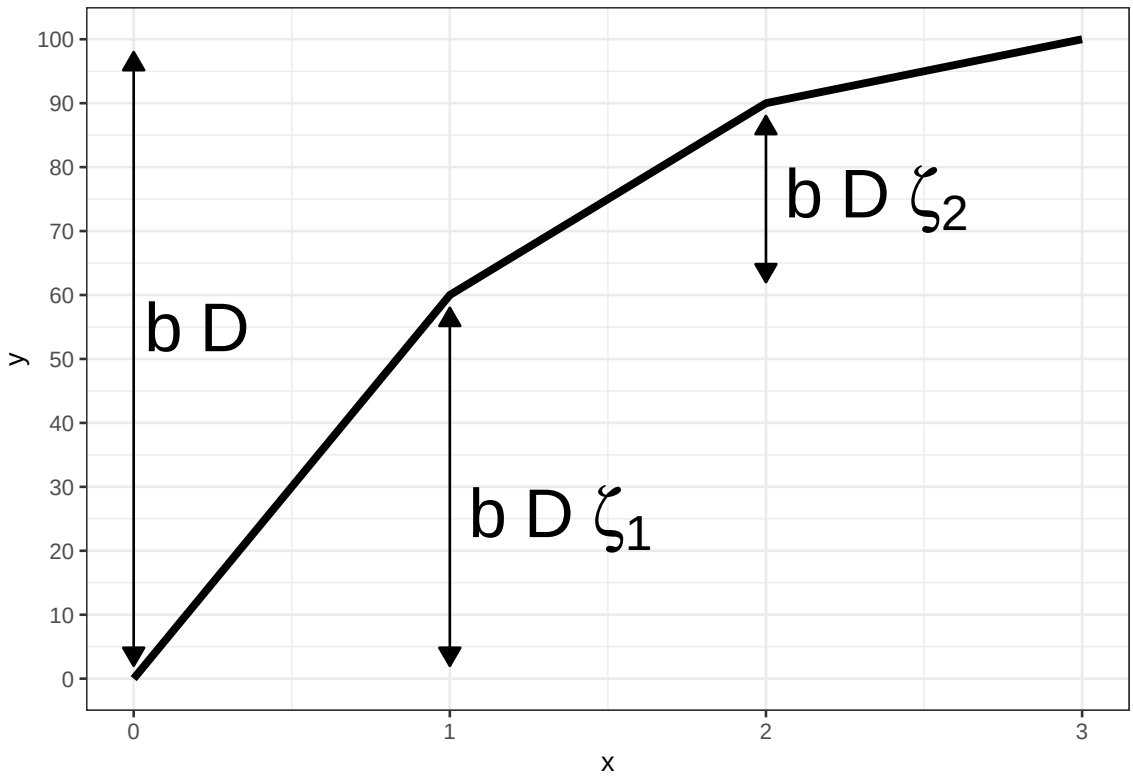

Figure 1. Visualization of a monotonic effect with $D+1 = 4$ predictor categories. Parameters were set to $b = 100/3$ and $\boldsymbol{\zeta} = (0.6, 0.3, 0.1)$.

values. For example, a linear predictor term consisting of an intercept as well as the main effects and two-way interaction between a monotonic predictor $\boldsymbol{x}$ and a continuous (or coded nominal) predictor $\boldsymbol{z}$ could be formulated as

$$\eta_n = b_0 + b_1\, z_n + b_2\, \mathrm{mo}(x_n, \boldsymbol{\zeta}_2) + b_3\, z_n\, \mathrm{mo}(x_n, \boldsymbol{\zeta}_3), \tag{5}$$

where $\boldsymbol{\zeta}_2$ and $\boldsymbol{\zeta}_3$ are two simplex parameters related to $\boldsymbol{x}$. If $\boldsymbol{\zeta}_2$ and $\boldsymbol{\zeta}_3$ are different, $\boldsymbol{x}$ may not necessarily be conditionally monotonic for all values of $\boldsymbol{z}$ (see Proof A4 in 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 (see Proof A3 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 too restrictive 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 overall effect scale (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.

1.1 Monotonic effects in a Bayesian framework

The present paper describes monotonic effects as embedded in a fully Bayesian framework. We consider every statistical model 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 $p(y)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{6}$$

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

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

Priors for b can be derived based on the *a-priori* expectation regarding the *average difference* between adjacent categories. Any family of prior distributions typically applied to regression coefficients can be applied on b, as well. As a weakly-informative prior for b, we can understand any location shift distribution – such as a normal of student-t distribution – centered around zero and with a scale parameter large enough to allow for large but plausible average differences, while penalizing implausibly large differences. This scale will necessarily depend on the scale of the response distribution and, the range of the monotonic predictor and also on the chosen link-function (Gelman, Simpson, & Betancourt, 2017). Alternatively, one may use an improper flat prior that treats all real values as being equally likely *a-priori* in the hope that the data alone is sufficient to identify b. Importantly, when setting up a prior on b, we do not need to take into account the individual differences between adjacent categories since the latter is fully handled by the simplex parameter $\boldsymbol{\zeta}$.

Setting a prior on the simplex parameter $\boldsymbol{\zeta}$ requires a different approach. A natural choice for a prior on simplex parameters is the Dirichlet distribution, a multivariate generalization of the beta distribution (Frigyik, Kapila, & Gupta, 2010). It is non-zero for all valid simplexes (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 single parameter vector $\boldsymbol{\alpha}$ of the same length as $\boldsymbol{\zeta}$. Its density is defined as

$$f(\boldsymbol{\zeta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{D} \zeta_i^{\alpha_i - 1}, \tag{7}$$

where $B(\boldsymbol{\alpha})$ is a normalizing constant (Balakrishnan, 2014). As the *a-priori* expectation 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 comparison to the sum over $\boldsymbol{\alpha}$ imply higher *a priori* values of ζ_i. Moreover, the higher the sum over $\boldsymbol{\alpha}$, the higher the certainty in each of the proportions w_i.

In the absence 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 $w_i = 1/D$, this prior centers $\boldsymbol{\zeta}$ around a linear trend with large uncertainty and thus appears to be a good default prior in the absence of any problem specific information.

If the prior on b is centered around zero and the prior of $\boldsymbol{\zeta}$ is centered around a linear trend, the implied joint prior of the monotonic effect is centered around zero, with potentally substantial uncertainty around it, depending how uncertain the priors on b and $\boldsymbol{\zeta}$ are. That is, the data has to provide enough evidence for a non-zero effect in order to overcome the prior. The stronger the prior in favor of a zero-effect, the more evidence we need from the data in order to be convinced of a non-zero effect. This property actually enables shrinkage priors for regression coefficient (e.g., Carvalho, Polson, & Scott, 2009; Piironen, Vehtari, & others, 2017) to be applied to monotonic effects.

1.2 Regularizing larger changes between categories

In a Bayesian framework, larger differences between adjacent categories can naturally be penalized—that is made less likely *a-priori*—by means of priors on b and $\boldsymbol{\zeta}$. Importantly, when we speak of penalizing larger differences, we do not mean making the overall functional form smoother. Since monotonic effects are piecewise linear, priors on b or $\boldsymbol{\zeta}$ will not make them look smoother (unless the effect turns out the be exactly linear across all predictor categories). This is an important difference to other approaches such as monotonic regression splines and should be taken into account when interpreting the influence of priors on the obtained parameter estimates.

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 logic behind this choice is straightforward as 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 (see Balakrishnan, 2014):

$$\mathrm{SD}(\zeta_i) = \sqrt{\frac{w_i(1 - w_i)}{\alpha_0 + 1}} \tag{8}$$

where $w_i = \alpha_i/\alpha_0$ is the expectation of the ith component. Although the standard deviation is an imperfect measure of variability for the Dirichlet distribution as the latter is not symmetric in general, we still believe the former to be helpful in better understanding the implications of one's chosen priors. For the default of $a = 1$ and a total of $D + 1 = 5$ categories, we get a rather large standard deviation of $\mathrm{SD}(\zeta_i) = 0.19$. If we set, for example, $a = 5$, we get $\mathrm{SD}(\zeta_i) = 0.09$ and thus much higher certainty in changes of equal size.

Of course, the process of increasing $\boldsymbol{\alpha}$ on average works equally well even if we do not expect all changes to be the same *a-priori*. For instance, if $D = 4$ and we expect a 3-times larger change between the first two categories than between all the other categories with some certainty, we may set $\boldsymbol{\alpha} = (9, 3, 3, 3)$. As a result, we get $w_1 = 1/2$ and esle $w_i = 1/6$. As standard deviations, we get $\mathrm{SD}(\zeta_1) = 0.11$ and $\mathrm{SD}(\zeta_i) = 0.09$ else.

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

$$\mathrm{Cor}(\zeta_i, \zeta_j) = -\frac{w_i w_j}{\sqrt{w_i(1 - w_i)w_j(1 - w_j)}} \tag{9}$$

A possible solution would be to plot the multivariate density of the Dirichlet prior, but this will become more difficult 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.

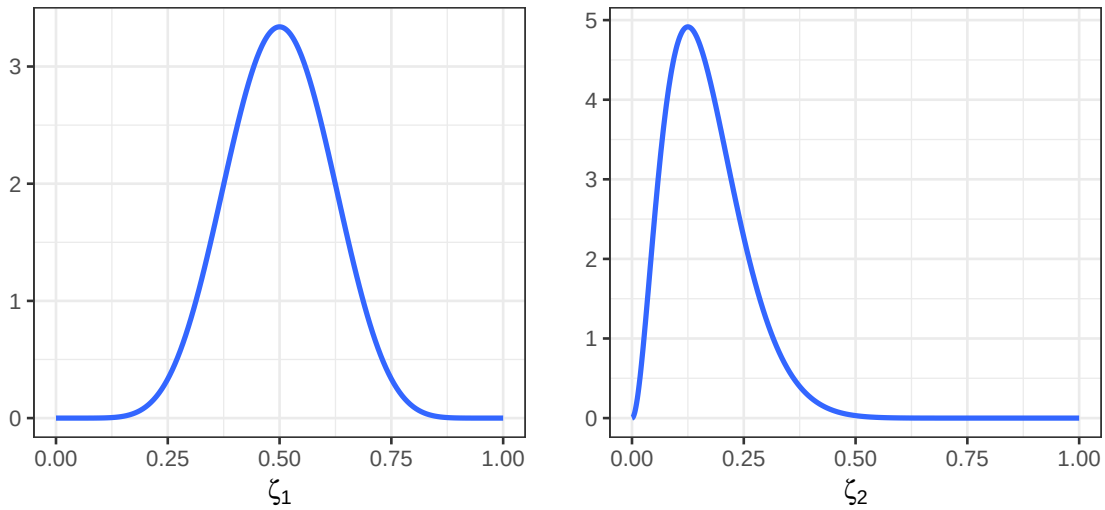

Figure 2. Densities of marginal priors of ζ_1 and ζ_2 for $\boldsymbol{\alpha} = (9, 3, 3, 3)$. The marginal priors of ζ_3 and ζ_4 are in this case identical to the one of ζ_2.

2 Implementation in brms

The brms package (Bürkner, 2017; Bürkner, 2018) provides an interface to fit Bayesian generalized (non-)linear (multilevel) regression models using Stan (Carpenter et al., 2017; Stan Development Team, 2019), 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 distributional 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 different predictor term, which is fully determined by and thus consistent with the basic R formula syntax. If `z` is monotonic as well, then `z` is simply replaced by `mo(z)`. Please note that for models including interactions with monotonic variables, brms will use *different* simplex parameters for different terms of the same monotonic variable (e.g., for the main effect of `x` and the interaction of `x` and `z`). This results is much greater modeling flexibility

as explained in the former section. The variable which should be modeled as monotonic may either be integer valued or an ordered factor. In the latter case, the ordered factor will be transformed to an integer variable with the lowest factor level being identified with zero as described above.

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

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

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

3 Simulations

To verify the correctness of our implementation of monotonic effects and to compare them to other approaches for ordinal predictors, we performed a simulation study. All simulations were done in R (R Core Team, 2018) via the RStudio interface (RStudio Team, 2018). For data preparation and plotting we used packages from the tidyverse (Wickham, 2017) in particular dplyr (Wickham, François, Henry, & Müller, 2019) and ggplot2 (Wickham, 2016).

3.1 Parameter recovery

Before applying a statistical model in practice, we should first make sure that it is able to recover its own parameters (e.g., Cook, Gelman, & Rubin, 2006). This means that if data is simulated from the model under consideration – so that it is the true data generating model – we should, on average, be able to recover the true parameters of the model. What is more, our parameter estimates should have just the right amount of uncertainty so that we are neither overly certain not overly uncertain about the location of the parameter. We may be tempted to just select a few parameter values to work as the ground truth, and evaluate parameter recovery on their basis. However, this is dangerous since we may (accidentally) select parameter values for which the algorithm works particularily well or particularily poorly (Talts, Betancourt, Simpson, Vehtari, & Gelman, 2018). A more robust approach is to sample the ground truth from a distribution of ground truths in each simulation trial and then evaluate the set of estimates against the true distribution.

If an estimation algorithm for a given model suceeds in the procedure described above, we call the algorithm *well calibrated*. In Bayesian statistics, we could equivalently say that

given the prior and the likelihood, we are able to estimate the true posterior distribution of the parameters. This can be formally tested by means of *simulation based calibration* (SBC; Talts et al., 2018), a procedure which works as follows. First, sample true parameter values $\tilde{\theta}$ from the prior, $\tilde{\theta} \sim p(\theta)$, second sample data $\tilde{y}$ from the likelihood, $\tilde{y} \sim p(y \,|\, \tilde{\theta})$. Third, using the algorithm which we want to validate, obtain L samples $\{\theta_1, \ldots, \theta_L\}$ from the estimated posterior distribution, $\{\theta_1, \ldots, \theta_L\} \sim p(\theta|\tilde{y})$. Fourth, for a quantity (or quantities) of interest $f(\theta)$, which can be computed on the basis of the posterior samples (e.g., the individual parameter estimates), count how many values in $\{f(\theta_1), \ldots, f(\theta_L)\}$ fall below the true value $f(\tilde{\theta})$. We call this count the *rank statistic* and denote it by $r(\theta_{1:L} \,|\, \tilde{\theta}, f)$.

If the algorithm is well calibrated for a given model, $\tilde{\theta}$ and $\{\theta_1, \ldots, \theta_L\}$ should be distributed according to the same distribution (Talts et al., 2018). We can verify this by repeating the above steps multiple times (say, $T = 1000$ times). For each repetition t, we compute the rank statistic $r_t(\theta \,|\, \tilde{\theta}, f)$. Afterwards, we create an histogram over $\{r_1(\theta_{1:L} \,|\, \tilde{\theta}, f), \ldots, r_T(\theta_{1:L} \,|\, \tilde{\theta}, f)\}$ and investigate its shape. If the histogram is approximately uniform over $[0, L]$, the algorithm is well calibrated to the model. If it is skewed, the algorithm is biased. If the histogram is U-shaped or inverse-U-shaped, the estimated posterior distribution is narrower or wider, respectively, than the true posterior distribution. We may add confidence intervals to the histograms to indicate the range in which we would the bars to be for a well calibrated quantity. This helps in differentiating actual estimation problems from random simulation noise. The SBC procedure needs to be adjusted slightly for use with autocorrelated samples such as those obtained by MCMC sampling. For more details see Talts et al. (2018).

To analyse the calibration of monotonic models using SBC in practically common settings, we performed a simulation study. We focused on normally distributed response variables:

$$y \sim \text{normal}(\mu, \sigma),$$

where the mean μ is regressed on some monotonic effects as detailed in the following, and σ is the residual standard deviation assumed constant across observations. This resembles a linear regression model except that the predictors were modeled as monotonic effects. We varied μ as either containing the main effect of a single monotonic predictor x:

$$\mu = b_0 + b_1 \,\text{mo}(x, \boldsymbol{\zeta}_x),$$

or the main effects of two monotonic predictors x and z plus their interaction:

$$\mu = b_0 + b_1 \,\text{mo}(x, \boldsymbol{\zeta}_1) + b_2 \,\text{mo}(z, \boldsymbol{\zeta}_2) + b_3 \,\text{mo}(x, \boldsymbol{\zeta}_{31}) \,\text{mo}(z, \boldsymbol{\zeta}_{32}).$$

Further, the dimension of all simplex parameters was $D \in \{4, 10, 50\}$ so that the number of predictor categories of x and z took on values of $D + 1 \in \{4, 11, 51\}$, respectively. In each simulation trial, the values of x and z where sampled uniformly from the set of possible categories $\{0, \ldots, D\}$. The number of observations took on values of $N \in \{50, 200, 1000\}$. As priors for the model parameters, we used Normal$(0, 1)$ distributions for all regression coefficients, uniform Dirichlet distributions of dimension D for all simplexes, and truncated Normal$_+(0, 1)$ distribution for the residual standard deviation σ. Further, regression coefficients were scaled to be independent of the number of predictor categories,

that is, b_1 and b_2 were divided by D and b_3 was divided by D^2. This ensures comparability of model predictions across different values of D. The monotonic models where fitted in Stan via the brms interface using 500 warmup and 500 post-warmup draws from the posterior obtained from a single Markov chain. For each of the $2 \times 3 \times 3 = 18$ conditions, the simulations were repeated $T = 1000$ times.

For brevity's sake, we only show results of selected simulation conditions that are representative of the overall findings. A complete overview of all results is available on Github (https://github.com/paul-buerkner/monotonic-effects-paper). The SBC results for the monotonic main effect and interaction models for $N = 200$ and $D = 4$ are displayed in Figure 3 and 4, respectively. We clearly see that all model parameters are well calibrated under these conditions. Even if the number of parameters P becomes substantially larger than the number of observations N, the model is may be well calibrated as examplified for the interaction model when $N = 50$ and $D = 50$ (see Figure 5).

However, this may not always be the case. In the interactiom model for $N = 1000$ and $D = 50$, we obsere spikes in the histograms at very small and very large parameter values, in particular for the simplex parameters (see Figure 6). This indicates strong autocorrelation in the chains and thus convergence problems of the model (Talts et al., 2018). A closer investigation of the fitted models revealed that most iterations exceeded the maximum treedepth (see Stan Development Team, 2019 for details). This indicates a highly complex posterior distribution which is hard to properly explore by the algorithm. For this model, the reason is the interaction term of two 50-dimensional simplex parameters, which the algorithm fails to explore efficiently (although model predictions are still accurate; see Section 3.2). Increasing the maximum treedepth can resolve this problem but increases the computation time noticeably.

Of course, monotonic effects can be applied in a lot of other modelling settings and so the present results provide no guarantee that they will be well calibrated in cases not studied in the present paper. Generally, we recommed building models specifically tuned to the study design, data, and subject matter knowledge. These models should then be validated as a natural part of the research process using SBC or other validation procedures.

3.2 Comparison to other approaches

To compare the predictive performance of monotonic effects to alternative approaches, which can be used under the same circumstances, we performed another simulation study. We used the same simulation conditions as in Section 3.1 with one exception described in following. As underlying data generating processes, we considered the main effects and interaction models described in Section 3.1 in three different variations: (1) simplex values fixed to $1/D$ implying a linear relationship, (2) simplex values sampled from a uniform dirichlet distribution of dimension D implying a non-linear but monotonic relationship, and (3) simplex values sampled from a uniform dirichlet distribution of dimension D with approximately half of the values having a negative sign implying a non-monotonic relationship.

As alternatives to the monotonic model (abbreviated as MO), we considered simple

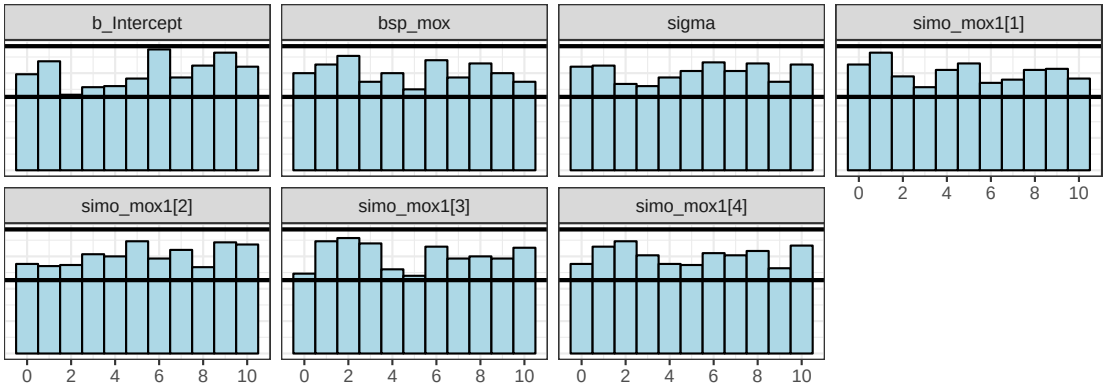

Figure 3. SBC results of the monotonic main effects model for $N = 200$ and $D = 4$. Facets show histograms of different model parameters whose names are taken from brms. Horizontal black lines indicate 99% confidence intervals under the assumption of correct calibration.

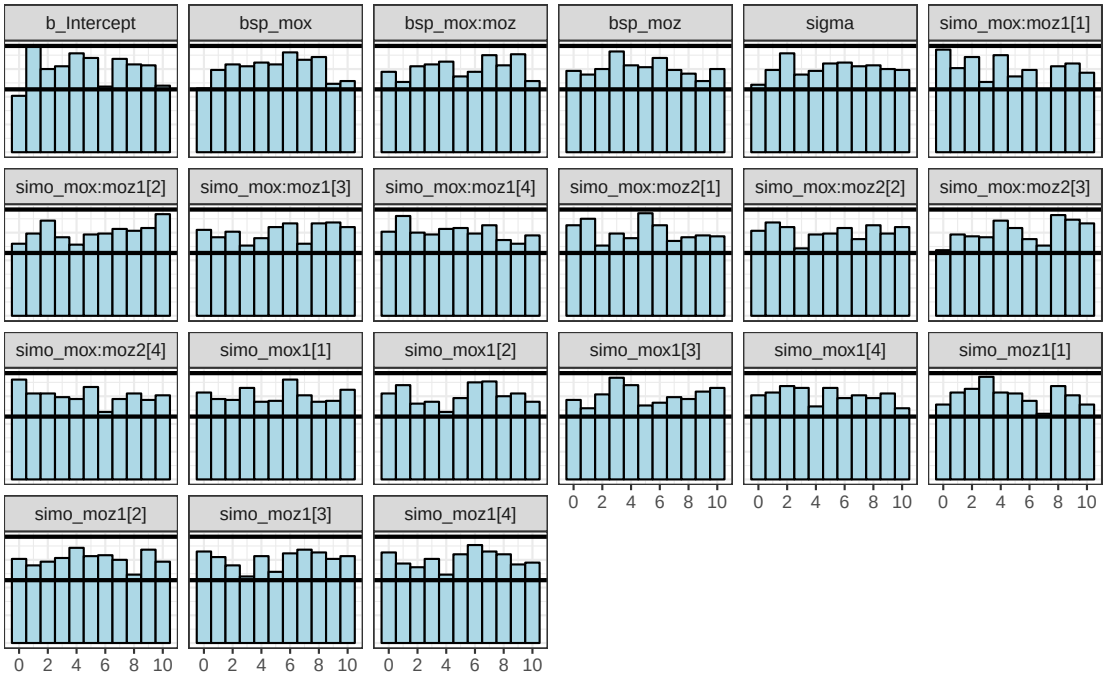

Figure 4. SBC results of the monotonic interaction model for $N = 200$ and $D = 4$. Facets show histograms of different model parameters whose names are taken from brms. Horizontal black lines indicate 99% confidence intervals under the assumption of correct calibration.

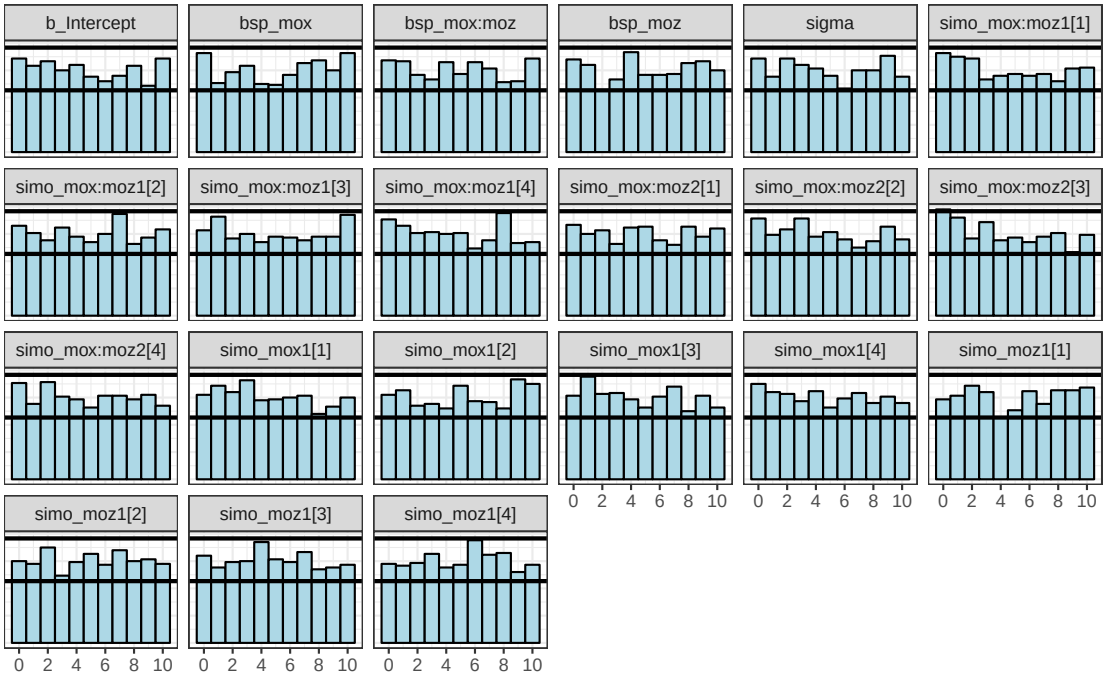

Figure 5. SBC results of the monotonic interaction model for $N = 50$ and $D = 50$. Facets show histograms of different model parameters whose names are taken from brms. For simplex parameters, only the first 4 elements are displayed. Horizontal black lines indicate 99% confidence intervals under the assumption of correct calibration.

linear (LIN) and categorical (CAT) regression[3], isotonic regression (ISO; Barlow et al., 1972; Robertson et al., 1988), penalized ordinal regression (OS; Gertheiss & Tutz, 2009; Gu, 2013), penalized ordinal regression with monotonicity constraint (OSMO; Helwig, 2017), as well as linear and cubic spline models (LS and CS; e.g., Gu, 2013; Helwig, 2016). The latter two are primarily designed for continuous responses but may still perform reasonably well for linear relationships or sufficiently large number of predictor categories. In the following, we will refer to the different approaches using the above introduced abbreviations.

The MO models were fitted with brms using its default priors, that is, without considering the true priors used in the data generating process. This avoids giving these models a possibly unfair advantage as in reality we are unlikely to be aware of the exact data generating process. LIN and CAT models were fitted via the `lm` function, while ISO models were fitted via the `isoreg` function. All penalized regression/splines approaches were fitted in the bigsplines package (Helwig, 2018) using the `bigspline` and `bigssp` functions.

For the main effect models, simulation results are displayed in Figure 7 and 8 showing the models' RMSE under true linearity and monotonicty, respectively. From Figure 7 we see that, under true linearity, LIN performed consistently better than all other models,

[3]Here, categorical refers to treating the predictor(s) as categorical, not the response variable which was assumed to be normally distributed under all simulation conditions.

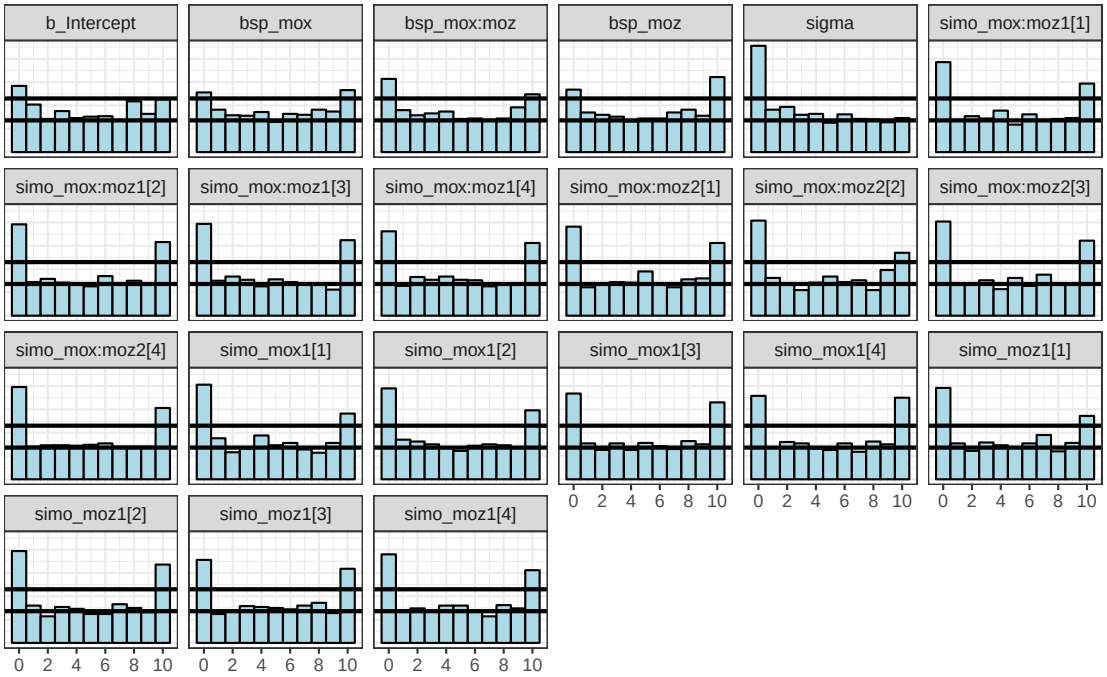

Figure 6. SBC results of the monotonic interaction model for $N = 1000$ and $D = 50$. Facets show histograms of different model parameters whose names are taken from brms. For simplex parameters, only the first 4 elements are displayed. Horizontal black lines indicate 99% confidence intervals under the assumption of correct calibration.

closely follows by CS and MO. Other penalized approaches had slightly but noticably higher RMSEs, while unpenalized approaches such as CAT or ISO models had even higher RMSEs in particular for larger D. From Figure 8 we see that, under true monotonicty, MO models had the same or better predictive performance than all other approaches although the difference to CS, OS, and OSMO models was generally quite small. As expected, under true non-monotonicty, MO models performed worse than models without monotonicity assumption but similarily to other monotonicity assuming models (see Figure 14 in Appendix B).

For the interaction models, simulation results are displayed in Figure 9 and 10 showing the models' RMSE under true linearity and monotonicty, respectively. We did not find implementations for interactions in ISO or OSMO models, which are thus not displayed in the figures. From Figure 9 we see that, under true linearity, LIN and CS models performed better than all other models, closely followed by MO, CS and LS. For small N and large D, MO was even on par with LIN. From 10 we see that, under true monotonicty, MO models had better predictive performance across all conditions compared all other approaches. As expected, under true non-monotonicty, MO models performed worse than models without monotonicity assumption (see Figure 15 in Appendix B).

In summary, in our simulations, MO yielded the same or better predictions than other

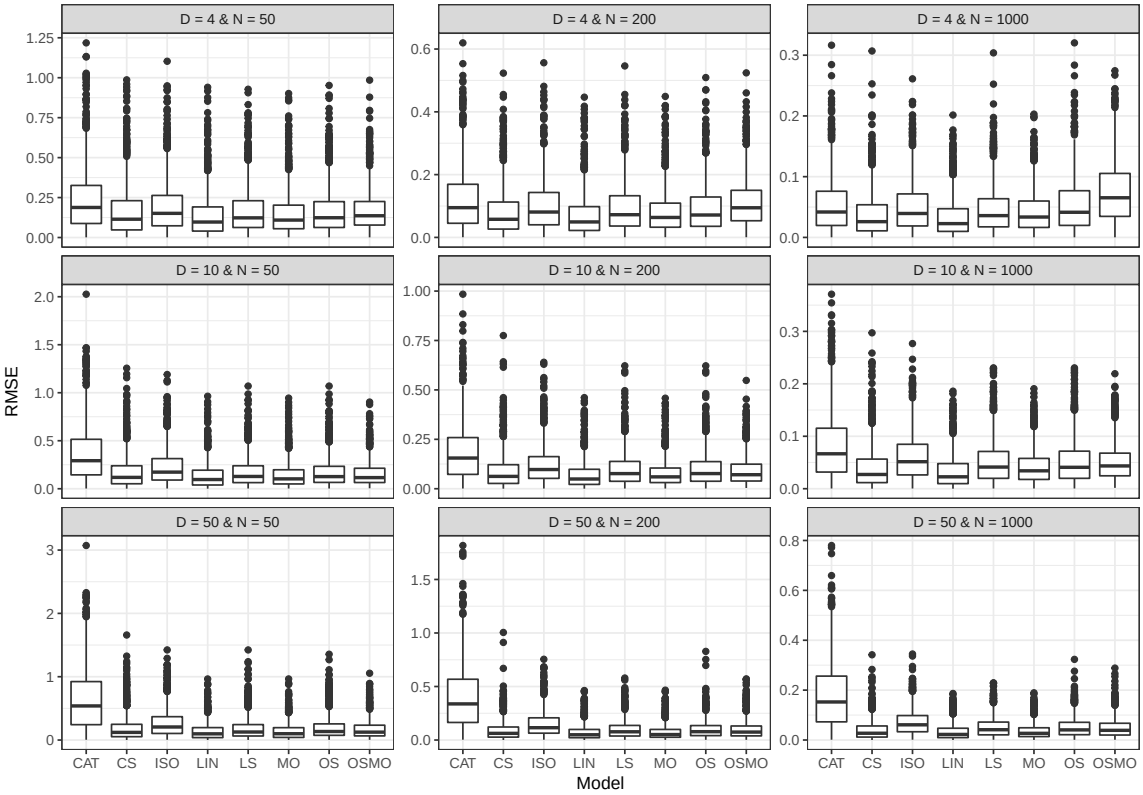

Figure 7. Simulation results for the main effects models under true linearity based on $T = 1000$ simulation trials. Abbreviations: N = number of observations; D = number of categories minus one; CAT = categorical model; CS = cubic spline model; ISO = Isotonic regression model; LIN = linear model; LS = Linear spline model; MO = monotonic model; OS = ordinal spline model; OSMO = ordinal monotonic spline model.

penalized or unpenalized ordinal approaches if the monotonicity assumption was justified. Intuitively, one may expect that MO models tend to overfit the data in cases of small N and comparably large D in particular for interaction models as they have considerably more parameters than observations. However, as evident in our simulations, this is not actually what happens, although we found convergence issues under some of these conditions. The reason for this lies in the joint Dirichlet prior on the simplex parameters: If one particular element of $\boldsymbol{\zeta}$ (i.e., one difference between two adjacent categories) is large, larger values of other elements are automatically penalized (i.e., made more unlikely) due to the sum-to-one constraint on $\boldsymbol{\zeta}$. The same property can be expressed in terms of the negative correlation between two distinct elements of $\boldsymbol{\zeta}$ (see Equation 9). This holds even if the Dirichlet prior is uniform over the set of possible simplexes, which is used as the default prior in brms.

There is also another aspect of the monotonic parameterization that can guard against overfitting. If the scale parameter b is close to zero, there is not much to learn about the corresponding simplex parameter $\boldsymbol{\zeta}$, which will thus have a posterior distribution close to its prior. Still, this uncertainty will not lead to overfitting as changes in $\boldsymbol{\zeta}$ do not influence predictions as long as b is small. This is because the latter controls the overall effect size of

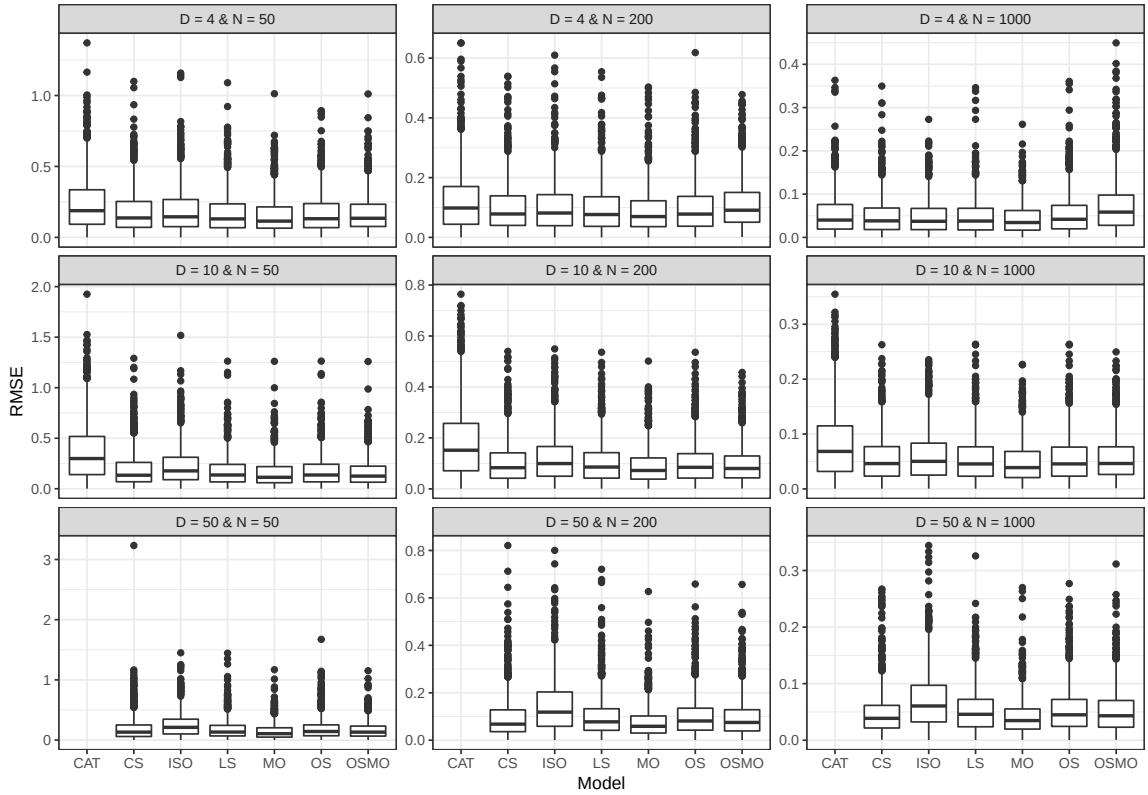

Figure 8. Simulation results for the main effects models under true monotonicity based on $T = 1000$ simulation trials. LIN is not displayed as its RMSE is too large and thus obscures differences between other models. For the same reason, CAT is not displayed for $D = 50$. Abbreviations: N = number of observations; D = number of categories minus one; CAT = categorical model; CS = cubic spline model; ISO = Isotonic regression model; LIN = linear model; LS = Linear spline model; MO = monotonic model; OS = ordinal spline model; OSMO = ordinal monotonic spline model.

the monotonic predictor while $\boldsymbol{\zeta}$ only controls the shape. In other words, the complexity of a monotonic predictor with a close to zero effect naturally reduces to the complexity of a simple linear predictor.

4 Case study: Measures of chronic widespread pain

To illustrate the application of monotonic effects in practice, we will reanalyze data used to validate measures of chronic widespread pain (CWP) from patients' point of view (Cieza et al., 2004; Gertheiss et al., 2011). There is not universally accepted definition of CWP, but "it may be characterized by pain involving several regions of the body, which causes problems in functioning, psychological distress, poor quality of sleep or difficulties in daily life" (Gertheiss et al., 2011, p. 378). The applied CWP measures stem 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

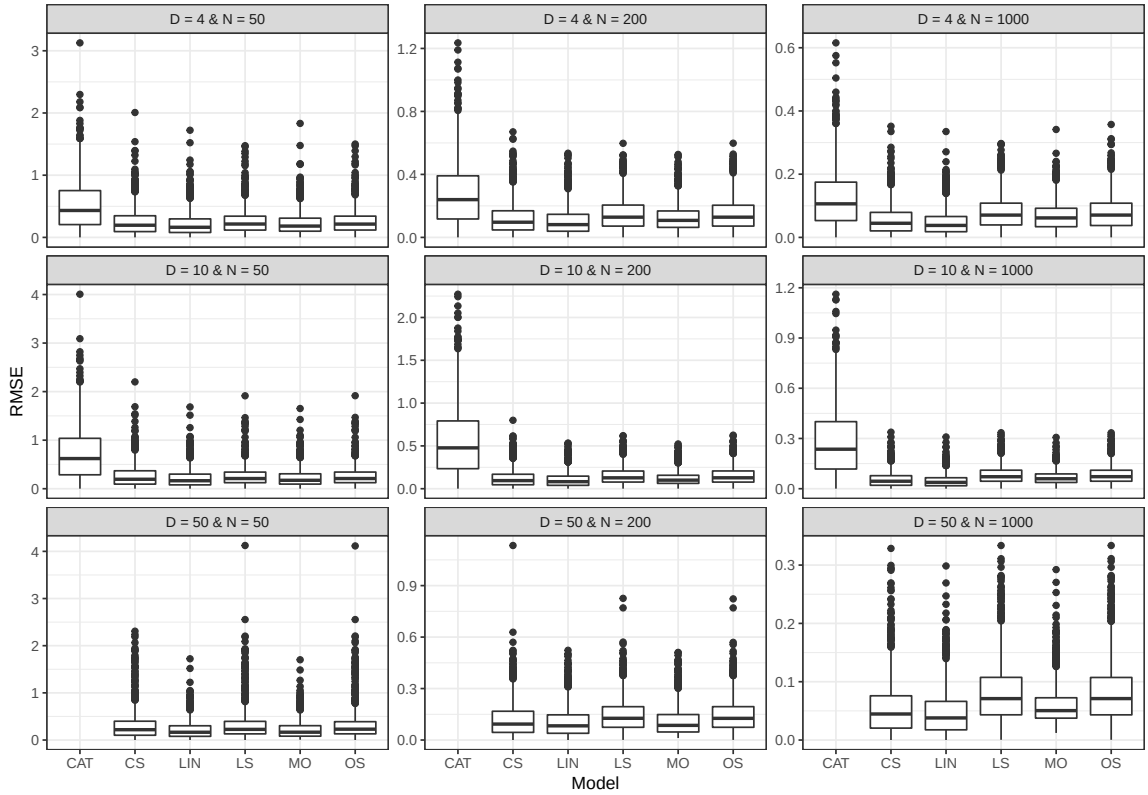

Figure 9. Simulation results for the interaction models under true linearity based on $T = 1000$ simulation trials. CAT is not displayed for $D = 50$ as its RMSE is too large and thus obscures differences between other models. ISO and OSMO are not displayed as they have no corresponding interaction model. Abbreviations: N = number of observations; D = number of categories minus one; CAT = categorical model; CS = cubic spline model; ISO = Isotonic regression model; LIN = linear model; LS = Linear spline model; MO = monotonic model; OS = ordinal spline model; OSMO = ordinal monotonic spline model.

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. (2011) and Cieza et al. (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., 2011 for explanation). Our fully reproducible analysis is available on OSF (https://osf.io/kvrsg/).

In the data set, the subjective physical health (variable `phcs`) ranges from 10.08 to 53.17 with a mean of 32.41 and a standard deviation of 8.17. For the purpose of this case study, we will predict `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

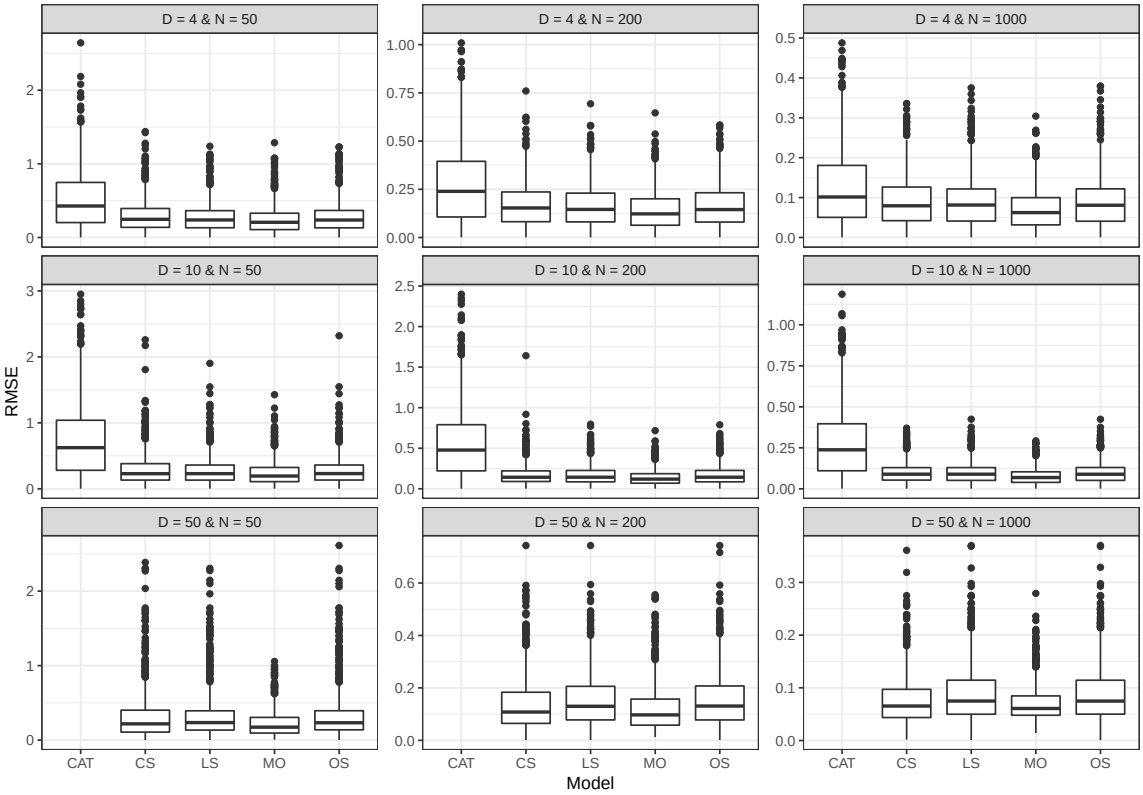

Figure 10. Simulation results for the interaction models under true monotonicity based on $T = 1000$ simulation trials. LIN is not displayed as its RMSE is too large and thus obscures differences between other models. For the same reason, CAT is not displayed for $D = 50$. ISO and OSMO are not displayed as they have no corresponding interaction model. Abbreviations: N = number of observations; D = number of categories minus one; CAT = categorical model; CS = cubic spline model; ISO = Isotonic regression model; LIN = linear model; LS = Linear spline model; MO = monotonic model; OS = ordinal spline model; OSMO = ordinal monotonic spline model.

of `phcs` in the analysis of Gertheiss et al. (2011). 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. Including more or even all of the 67 predictors would be possible as well in theory but barely sensible without principled variable selection techniques. Such techniques have yet to be developed for monotonic effects and are out of the scope of the present paper.

We will start by predicting subjective physical health only by impairments in moving around. For the present example (and also more generally; see Section 3.2), 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 even if priors are similar to the default ones. Based on knowledge about the outcome scale, it is unlikely that a one-point change in any WCP measure will imply a change in the subjective

Table 1
Summary of parameter estimates for impairments in moving around.

	Estimate	l-95% CI	u-95% CI
intercept	37.19	35.80	38.64
slope	-2.70	-3.34	-2.08
simo[1]	0.38	0.23	0.55
simo[2]	0.11	0.01	0.27
simo[3]	0.36	0.15	0.55
simo[4]	0.15	0.01	0.35

Note. simo = simplex parameter of the monotonic effect; Estimate = posterior mean; CI = credible interval based on quantiles.

physical health by more than 5 points. We code this expectation as a normal(0, 2.5) prior on the scale parameters b. That way, $|b|$ will only exceed 2.5 and 5 outcome points with probabilities of roughly 32% and 5%, respectively. With regard to the shape of the effect of "moving around", we have no particular prior expectations and thus assume a uniform Dirichlet prior as explained in Section 1.2, which is also the default in brms. In brms, we can specify the above priors by means of the following code:

```
library(brms)
prior_b <- prior(normal(0, 2.5), class = "b")
prior_s1 <- 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 name `"mod4551"` are constructed as `mo<variable><index>`, where `<index> = 1` unless a single regression term contains multiple simplexes – which is only the case for interactions of monotonic effects. Finally, we fit the model in brms via

```
fit1 <- brm(phcs ~ mo(d455), data = cwp, prior = prior_b + prior_s1)
```

As illustrated in the center of Figure 11, impairments in moving around show a strong negative relationship to subjective physical health. Moreover, this relationship is clearly (at least visually) non-linear. Changes in the outcome are strongest between the first two categories and the third to fourth category. This impression is confirmed by the summary estimates of the regression parameters (see Table 1) as `simo[1]` and `simo[3]` have the largest estimates. For example, the estimate of `simo[1] =` 0.38 implies that 38% of the total change in subjective physical health due to impairments in walking happens between the first two predictor categories. Further, the estimate of `slope =` -2.70 implies that *on average* the subjective physical health *decreases* by 2.70 per increase of impairments in walking by one category.

Next, let us compare the monotonic model to a linear and an unordered categorical model, which are fitted as follows:

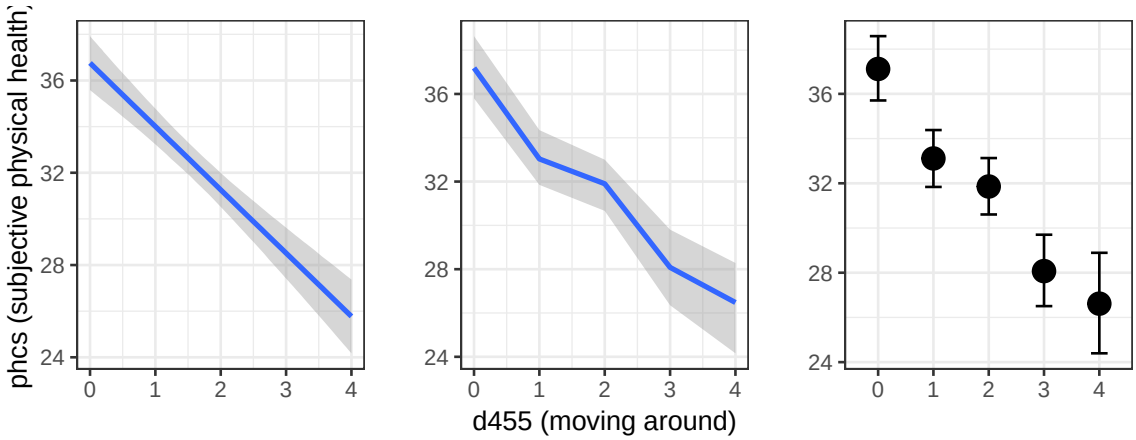

Figure 11. Effects of impairments in moving around on subjective physical health. Left: linear model; center: monotonic model; right: categorical model.

```
fit2 <- brm(phcs ~ d455, data = cwp, prior = prior_b)
fit3 <- brm(phcs ~ d455c, data = cwp, prior = prior_b)
```

The variable `d455c` denotes the categorical version of `d455` to which we applied sequential difference coding. Results are visualized on the left and right-hand side of Figure 11. To compare models, we use approximate leave-one-out cross-validation (Vehtari, Gelman, & Gabry, 2017) together with corresponding information criteria and Akaike model weights (Vehtari et al., 2017; Wagenmakers & Farrell, 2004)[4]:

```
loo_compare(loo(fit1), loo(fit2), loo(fit3))
model_weights(fit1, fit2, fit3, weights = "loo")
```

As shown in Table 2, the monotonic models fits better than the categorical model followed by the linear model although the differences between the three models is not that substantial. Looking closer at the results, we see that the effective number of parameters is somewhat smaller for the monotonic model than those of the categorical model; about the same difference as we see in the corresponding ELPD difference. Thus, the better predictive performance of the monotonic model is primarily driven by it being more parisimonous than the categorical model. Together, this provides evidence that the monotonicity assumption for the effect of predictor `d455` is justified by the data.

Next, we will use both impairments in walking (variable `d450`) and in moving around (variable `d455`) to predict the subjective physical health. 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. Thus, the corresponding prior requires a vector of reduced size.

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

Table 2
Comparison of models fit1 to fit3 based on approximate leave-one-out cross-validation.

	ELPD-LOO	ELPD-Diff	SE-Diff	P-LOO	LOOIC	Akaike-Weight
fit1	-1,439.54	0.00	0.00	4.79	2,879.09	0.44
fit3	-1,439.85	-0.30	0.23	4.96	2,879.70	0.32
fit2	-1,440.14	-0.59	1.90	2.90	2,880.27	0.24

Note. ELPD-LOO = expected log posterior predictive density (higher is better); ELPD-DIFF = difference in ELPD values compared to the best model. SE-DIFF = standard error of the ELPD difference. P-LOO = effective number of model parameters (lower is better); LOOIC: leave-one-out information criterion (lower is better); Akaike-Weight = Model weight based on the LOOIC values (higher is better).

```
prior_s2 <-
  prior(dirichlet(1, 1, 1), class = "simo", coef = "mod4501") +
  prior(dirichlet(1, 1, 1, 1), class = "simo", coef = "mod4551")
```

We fit the monotonic, linear, and categorical models as follows:

```
fit4 <- brm(phcs ~ mo(d450) + mo(d455), data = cwp,
            prior = prior_b + prior_s2)
fit5 <- brm(phcs ~ d450 + d455, data = cwp, prior = prior_b)
fit6 <- brm(phcs ~ d450c + d455c, data = cwp, prior = prior_b)
```

Conditional predictions of the three models are visualized in Figure 12. As visible on the right-hand side of Figure 12, the effect of moving around seems to be no longer monotonic when controlling for the effect of walking. Thus, we would expect the categorical model to now show better predictions than the monotonic model. As can be seen in Table 3, this is indeed what happens as the categorical model has a higher ELPD value and corresponding model weight. That is, from a purely predictive perspective, we will likely prefer the categorical model. However, from a theoretical perspective, the situation may be different as it is more plausible that a change to worse in moving around always leads (in expectation) to a reduction in subjective physical health no matter the impairments in walking individuals may have. That is, even for impairments in walking held constant, the effect of impairments in moving around should still be monotonically decreasing. The fact that this is not strictly the case in the present data is clearly the result of the dependency structure between the two predictors as well as a large number of possible confounders that we did not account for in the present analysis. Just from the present data, it remains unclear how well the (non-)monotonicty will generalize to other samples or populations of impaired individuals.

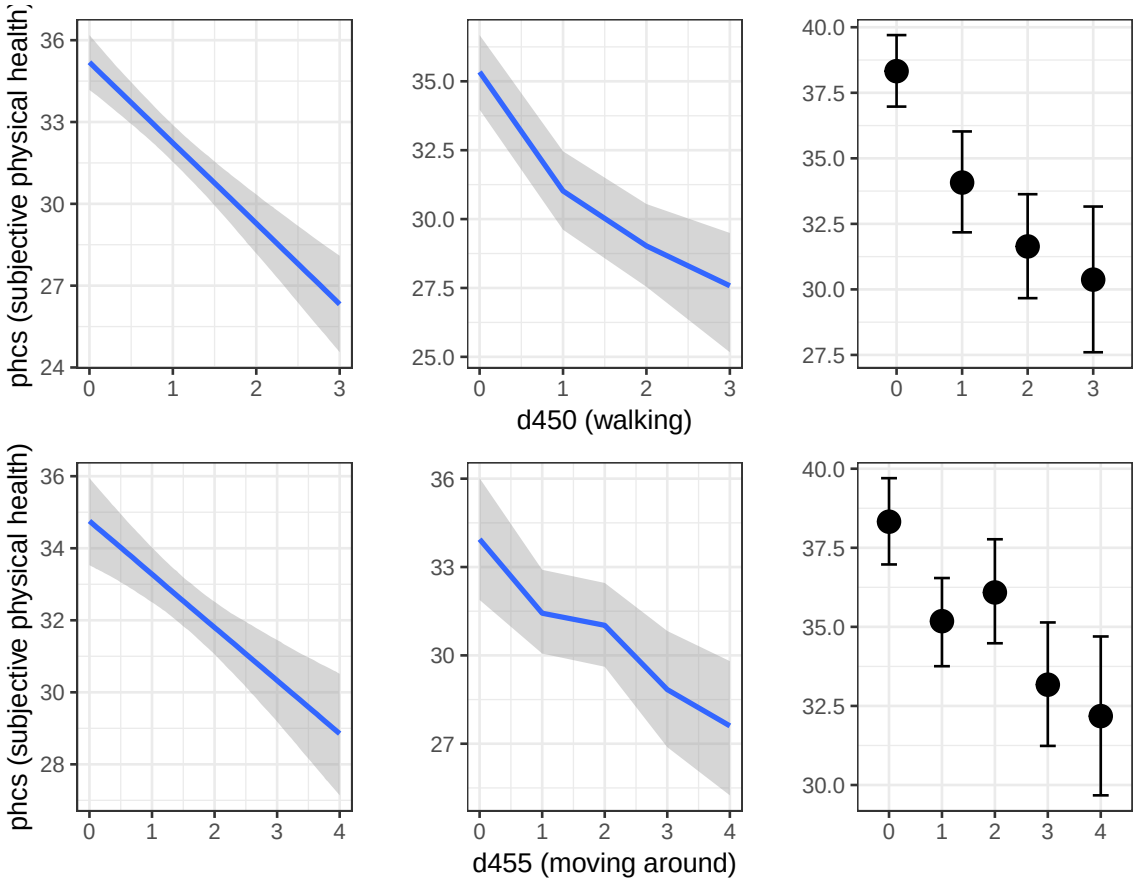

Figure 12. Effects of impairments in walking and in moving around on subjective physical health. Left: linear model; center: monotonic model; right: categorical model.

Table 3
Comparison of models fit4 to fit6 based on approximate leave-one-out cross-validation.

	ELPD-LOO	ELPD-Diff	SE-Diff	P-LOO	LOOIC	Akaike-Weight
fit6	-1,411.48	0.00	0.00	7.08	2,822.97	0.71
fit4	-1,412.45	-0.96	1.47	6.22	2,824.89	0.27
fit5	-1,415.01	-3.53	3.36	3.62	2,830.03	0.02

Note. ELPD-LOO = expected log posterior predictive density (higher is better); ELPD-DIFF = difference in ELPD values compared to the best model. SE-DIFF = standard error of the ELPD difference. P-LOO = effective number of model parameters (lower is better); LOOIC: leave-one-out information criterion (lower is better); Akaike-Weight = Model weight based on the LOOIC values (higher is better).

5 Discussion

In the present paper, we proposed a new approach to including *monotonic effects* of ordinal predictors in regression models. The proposed parameterization not only ensures monotonicity but also naturally regularizes the model and its predictions even without the usage of strong prior information. Thus, monotonic effects share important aspects with existing methods for modeling ordinal predictors. Moreover, monotonic effects nicely integrate into the framework of generalized linear regression and can even be used within multilevel models. By making an informed decision about the parameterization of interactions with monotonic effects, different kinds of monotonicity can be modeled depending on the research question and a-priori information available. Monotonic effects are fully supported in the brms R package, which fits Bayesian regression models using Stan and provides an intuitive user interface based on widely known R formula syntax. To date, ordinal predictors are still mostly treated as either nominal or metric thus under- or overstating the contained information. Monotonic effects avoid these problems but still allow for an intuitive interpretation of the estimated parameters. In summary, we think that monotonic effects provide a useful tool for handling of ordinal predictors in regression models in situation where the monotonicity assumption is justified.

One potential problem in the Bayesian estimation of monotonic effects is that elements of a simplex tend to be negatively correlated, sometimes rather strongly, thus making MCMC sampling more difficult (Hoffman & Gelman, 2014). However, due to 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), these problems may be alleviated when fitting monotonic effects in Stan – either directly or indirectly through brms. Indeed, in the models estimated for the purpose of this paper, sampling efficiency and convergence were good and rarely worse that when using a purely linear approach. The only exceptions were monotonic interaction models with a very high number of predictor categories (i.e., 51 in our simulation). For that many predictor categories, it may be easier to fit (monotonic) splines or similar models which require fewer parameters.

With respect to predictions, monotonic effects performed very well under all simulation conditions where the monotonicity assumption was justified, even in those where convergence was an issue. More precisely, monotonic effects made similar or better predictions than other penalized ordinal approaches such as ordinal splines and much better predictions than unpenalized approaches such as standard isotonic regression or approaches treating the predictors as categorical. This nicely illustrates the advantages of a fully Bayesian approach, where joint priors, even weakly informative ones, can regularize the model parameters and ultimately lead to improved predictions. It has to be noted that all penalized/regularized approaches generally performed well in our simulations and differences between them were comparably small (which we believe is a good sign for the validity of these methods in general). Where monotonic effects differ from previously proposed approaches is their conceptualization as generalizations of continuous predictor terms, rather then as (penalized) categorial predictor means. This allows for the intuitive interpretation of the scale parameter as an ordinary regression coefficient, except that we do not assume the shape of the

relationship as linear, but more generally as monotonic. It is this clear separation between the strength of the relationship and its shape that, in our opinion, makes monotonic effects very appealing for interpretation and communication.

This separation is especially advantageous when dealing with interactions of ordinal predictors. If we used any type of categorical coding (e.g., dummy coding) for the interaction of two ordinal predictors, the number of regression coefficients would increase quadratically with the number of predictor categories, which complicates interpretation. When working with monotonic effects, in contrast, we would only have three related regression coefficients (two for the main effects and one for the interaction), which would essentially have the same interpretative complexity as a linear model with the ordinal predictors treated as continuous. Of course, intepreting the shape parameters of the main effects and interactions will again increase the complexity to a level similar to what is implied by categorical coding. However, often we may not be interested in the exact shape of the effect, in which case it would simply be sufficient to know that the shape has been taken into account by the model. Of course, there is nothing wrong with directly reporting and interpreting the estimated category means (or their contrasts), which, depending on one's preferences and the overall complexity of the model, may also be seen as more intuitive (e.g., see Barlow et al., 1972; Danaher et al., 2012; Gertheiss & Tutz, 2009; Klugkist & Mulder, 2008; Mulder & Raftery, 2019 for related approaches).

Regardless of what formulation one chooses, the assumption of monotonicity is critical and needs to be theoretically justified and/or statistically investigated. A general approach to the latter is to fit one model with and one without monotonicity constraint and then compare the two models using one's prefered criteria of model fit. From a Bayesian perspective, we could use, for example, Bayes factors or cross-validation procedures/information critera depending on whether we aim to penalize prior or posterior complexity, respectively (Gelman et al., 2013; Hoijtink, Klugkist, & Boelen, 2008). In our case study, we applied approximate leave-one-out cross-valdiation for this purpose but other criteria would have been possible as well (e.g., see Mulder & Raftery, 2019).

As already pointed out earlier, one important distinction within the class of models dealing with ordinal predictors is the assumption about the generative process of these variables. In our monotonic effects approach, as well as in a lot of other prominent approaches (e.g., Klugkist & Mulder, 2008; Gertheiss & Tutz, 2009; Gu, 2013), we treat ordinal predictors as manifest variables. Specifically, we do not make the assumption that the observations originate from the categorization of a latent continuous variable as is commonly the case in latent variable models (e.g., Winship & Mare, 1984; Finney & DiStefano, 2006; Jöreskog, 1994; Lei, 2009). Such a latent variable assumption may be sensible if the ordinal predictor represents the chosen categories of a Likert item, for instance, with the intention of measuring a latent psychological construct, but would certainly not be sensible in some other cases, for instance, if the categories were known discrete points in time. Thus, both approaches seem valid in our opinion and are called for in different settings and research questions. We would like to note again that our proposed approach is targeted towards data settings where modeling manifest observations is desired, and that if researchers desire to examine a latent relationship, they are advised to use other methods.

Although our primary focus was the use of monotonic effects for modeling strictly ordinal predictors, we want to point out that they 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.

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). As we prefer a fully Bayesian approach to statistical modeling, we did not dive deeper into this direction. However, we still believe that developing frequentist estimators and corresponding uncertainty estimates for more complex monotonic models including, for instance, interactions or multilevel structure, may be a worthwhile endeavor for future research. Lastly, we want to note that the general idea of monotonic effects should also generalize 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.

6 Acknowledgements

We would like to thank Marie Beisemann, Andrew Rutherford, and three anonymous reviewers for valuable comments on earlier versions of the paper.

References

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

Alvarez, R. M., Bailey, D., & Katz, J. N. (2011). An empirical bayes approach to estimating ordinal treatment effects. *Political Analysis*, *19*(1), 20–31.

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*, *10*(1), 395–411.

Bürkner, P.-C., & Vuorre, M. (2018). Ordinal regression models in psychological research: A tutorial. Retrieved from https://psyarxiv.com/x8swp/

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

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. (1992). *Statistical models in S* (Vol. 251). Wadsworth & Brooks/Cole Advanced Books & Software.

Christensen, R. H. B. (2018). ordinal—Regression models for ordinal data. Retrieved from http://www.cran.r-project.org/package=ordinal/

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*(1), 63–68.

Cook, S. R., Gelman, A., & Rubin, D. B. (2006). Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, *15*(3), 675–692.

Danaher, M. R., Roy, A., Chen, Z., Mumford, S. L., & Schisterman, E. F. (2012). Minkowski–weyl priors for models with parameter constraints: An analysis of the biocycle study. *Journal of the American Statistical Association*, *107*(500), 1395–1409.

Dykstra, R. L., & Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, *10*(3), 708–716.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural Equation Modeling: A Second Course*, *10*(6), 269–314.

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.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555–567. doi:10.3390/e19100555

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. (2011). 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. (2011). 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.

Gu, C. (2013). *Smoothing spline anova models* (Vol. 297). Springer Science & Business Media.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in s* (pp. 249–307). Routledge.

He, X., & Shi, P. (1998). Monotone b-spline smoothing. *Journal of the American Statistical Association*, *93*(442), 643–650.

Helwig, N. E. (2016). Efficient estimation of variance components in nonparametric mixed-effects models with large samples. *Statistics and Computing*, *26*(6), 1319–1336.

Helwig, N. E. (2017). Regression with ordered predictors via ordinal smoothing splines. *Frontiers in Applied Mathematics and Statistics*, *3*, 15.

Helwig, N. E. (2018). *Bigsplines: Smoothing splines for large samples.* Retrieved from https://CRAN.R-project.org/package=bigsplines

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.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses.* Springer Science & Business Media.

Jöreskog, K. G. (1994). Structural equation modeling with ordinal variables. *Lecture Notes-Monograph Series*, 297–310.

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*, *46*(4), 1071–1085.

Klugkist, I., & Mulder, J. (2008). Bayesian estimation for inequality constrained analysis of variance. In *Bayesian evaluation of informative hypotheses* (pp. 27–52). Springer.

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial Introduction with R* (2nd Edition.). 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.

Lei, P.-W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*(3), 495.

Leitenstorfer, F., & Tutz, G. (2006). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, *8*(3), 654–673.

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)*, *42*(2), 109–142.

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

Mulder, J., & Raftery, A. E. (2019). BIC extensions for order-constrained model selection. *Sociological Methods & Research*.

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

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.

RStudio Team. (2018). RStudio: Integrated development for R. *RStudio, Inc., Boston, MA URL Http://Www.rstudio.com*, *42*.

Rufibach, K. (2010). An active set algorithm to estimate parameters in generalized linear models with ordered predictors. *Computational Statistics & Data Analysis*, *54*(6), 1442–1456.

Stan Development Team. (2019). *Stan modeling language: User's guide and reference manual*. Retrieved from http://mc-stan.org/manual.html

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv Preprint arXiv:1804.06788*.

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.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Wang, W., & Small, D. S. (2015). Monotone b-spline smoothing for a generalized linear model response. *The American Statistician*, *69*(1), 28–33.

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*, *30*(6), 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*, 3571–3594.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, 512–525.

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. *Journal of Statistical Software*, *65*(5), 1–33.

Appendix

Appendix A: Mathematical Proofs

Proof. A1: Monotonicity. For all values x between 0 and $D-1$, we have

$$b\,\mathrm{mo}(x+1,\boldsymbol{\zeta}) - b\,\mathrm{mo}(x,\boldsymbol{\zeta}) = bD\sum_{i=1}^{x+1}\zeta_i - bD\sum_{i=1}^{x}\zeta_i = bD\zeta_{x+1}. \tag{10}$$

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

Proof. A2: Equivalence to categorical isotonic regression. Consider a simple linear model of a continuous response y regressed on a categorical predictor x with categories $j\in\{0,\ldots,D\}$. Further, let μ_j be the group mean of category j with respect to the response variable. Then the model for observation n can be written as

$$y_n = \mu_{x_n} + e_n, \tag{11}$$

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

$$y_n = b_0 + b_1\,D\sum_{i=1}^{x_n}\zeta_i + e_n. \tag{12}$$

Hence, we can identify μ_0 with b_0 and μ_j with $b_0+b_1D\sum_{i=1}^{j}\zeta_i$ for $j>0$. This identification is bijective within the set of order-constraint μ. □

Proof. A3: Proposition 2.1. Under the stated assumptions, we can, without loss of generality, write the linear predictor $\eta=\eta(x)$ as

$$\eta(x) = b_0 + \sum_{k=1}^{K} b_k D_k \sum_{i=1}^{x}\zeta_i = b_0 + \left(\sum_{k=1}^{K} b_k D_k\right)\left(\sum_{i=1}^{x}\zeta_i\right). \tag{13}$$

Since all other predictors have been fixed to some constants, their contribution to η can be absorbed by the intercept b_0 and the regression coefficients b_1 to b_K which are all related to x. If we define $b=\sum_{i=1}^{K} b_i D_i$ we see that $\eta(x)$ is monotonic in x with the sign of the effect determined by the sign of b. □

Proof. A4: Counter example to conditional monotonicity for varying simplex parameters. Consider the situation shown in Figure 13, where quite clearly, the effect of $\boldsymbol{x}$ is monotonic for group G, but non-monotonic for group H. Suppose further that we named the grouping variable $\boldsymbol{z}$ and applied dummy coding such that $G=0$ and $H=1$. Using different simplex parameters for the main effect of $\boldsymbol{x}$ and the interaction effects between $\boldsymbol{x}$ and $\boldsymbol{z}$, the linear predictor reads as follows:

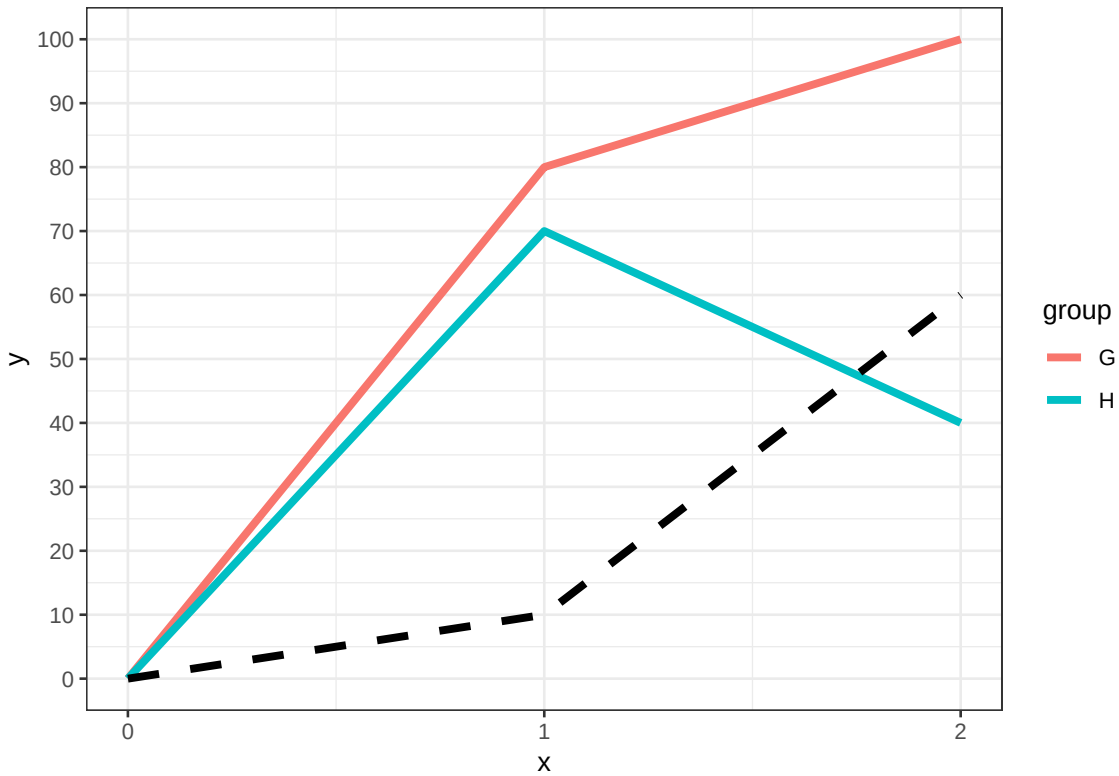

Figure 13. Counter example to the conditional monotonicity for varying simplex parameters. The dashed line shows the difference between the groups G and H as a function of $\boldsymbol{x}$.

$$\eta(x, z) = b_0 + b_1\, z + b_2\ \mathrm{mo}(x, \boldsymbol{\zeta}_2) + b_3\, z\ \mathrm{mo}(x, \boldsymbol{\zeta}_3) \tag{14}$$

Clearly, $b_0 = 0$. For group G this implies $\eta(x, 0) = b_2\ \mathrm{mo}(x, \boldsymbol{\zeta}_2)$ so that $b_2 = 50$ as well as $\boldsymbol{\zeta}_2 = (0.8, 0.2)$ are completely defined by the curve of group G. For group H, we have

$$\eta(x, 1) = b_0 + b_1 + b_2\ \mathrm{mo}(x, \boldsymbol{\zeta}_2) + b_3\, \mathrm{mo}(x, \boldsymbol{\zeta}_3). \tag{15}$$

As the curve of group H starts at the origin, we have $b_1 = 0$. Due to the chosen parameterization of $\boldsymbol{z}$, the term $b_3\, \mathrm{mo}(x, \boldsymbol{\zeta}_3)$ models the *difference* between in the effect of $\boldsymbol{x}$ between the two groups, which visualized as a dashed line in Figure 13 and is clearly monotonic. Consequently, we have $b_3 = 30$ and $\boldsymbol{\zeta}_3 = (\frac{1}{6}, \frac{5}{6})$. Although the assumptions of the monotonic effects are fully met, the effect of $\boldsymbol{x}$ in group H is non-monotonic. Thus, $\boldsymbol{x}$ is not conditionally monotonic given $\boldsymbol{z}$.

□

Appendix B: Additional Simulation Results

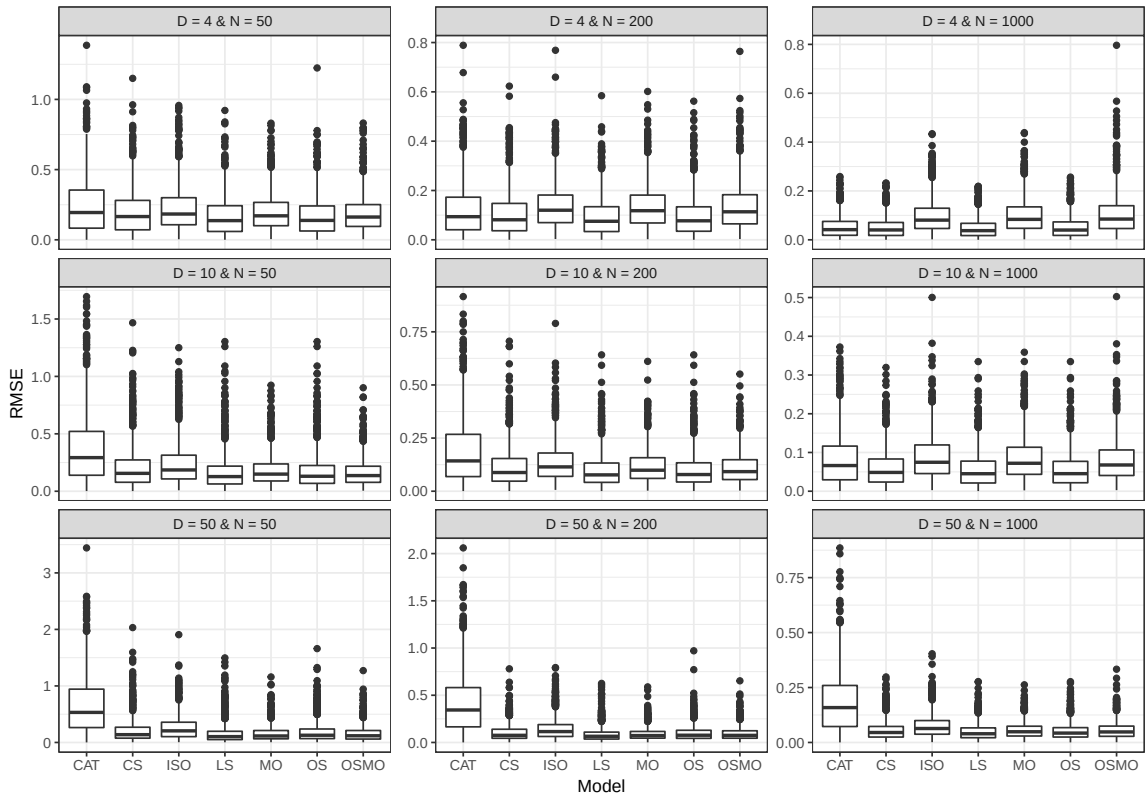

Figure 14. Simulation results for the main effects models under true non-monotonicity based on $T = 1000$ simulation trials. LIN is not displayed as its RMSE is too large and thus obscures differences between other models. Abbreviations: N = number of observations; D = number of categories minus one; CAT = categorical model; CS = cubic spline model; ISO = Isotonic regression model; LIN = linear model; LS = Linear spline model; MO = monotonic model; OS = ordinal spline model; OSMO = ordinal monotonic spline model.

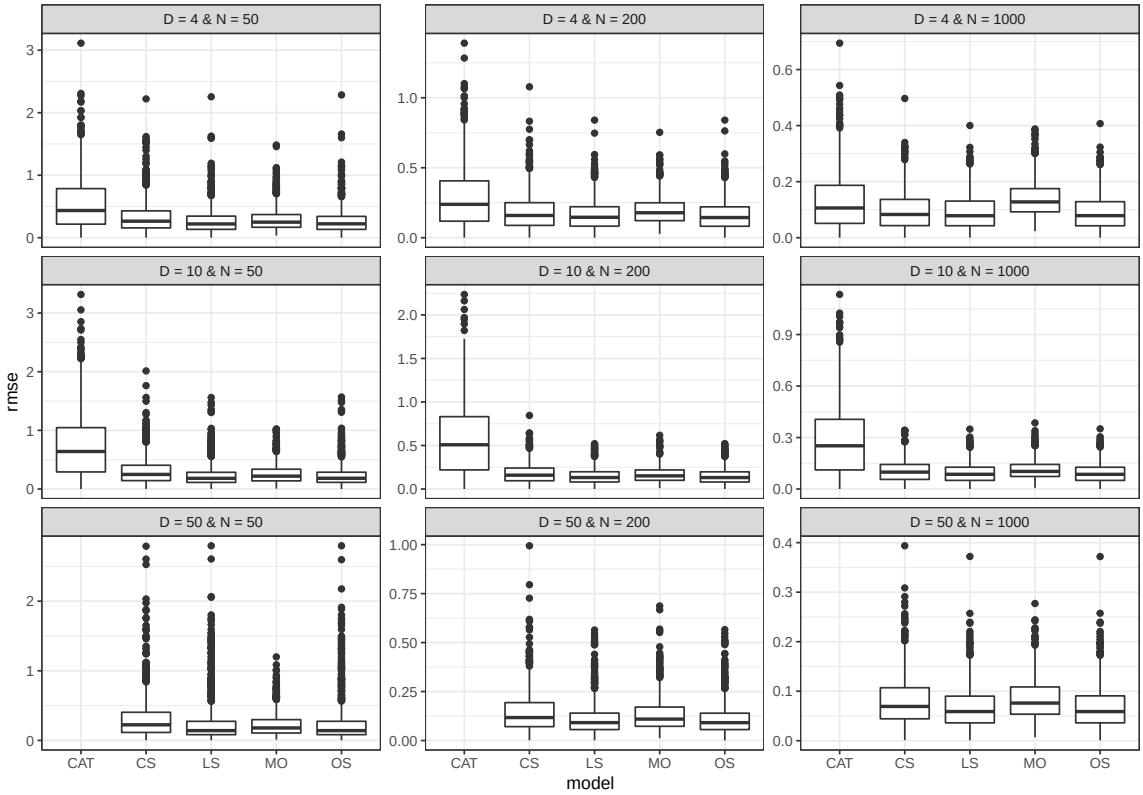

Figure 15. Simulation results for the interaction models under true non-monotonicity based on $T = 1000$ simulation trials. LIN is not displayed as its RMSE is too large and thus obscures differences between other models. For the same reason, CAT is not displayed for $D = 50$. ISO and OSMO are not displayed as they have no corresponding interaction model. Abbreviations: N = number of observations; D = number of categories minus one; CAT = categorical model; CS = cubic spline model; ISO = Isotonic regression model; LIN = linear model; LS = Linear spline model; MO = monotonic model; OS = ordinal spline model; OSMO = ordinal monotonic spline model.