

# Bayesian Model Choice in Cumulative Link Ordinal Regression Models

Trevelyan J. McKinley\*, Michelle Morters\*, and James L. N. Wood\*

**Abstract.** The use of the proportional odds (PO) model for ordinal regression is ubiquitous in the literature. If the assumption of parallel lines does not hold for the data, then an alternative is to specify a non-proportional odds (NPO) model, where the regression parameters are allowed to vary depending on the level of the response. However, it is often difficult to fit these models, and challenges regarding model choice and fitting are further compounded if there are a large number of explanatory variables. We make two contributions towards tackling these issues: firstly, we develop a Bayesian method for fitting these models, that ensures the stochastic ordering conditions hold for an arbitrary finite range of the explanatory variables, allowing NPO models to be fitted to any observed data set. Secondly, we use reversible-jump Markov chain Monte Carlo to allow the model to choose between PO and NPO structures for each explanatory variable, and show how variable selection can be incorporated. These methods can be adapted for any monotonic increasing link functions. We illustrate the utility of these approaches on novel data from a longitudinal study of individual-level risk factors affecting body condition score in a dog population in Zenzele, South Africa.

**Keywords:** Bayesian inference, ordinal regression, Markov chain Monte Carlo, reversible-jump, Bayesian model choice.

## 1 Introduction

The most common regression models for analysing ordinal data fall under the set of cumulative link models, in which the categories of the response variable can be modelled as contiguous intervals on some continuous scale (McCullagh, 1980). A general monotonic increasing link function is then used to map these intervals from the continuous scale onto the interval  $(0, 1)$ . The choice of link function will generally lead to qualitatively similar model fits, and so can be chosen on the basis of interpretability (McCullagh, 1980) or convenient mathematical properties (e.g. Albert and Chib, 1993). To this end we concentrate on the logistic link, leading to comparisons of the cumulative odds.

A popular implementation assumes that the relationship between the cumulative log-odds and the explanatory variables does not depend on the response category (the *proportional odds* [PO] model—McCullagh, 1980). Under simple constraints, this implementation guarantees that the model exhibits *stochastic ordering* (i.e. it ensures that for  $J$  ordered groups, the cumulative probability of belonging to group  $j$  is less than or equal to the cumulative probability of belonging to group  $j + 1$ , for  $j = 1, \dots, J - 1$ ). As

---

\*Disease Dynamics Unit, Department of Veterinary Medicine, University of Cambridge, UK, [tjm44@cam.ac.uk](mailto:tjm44@cam.ac.uk)

a direct result of its simplicity and ease-of-interpretation, the PO model is commonly used in the literature.

An alternative is to allow the relationship between the response and explanatory variables to vary by response category (the *non-proportional odds* [NPO] model). Whilst more flexible than the PO model, the use of NPO models in the literature is limited, since the stochastic ordering conditions will only hold for a limited range of values of the explanatory variables (Agresti, 2010). This means that many traditional fitting mechanisms can fail to fit (Tutz and Scholz, 2003), and in the case where the PO assumption fails to hold, it may be difficult to find estimates for the general NPO model. A useful alternative is to fit a set of  $J - 1$  separate binary logistic regression models to each cumulative logit separately (Bender and Grouven, 1998; Cole et al., 2004), however, the parameter estimates for the regression parameters do not always correspond to those obtained from the general model (Tutz and Scholz, 2003). The *partial proportional odds* (PPO) model (Peterson and Harrell, 1990) allows for a mixture of PO and NPO variables to be included, though the structure for each variable must be specified in advance. Various approaches have been developed as a means of assessing whether the PO assumption is appropriate for a set of given variables (see e.g. Brant, 1990; Agresti, 2010), but these are difficult to apply to large numbers of variables, particularly if there are interactions between some of them that may impact the relationship.

Another alternative approach—when the response variable is such that to belong to a particular category it is necessary to pass through all previous categories in turn—is to use continuation ratios (Feinberg, 1980). Good reviews of general ordinal regression frameworks can be found in Ananth and Kleinbaum (1997) and Lall et al. (2002).

The motivating example for this paper is a longitudinal, individual animal-level study of risk factors associated with body condition score in a population of dogs. These data form part of a wider study to examine the impact of immunological and demographic factors on canine rabies vaccination coverage, which covered four locations: Braamfischerville and Zenzele in Gauteng province, South Africa; and Antiga and Kelusa in Bali province, Indonesia. Full details of the wider study, and a comprehensive analysis of all the data collected from each of the sites is provided in Morters et al. (2014).

To illustrate the requirement and performance of the methodology, we focus attention on one particular data set from Zenzele. Full details of these data are given in Section 5. The response variable is body condition score (BCS)—defined on a scale from 1–9 where a score of 1 is highly underweight, 5 is healthy, and 9 is highly overweight. As such it seems sensible to consider using cumulative link models.

There are various challenges with modelling this system, and we expand on each point in the subsequent discussion:

1. We wish to perform variable selection, in order to assess the relative impact and importance of a series of potential risk factors on BCS.
2. The data are longitudinal, and so it is necessary to account for clustering due to repeated measurements on individual animals.

3. We wish to assess the weight-of-evidence for PO versus NPO structures for each of the variables. This information is useful in helping to build up a picture (along with other indirect sources of evidence—see Morters et al., 2014) of the environmental processes driving canine demography in these regions (see Section 5 for a more detailed discussion of this point).
4. In order to tackle point 3, it is necessary to overcome some of the challenges regarding the fitting of NPO models when the stochastic ordering conditions may not hold.

In a classical statistical framework, model choice is usually performed using some form of model comparison criteria, such as Akaike’s Information Criterion (AIC; Akaike, 1974), or likelihood ratio testing. These procedures use information from a single point estimate of the parameters,  $\hat{\theta}$ , and neglect the uncertainty in  $\theta$ . In addition, inference is made conditionally on the selected model, and does not incorporate uncertainty in the choice of model. This can be important in cases where explanatory variables show consistent evidence of an effect across a range of models, but is not selected in the ‘final’ model (see e.g. Viallefont et al., 2001). Here we propose to use a Bayesian framework, and implement model selection using posterior probabilities of association (see e.g. Kass and Raftery, 1995; Viallefont et al., 2001; O’Hara and Sillanpää, 2009). This has the advantage that it allows us to assess the weight-of-evidence in favour of a given model, and also allows us to assess the evidence for a particular variable being associated with the response after averaging across all possible models. This is particularly important to this study, since we also wish to assess the conditional evidence of a PO or NPO structure for the relationship between an explanatory variable and the response, given that an association exists.

A common method to account for clustering due to repeated measurements is to use mixed effects models (see e.g. Diggle et al., 2002), in which the error term is split into different components in order to model the variance-covariance structure at different hierarchies (Hedeker and Gibbons, 1994; Gibbons and Hedeker, 1997; Hartzel et al., 2001; Hedeker, 2003; Liu and Hedeker, 2006). These techniques are well characterised in the literature, though there is some debate about how to perform model selection in the presence of random effects in a classical setting (see e.g. Vaida and Blanchard, 2005; Liang et al., 2008). In the Bayesian setting, all parameters are considered random variables, and it is straightforward to incorporate *a priori* clustering into the prior. The model choice problem then remains the same, as the parameters are simply integrated over when estimating the posterior probabilities of association.

The literature surrounding the development of ordinal regression frameworks is large and varied, applied in a wide range of fields. Focussing on Bayesian models; probit link functions are frequently used (e.g. Albert and Chib, 1993; Chu and Ghahramani, 2005; Yi et al., 2007) and there are various recent developments in modelling the link function using mixture distributions (e.g. Lang, 1999; Leon-Novelo et al., 2010). Different fitting mechanisms have also been developed, including Markov chain Monte Carlo (MCMC—Lang, 1999; Ishwaran and Gatsonis, 2000; Holmes and Held, 2006; Yi et al., 2007; Webb and Forster, 2008; Leon-Novelo et al., 2010), Laplace approximations (e.g. Chu and

Ghahramani, 2005; Paquet et al., 2005) and expectation-propagation algorithms (Chu and Ghahramani, 2005). Holmes and Held (2006) develop efficient MCMC samplers for logistic multinomial regression models, and O’Brien and Dunson (2004) develop a multivariate logistic regression framework that provides a marginal logistic structure for each of the outcomes. Some work has also been done on model selection using probit models (e.g. Albert and Chib, 1997; Chu and Ghahramani, 2005; Webb and Forster, 2008), and Mwalili et al. (2005) extend a PO model to account for interobserver measurement error. This list is by no means exhaustive, but as far as we are aware no method has been developed that accounts for all four of the challenges we highlighted earlier within the same framework. This manuscript is an attempt to provide an alternative approach for fitting logistic regression models, which allows both PO and NPO structures to be used (and provides a model-driven means of assessing which structure is most relevant for each variable in the presence of other variables), and which can be extended to deal with repeated or clustered measurements, as well as variable selection.

The first challenge we address is to provide a framework in which the stochastic ordering conditions can be made to hold for any given data set. This provides a means to explore the fitting of NPO models to any data set, and facilitates the development of a more general approach in which the relationship between the response and explanatory variables (e.g. PO or NPO) can be allowed to vary according to the data. The latter is our second contribution, and is useful because often we do not know which explanatory variables are best modelled using PO or NPO structures in advance, particularly when there are a large number of variables. To this end, Tutz and Scholz (2003) propose a method that switches between the PO, PPO and NPO models, fitting the model via a penalised likelihood approach. However, in this case we also would like to produce an estimate of the support under the data for these competing structures for each of the variables, which can provide some indirect evidence regarding the mechanisms at play in the underlying system. Although the PO model could be viewed as a special case of the NPO model, the naïve use of a straight NPO model could result in overfitting.

The challenge with comparing PO and NPO structures is that the dimensionality of the system is different in each case (a single regression parameter for the PO model corresponds to  $J - 1$  parameters for the NPO model). To deal with this issue we use reversible-jump Markov chain Monte Carlo (RJ-MCMC—Green, 1995) to fit the model, and Bayesian model averaging (BMA—e.g. Kass and Raftery, 1995) to produce posterior probabilities of association (PPAs) for the support under the prior and the data for the choice of PO or NPO structure, averaged across the set of possible models. Finally we extend these ideas to incorporate variable selection (see e.g. Dellaportas et al., 2002; O’Hara and Sillanpää, 2009). Note that an implementation of model choice based on Bayes Factors for ordinal regression models was developed in Albert and Chib (1997), though each competing model needs to be fitted separately in order to be compared (Chib, 1995). Here we integrate across the competing models in one framework, which is likely to be much more efficient when searching across a large model space. An alternative Bayesian RJ-MCMC approach to model choice for ordinal probit models is developed in Webb and Forster (2008).

In Section 2 we discuss the Bayesian paradigm and (RJ-)MCMC. In Section 3 we introduce the general cumulative link model, and more specifically the PO, NPO and

PPO models. In Section 3.2 we discuss how the Bayesian framework can be used to ensure stochastic ordering for specified variable ranges, and we justify this approach in practice in Section 3.4. The specific RJ-MCMC sampler for this ordinal model choice problem is described in Section 4. We then apply these methods to both simulated data, as well as data from a longitudinal study of individual-level risk factors affecting body condition score in a dog population in Zenzele, South Africa (Section 5). We conclude with a discussion (Section 6).

## 2 Bayesian inference and Markov chain Monte Carlo

All of the models described in this paper will be formulated in a Bayesian framework, and fitted using Markov chain Monte Carlo (MCMC). We assume that readers are familiar with the Bayesian framework, but otherwise they are referred to various excellent texts available, such as Gilks et al. (1996); Gelman et al. (2004) and Gamerman and Lopes (2006). The model fitting algorithms described in this manuscript are specifically variations of the Metropolis-Hastings (M-H) algorithm (Metropolis et al., 1953; Hastings, 1970).

Reversible-jump MCMC (Green, 1995) is an extension to the classic M-H routine that allows the Markov chain to jump between models with different dimensionality. Again, we do not discuss the full details of RJ-MCMC here, but for good introductions to the method the reader is referred to papers by Waagepetersen and Sorensen (2001) and Hastie and Green (2012).

### 2.1 Bayesian model choice using reversible-jump MCMC

Assume that we have  $V$  competing models to choose between. We can formulate the Bayesian model choice problem as one of estimating the posterior probability that a model ( $M_v$ ) is true, given the choice of competing models ( $M_1, \dots, M_V$ ). Formally, this quantity is defined as

$$P(M_v | \mathbf{D}) = \frac{f(\mathbf{D} | M_v) P(M_v)}{\sum_{u=1}^V f(\mathbf{D} | M_u) P(M_u)}, \quad (1)$$

with  $P(M_v)$  the *prior* probability of association for model  $M_v$ , and

$$f(\mathbf{D} | M_v) = \int_{\Omega_v} f(\mathbf{D} | \boldsymbol{\omega}_v, M_v) f(\boldsymbol{\omega}_v | M_v) d\boldsymbol{\omega}_v, \quad (2)$$

the *integrated likelihood*; where  $\mathbf{D}$  is the data, and  $\Omega_v$  is the (multidimensional) parameter space for the unknown parameters  $\boldsymbol{\omega}_v$  corresponding to model  $M_v$ . The quantity (1) is sometimes referred to as the *posterior probability of association* (PPA). These ideas for Bayesian model choice go back originally to Jeffreys (1935, 1961), and for a detailed introduction see Kass and Raftery (1995) and O'Hara and Sillanpää (2009).

To implement model choice in an RJ-MCMC framework, let  $v = i$  be the model indicator at a given iteration (i.e. the chain is in model  $M_i$ ), then let  $p(M_i \rightarrow M_j)$

denote the probability that a jump from  $M_i$  to  $M_j$  is proposed. In order to jump between models of differing dimensionality, the parameters  $\boldsymbol{\omega}_i$  are mapped to a set of parameters  $\boldsymbol{\omega}_j$  via the inclusion of a set of dummy parameters,  $\mathbf{u}_i$  and  $\mathbf{u}_j$ , that are chosen to ensure that  $\dim(\boldsymbol{\omega}_i, \mathbf{u}_i) = \dim(\boldsymbol{\omega}_j, \mathbf{u}_j)$ . Once these dummy parameters are chosen,  $(\boldsymbol{\omega}_i, \mathbf{u}_i)$  is mapped to  $(\boldsymbol{\omega}_j, \mathbf{u}_j)$  through a deterministic bijective function  $g_{ij}$ , such that  $g_{ij}(\boldsymbol{\omega}_i, \mathbf{u}_i) = (\boldsymbol{\omega}_j, \mathbf{u}_j)$ , and the reverse move is  $g_{ji}(\boldsymbol{\omega}_j, \mathbf{u}_j) = (\boldsymbol{\omega}_i, \mathbf{u}_i)$ . The acceptance probability of the move is then given by:

$$\alpha = \min \left[ 1, \frac{f(\mathbf{D} | \boldsymbol{\omega}_j, M_j)}{f(\mathbf{D} | \boldsymbol{\omega}_i, M_i)} \times \frac{f(\boldsymbol{\omega}_j | M_j)}{f(\boldsymbol{\omega}_i | M_j)} \times \frac{P(M_j)}{P(M_i)} \times \frac{p(M_j \rightarrow M_i) q_u(\mathbf{u}_j)}{p(M_i \rightarrow M_j) q_u(\mathbf{u}_i)} \times \left| \frac{\partial(\boldsymbol{\omega}_j, \mathbf{u}_j)}{\partial(\boldsymbol{\omega}_i, \mathbf{u}_i)} \right| \right], \quad (3)$$

where  $q_u(\mathbf{u}_i)$  is the proposal density for the dummy parameters  $\mathbf{u}_i$ , and likewise for  $q_u(\mathbf{u}_j)$ . The final quantity in (3) is the absolute value of the determinant of the Jacobian matrix.

One advantage of using RJ methodology is that for a well-mixing model, the PPA defined in (1) for a model  $M_v$  can be simply estimated as the proportion of time that the chain spends in model  $v$ . In Section 4 we show how this routine can be implemented for variable selection, as well as choosing between PO and NPO structures for individual variables. For other applications of RJ-MCMC in Bayesian model choice see e.g. Richardson and Green (1997) and Dellaportas et al. (2002).

### 3 Cumulative link models

Let  $\mathbf{Y} = (Y_1, \dots, Y_J)$  be a set of counts of individuals in  $j = 1, \dots, J$  ordered categories, which can be modelled as

$$\mathbf{Y} \sim \text{Mult}(n, \mathbf{p}), \quad (4)$$

where  $n$  is the number of individuals, and  $\mathbf{p} = (p_1, \dots, p_J)$  correspond to the probabilities of each individual being in any given category  $j$  (such that  $\sum_{j=1}^J p_j = 1$ ). If we have a set of  $K$  explanatory variables,  $\mathbf{X}_i = (X_{i1}, \dots, X_{iK})$ , associated with subset  $i$  of the  $n$  individuals (where  $i = 1, \dots, I$ , such that  $n = n_1 + \dots + n_I$ ), then

$$\mathbf{Y}_i \sim \text{Mult}(n_i, \mathbf{p}_i), \quad (5)$$

where  $\mathbf{p}_i = (p_{i1}, \dots, p_{iJ})$  and  $\sum_{j=1}^J p_{ij} = 1$ . For a fully individual-based model then  $I = n$  and  $n_i = 1$  for all  $i$ . Letting  $C_i$  correspond to the category that an individual  $i$  belongs to (such that  $C_i$  takes values  $1, \dots, J$ ), then following McCullagh (1980), we can model the cumulative probabilities,  $P(C_i \leq j) = \gamma_{ij}$  through a monotonic increasing link function  $h(\cdot)$ , mapping the interval  $(0, 1) \rightarrow (-\infty, \infty)$ , as

$$h(\gamma_{ij}) = \theta_j - \mu_i, \quad (6)$$

where  $\mu_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}$  is a linear regression term, and  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_K)$  is a vector of  $K+1$  regression parameters. In this framework the  $\theta_j$  parameters correspond

to a set of latent continuous ‘cut-points’, such that  $-\infty < \theta_1 < \dots < \theta_{J-1} < \infty$ . For identifiability we set  $\beta_0 = 0$ . The probabilities of category membership are then given by

$$p_{ij} = \begin{cases} \gamma_{ij} & \text{for } j = 1, \\ \gamma_{ij} - \gamma_{i(j-1)} & \text{for } j = 2, \dots, J-1, \\ 1 - \gamma_{i(J-1)} & \text{for } j = J. \end{cases} \quad (7)$$

In (6), the effect of the explanatory variables is independent of the grouping, and so regardless of the choice of link function the models display *strict stochastic ordering* (McCullagh, 1980). This means that subject to the constraint  $-\infty < \theta_1 < \dots < \theta_{J-1} < \infty$ , the cumulative probabilities will be such that  $0 < \gamma_{i1} < \gamma_{i2} < \dots < \gamma_{i(J-1)} < 1$ .

A more general model would allow the effect of the covariates to vary between the groups, such that

$$h(\gamma_{ij}) = \theta_j - \mu_{ij}. \quad (8)$$

In this case the models are only stochastically ordered for certain ranges of explanatory variables (see e.g. Agresti, 2010; Congdon, 2005; Tutz and Scholz, 2003). We discuss both forms of these models in detail for the case of the logistic-link function, but also extend the discussion to more general cases.

### 3.1 Proportional odds model

A common form for the link function is the *logistic* link:

$$h(\gamma_{ij}) = \log\left(\frac{\gamma_{ij}}{1 - \gamma_{ij}}\right) = \theta_j - \mu_i. \quad (9)$$

This is known as the *proportional odds* (PO) model (McCullagh, 1980), so-called because the cumulative log-odds ratio for two sets of explanatory variables,  $\mathbf{X}_1$  and  $\mathbf{X}_2$  is given by

$$\begin{aligned} \text{logit}(\gamma_{1j}) - \text{logit}(\gamma_{2j}) &= \theta_j - \beta^T \mathbf{X}_1 - \theta_j + \beta^T \mathbf{X}_2 \\ &= \beta^T (\mathbf{X}_2 - \mathbf{X}_1). \end{aligned} \quad (10)$$

Hence the cumulative log-odds ratio is proportional to the distance between  $\mathbf{X}_1$  and  $\mathbf{X}_2$  (see also Agresti, 2010).

In the case of the PO model (9), the  $\theta$  and  $\beta$  parameters are *a priori* independent, and so the joint prior distribution can be written as  $f(\beta, \theta) = f(\theta) f(\beta)$ , where we let

$$f(\beta) = \prod_{k=1}^K f(\beta_k) \quad \text{and} \quad f(\theta) = f(\theta_1) \prod_{j=2}^{J-1} f(\theta_j | \theta_{j-1}), \quad (11)$$

where  $f(\theta_1)$  is defined in  $(-\infty, \infty)$ , and  $f(\theta_j | \theta_{j-1})$  is defined in the range  $(\theta_{j-1}, \infty)$  for  $j = 2, \dots, J-1$  (see also Albert and Chib, 1993; Johnson and Albert, 1999; Congdon,

2005). This ensures stochastic ordering for any values of  $\boldsymbol{\beta}$ . We let  $\theta_1 \sim N(0, \sigma_\theta^2)$ ,  $\beta_k \sim N(0, \sigma_\beta^2)$  (for  $k = 1, \dots, K$ ) and

$$\theta_j | \theta_{j-1} \sim N(0, \sigma_\theta^2) I(T_{j-1}, \infty) \text{ for } j = 2, \dots, J-1, \quad (12)$$

where  $I(T_{j-1}, \infty)$  signifies that the distribution is truncated in the region  $(T_{j-1}, \infty)$  (i.e. it is a *lower-truncated* normal distribution) with  $T_{j-1} = \theta_{j-1}$ . Other alternative choices for the prior distributions include doubly-truncated normals (Congdon, 2005), an ordered uniform distribution (Ishwaran, 2000), or a re-parameterisation which maps the constrained variables  $\boldsymbol{\theta}$  to a set of unconstrained variables,  $\boldsymbol{\alpha}$ , which can be given, for example, a multivariate normal prior (Fahrmeier and Tutz, 1994; Albert and Chib, 1997). We choose normal random walk proposal distributions for each  $\beta_k$ , such that

$$\beta'_k | \beta_k^{(i)} \sim N(\beta_k^{(i)}, \sigma_{P\beta}^2), \quad (13)$$

where  $\sigma_{P\beta}^2$  is the proposal variance. For the cut-point parameters,  $\theta_j$ , we choose truncated uniform random-walk proposals, such that

$$\theta'_j | \boldsymbol{\theta}^{(i)} = \begin{cases} U(\theta_j^{(i)} - \tau_\theta, \min[\theta_j^{(i)} + \tau_\theta, \theta_{j+1}^{(i)}]) & \text{if } j = 1, \\ U(\max[\theta_j^{(i)} - \tau_\theta, \theta_{j-1}^{(i)}], \min[\theta_j^{(i)} + \tau_\theta, \theta_{j+1}^{(i)}]) & \text{if } j = 2, \dots, J-2, \\ U(\max[\theta_j^{(i)} - \tau_\theta, \theta_{j-1}^{(i)}], \theta_j^{(i)} + \tau_\theta) & \text{if } j = J-1, \end{cases} \quad (14)$$

where  $\tau_\theta > 0$  controls the size of the maximum unconstrained move away from the current value at each iteration.

### 3.2 Non-proportional odds model

The non-proportional odds (NPO) model is specified as

$$\log\left(\frac{\gamma_{ij}}{1 - \gamma_{ij}}\right) = \theta_j - \mu_{ij}. \quad (15)$$

In this case the regression parameters are allowed to vary with category level, such that  $\mu_{ij} = \boldsymbol{\beta}_j^T \mathbf{X}_i$  (see e.g. Agresti, 2010; Bender and Grouven, 1998; Tutz and Scholz, 2003; Congdon, 2005). The key challenge is that in order for stochastic ordering to hold, it is necessary that

$$-\infty < \theta_1 - \boldsymbol{\beta}_1 \mathbf{X} < \theta_2 - \boldsymbol{\beta}_2 \mathbf{X} < \dots < \theta_{J-1} - \boldsymbol{\beta}_{J-1} \mathbf{X} < \infty \quad (16)$$

for all  $\mathbf{X}$ . For identifiability we set each of the intercept parameters  $\beta_{0j} = 0$ . If we have  $K$  explanatory variables, then after expanding out the regression in the stochastic ordering constraints (16), for any  $j = 1, \dots, J-2$ , we have that

$$\theta_j - \theta_{j+1} < \sum_{k=1}^K (\beta_{kj} - \beta_{k(j+1)}) X_k, \quad (17)$$

which must hold for any value of  $X_k$ .

In the first instance, assume that we have a lower and upper bound for the possible values of  $X_k$ , such as would be the case if  $X_k$  were categorical. Denote the minimum and maximum values of  $X_k$  as  $X_k^m$  and  $X_k^M$  respectively. The condition (17) then becomes

$$\theta_j - \theta_{j+1} < \sum_{k=1}^K \min([\beta_{kj} - \beta_{k(j+1)}] X_k^m, [\beta_{kj} - \beta_{k(j+1)}] X_k^M). \quad (18)$$

For brevity, let

$$C_{kj} = \min[X_k^m (\beta_{kj} - \beta_{k(j+1)}), X_k^M (\beta_{kj} - \beta_{k(j+1)})] \text{ and } C_j = \sum_{k=1}^K C_{kj}. \quad (19)$$

In a similar manner to the PO model, we can therefore specify the joint prior distribution of  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$ ; however this time we do not assume both sets of parameters are independent, hence

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\theta}) &= f(\boldsymbol{\theta} | \boldsymbol{\beta}) f(\boldsymbol{\beta}) \\ &= f(\theta_1) \left[ \prod_{j=2}^{J-1} f(\theta_j | \theta_{j-1}, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1}) \right] \prod_{k=1}^K f(\beta_k), \end{aligned} \quad (20)$$

where  $f(\beta_k)$  ( $k = 1, \dots, K$ ) and  $f(\theta_1)$  are defined as before, and  $f(\theta_j | \theta_{j-1}, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$  is the probability density function for a truncated normal distribution,  $N(0, \sigma_\theta^2) I(T_{j-1}, \infty)$  with  $T_{j-1} = \theta_{j-1} - C_{j-1}$ . We choose to update each  $\beta_{kj}$  parameter in turn, conditional on all other parameters remaining fixed. It is tricky to define a simple mechanism for truncated sampling of the regression parameters, due to the fact that the conditions (16) change according to whether we propose  $\beta'_{kj} < \beta_{k(j+1)}^{(i)}$  or  $\beta'_{kj} > \beta_{k(j+1)}^{(i)}$ . Instead we opt here for a simple random-walk proposal, such that

$$\beta'_{kj} | \beta_{kj}^{(i)} = U\left(\beta_{kj}^{(i)} - \tau_\beta, \beta_{kj}^{(i)} + \tau_\beta\right), \quad (21)$$

where  $\tau_\beta > 0$  controls the size of the maximum move away from the current value at each iteration. For the cut-point parameters,  $\theta_j$ , we choose truncated uniform random-walk proposals, such that

$$\theta'_j | \boldsymbol{\theta}^{(i)} = \begin{cases} U\left(\theta_j^{(i)} - \tau_\theta, \min\left[\theta_j^{(i)} + \tau_\theta, \theta_{j+1}^{(i)} + C_j^{(i)}\right]\right) & \text{if } j = 1, \\ U\left(\max\left[\theta_j^{(i)} - \tau_\theta, \theta_{j-1}^{(i)} - C_{j-1}^{(i)}\right], \min\left[\theta_j^{(i)} + \tau_\theta, \theta_{j+1}^{(i)} + C_j^{(i)}\right]\right) & \text{if } j = 2, \dots, J-2, \\ U\left(\max\left[\theta_j^{(i)} - \tau_\theta, \theta_{j-1}^{(i)} - C_{j-1}^{(i)}\right], \theta_j^{(i)} + \tau_\theta\right) & \text{if } j = J-1, \end{cases} \quad (22)$$

where  $\tau_\theta > 0$  controls the size of the maximum unconstrained move.

### 3.3 Partial proportional odds

The *partial proportional odds* (PPO) model, proposed by Peterson and Harrell (1990), allows some variables to have a proportional odds structure and some to not. It takes the form

$$\log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) = \theta_j - \boldsymbol{\beta}^T \mathbf{X}_i - \boldsymbol{\eta}_j^T \mathbf{U}_i, \quad (23)$$

where  $i = 1, \dots, n$  and  $j = 1, \dots, J - 1$ . The regression parameters  $\boldsymbol{\beta}$  correspond to the set of explanatory variables,  $\mathbf{X}_i$ , that have a proportional odds structure, and the regression parameters  $\boldsymbol{\eta}_j$  correspond to the set of explanatory variables,  $\mathbf{U}_i$ , that have a non-proportional odds structure. The approaches described in Sections 3.1 and 3.2 can be combined in order to fit a PPO model, where the PO and NPO variables ( $\mathbf{X}$  and  $\mathbf{U}$ ) are known in advance. In subsequent sections we formulate an approach whereby the optimal choice of PO or NPO structure for each explanatory variable can instead be directly estimated by the model.

### 3.4 Justification of approach

The approach described in the previous section assumes that each of the  $X_k$  variables is bounded in some finite region, which is true for any set of categorical explanatory variables, since for a categorical variable  $Z_i$  with  $L$  levels ( $0, \dots, L - 1$ ), it is straightforward to represent  $Z_i$  as a set of  $L - 1$  dummy variables,  $X_{i1}, \dots, X_{i(L-1)}$ , such that

$$X_l = \begin{cases} 1 & \text{if } Z_i = l, \\ 0 & \text{otherwise.} \end{cases} \quad (24)$$

In the case of continuous or discrete explanatory variables that are bounded in a finite range, then the approach described in Section 3.2 will also ensure stochastic ordering holds. However, if  $X_k$  is defined over an infinite range, then these conditions will break for some values of  $X_k$  if the  $\beta_{kj}$  parameters are also unbounded.

Here we argue for a pragmatic solution to this problem, by considering that it is possible to define an upper and lower bound for  $X_k$  based on the observed data, and then use (18) to set boundary conditions for the conditional priors in (20). Although this does not mean that the stochastic ordering will hold for all possible theoretical values of  $X_k$ , it does ensure that the stochastic ordering will hold for the range of values found in the observed data. We make two main arguments to justify this approach:

1. Although theoretically the values for  $X_k$  might be infinite, for any practical applications of the model there will almost certainly be a finite range of possible values. If the observed data are a fair representation of the underlying population, then provided the model is a good fit, any population-level inference made from the model is likely to be fairly robust (i.e. the posterior distribution is the distribution of the parameters *given* the observed data, so this is explicitly represented within the Bayesian paradigm).

2. The assumptions underlying *any* statistical model can only be assessed across the range of values used to fit the data, there is no guarantee that the assumptions will hold beyond this range, even if it is possible to extrapolate without breaking any conditions of the model.

## 4 Reversible-jump algorithm for variable selection in cumulative odds ordinal regression models

Consider that we have  $K$  parameters describing the explanatory variables. In the first instance assume that each parameter measures the effect of a single variable (i.e. there are no categorical variables with  $> 2$  levels, or any interaction effects). We will extend discussion to these more complex variables in due course. We can model the relationship between the response variable  $Y$  and each variable  $X_k$  in one of three ways: either with a PO structure, an NPO structure or no relationship at all; in this example giving  $3^K$  possible models. Here we will assume that we have no prior information to distinguish between which of these models is most likely, and so assume equal prior probabilities of association for each competing model. To model these structures we introduce an indicator variable  $S_k$ , for  $k = 1, \dots, K$ , where

$$S_k = \begin{cases} 0 & \text{if } X_k \text{ has a PO structure,} \\ 1 & \text{if } X_k \text{ has an NPO structure,} \\ 2 & \text{if } X_k \text{ is excluded.} \end{cases} \quad (25)$$

To ease programming, it is helpful to treat each of these three possibilities as special cases of the NPO-structure, such that if a variable has a PO-structure then this is equivalent to setting  $\beta_{kj} = \beta_k$  for  $j = 1, \dots, J - 1$ , with independent point-mass priors on  $\beta_{k2}, \dots, \beta_{k(J-1)}$  such that  $f(\beta_{kj} = 0) = 1$ . If a variable is excluded then this is equivalent to setting  $\beta_{kj} = 0$  with a point-mass prior  $f(\beta_{kj} = 0) = 1$ . This enables us to use the conditions in (16) to ensure general stochastic ordering.

### 4.1 Adding or removing variables

Our stochastic search routine updates each variable  $X_k$  in a random order, by proposing to add the variable (if currently excluded) or to remove the variable (if currently included) with a probability  $p_{\text{jump}}$  (hence we do nothing with a probability  $1 - p_{\text{jump}}$ ).

To add a variable into the model, we sample whether to use a PO or NPO structure with probability  $p_{\text{PO}}$  and  $p_{\text{NPO}} = 1 - p_{\text{PO}}$  respectively. To add a variable to the model with a PO structure, we define a bijective function

$$g_{0 \rightarrow \text{PO}}(u_1) = u_1 = \beta_k, \quad (26)$$

where  $u_1$  is sampled from some distribution with p.d.f.  $q_u(\cdot)$ . To add a variable with an NPO structure, we define a bijective function

$$g_{0 \rightarrow \text{NPO}}(u_1, \dots, u_{J-1}) = (u_1, \dots, u_{J-1}) = (\beta_{k1}, \dots, \beta_{k(J-1)}), \quad (27)$$

where  $u_1, \dots, u_{J-1}$  are independent and identically distributed (i.i.d.) samples from a distribution following  $q_u(\cdot)$ . For a  $0 \rightarrow \text{PO}$  move, the acceptance probability is

$$\alpha = \min \left[ 1, \frac{f(\mathbf{Y} | \boldsymbol{\beta}', \boldsymbol{\theta}^{(i)})}{f(\mathbf{Y} | \boldsymbol{\beta}^{(i)}, \boldsymbol{\theta}^{(i)})} \times \frac{f(\beta'_k)}{1} \times \frac{1}{q_u(u_1)} \times \frac{1}{p_{\text{PO}}} \right]. \quad (28)$$

The probability of adding or dropping a variable,  $p_{\text{jump}}$ , is the same for the forwards and reverse moves, and so cancel in the acceptance ratio. The determinant of the Jacobian matrix is 1. For a  $0 \rightarrow \text{NPO}$  move, the acceptance probability is

$$\begin{aligned} \alpha = \min & \left[ 1, \frac{f(\mathbf{Y} | \boldsymbol{\beta}', \boldsymbol{\theta}^{(i)})}{f(\mathbf{Y} | \boldsymbol{\beta}^{(i)}, \boldsymbol{\theta}^{(i)})} \times \frac{[\prod_{j=1}^{J-1} f(\beta'_{kj})]}{[\prod_{j=2}^{J-1} f(\theta_j^{(i)} | \theta_{j-1}^{(i)}, \boldsymbol{\beta}'_j, \boldsymbol{\beta}'_{j-1})]} \right. \\ & \left. \times \frac{1}{q_u(u_1, \dots, u_{J-1})} \times \frac{1}{1 - p_{\text{PO}}} \right]. \quad (29) \end{aligned}$$

We let  $u_j$  be i.i.d. random variables such that  $u_j \sim N(0, \sigma_{P\beta}^2)$ , where  $\sigma_{P\beta}^2$  is the proposal variance. To remove a variable that is currently included we can simply reverse this process, amending the acceptance probabilities accordingly.

## 4.2 Updating included variables

The second stage of our MCMC routine involves updating the values for any parameters that are currently included in the model. In a random order, we select each of the  $K$  variables in turn, and with probability  $p_{\text{move}}$  we propose new values for the associated parameter(s), and with probability  $1 - p_{\text{move}}$  we propose a shift from  $\text{PO} \rightarrow \text{NPO}$  (if variable  $k$  has a  $\text{PO}$  structure), or  $\text{NPO} \rightarrow \text{PO}$  (if variable  $k$  has an  $\text{NPO}$  structure).

If variable  $k$  has a  $\text{PO}$  structure, then to update the value of  $\beta_k$  we simply propose a new value from some proposal distribution with p.d.f.  $q_\beta(\beta'_k)$ . The update is then a standard Metropolis-Hastings step. Likewise for  $\beta_{kj}$  ( $j = 1, \dots, J-1$ ) if variable  $k$  has an  $\text{NPO}$  structure.

To switch structures we require a reversible-jump step. To make an  $\text{NPO} \rightarrow \text{PO}$  move—i.e.  $\text{map}(\beta_{k1}, \dots, \beta_{k(J-1)}) \rightarrow \beta_k$ —we define a bijective function

$$\begin{aligned} g_{\text{NPO} \rightarrow \text{PO}}(\beta_{k1}, \dots, \beta_{k(J-1)}) &= (\bar{\beta}_k, \bar{\beta}_k - 2\beta_{k2}, \dots, \bar{\beta}_k - 2\beta_{k(J-1)}) \\ &= (\beta_k, u_1, \dots, u_{J-2}), \end{aligned} \quad (30)$$

where  $\bar{\beta}_k = (J-1)^{-1} \sum_{j=1}^{J-1} \beta_{kj}$ . To make the reverse move we do not have to propose any new values, and simply use the inverse function

$$\begin{aligned} &g_{\text{PO} \rightarrow \text{NPO}}(\beta_k, u_1, \dots, u_{J-2}) \\ &= \frac{J-1}{2} \left( \beta_k(4-J) + \sum_{j=1}^{J-2} u_j, (\beta_k - u_1), \dots, (\beta_k - u_{J-2}) \right) \\ &= (\beta_{k1}, \dots, \beta_{k(J-1)}). \end{aligned} \quad (31)$$

These choices are based around the moment matching approach of Brooks et al. (2003). The acceptance probability for a PO  $\rightarrow$  NPO move is:

$$\begin{aligned} \alpha = \min & \left[ 1, \frac{f(\mathbf{Y} \mid \boldsymbol{\beta}', \boldsymbol{\theta}^{(i)})}{f(\mathbf{Y} \mid \boldsymbol{\beta}^{(i)}, \boldsymbol{\theta}^{(i)})} \times \frac{\prod_{j=1}^{J-1} f(\beta'_{kj})}{f(\beta_k^{(i)})} \right. \\ & \times \frac{\prod_{j=2}^{J-1} f(\theta_j^{(i)} \mid \theta_{j-1}^{(i)}, \boldsymbol{\beta}'_j, \boldsymbol{\beta}'_{j-1})}{\prod_{j=2}^{J-1} f(\theta_j^{(i)} \mid \theta_{j-1}^{(i)}, \boldsymbol{\beta}_j^{(i)}, \boldsymbol{\beta}_{j-1}^{(i)})} \times \frac{1}{q_u(u_1, \dots, u_{J-2})} \\ & \left. \times (J-1) \left( \frac{J-1}{2} \right)^{J-2} \right], \end{aligned} \quad (32)$$

where the final term is the absolute value for the determinant of the Jacobian. Similarly, the acceptance probability for an NPO  $\rightarrow$  PO move is

$$\begin{aligned} \alpha = \min & \left[ 1, \frac{f(\mathbf{Y} \mid \boldsymbol{\beta}', \boldsymbol{\theta}^{(i)})}{f(\mathbf{Y} \mid \boldsymbol{\beta}^{(i)}, \boldsymbol{\theta}^{(i)})} \times \frac{f(\beta'_k)}{\prod_{j=1}^{J-1} f(\beta_{kj}^{(i)})} \right. \\ & \times \frac{\prod_{j=2}^{J-1} f(\theta_j^{(i)} \mid \theta_{j-1}^{(i)}, \boldsymbol{\beta}'_j, \boldsymbol{\beta}'_{j-1})}{\prod_{j=2}^{J-1} f(\theta_j^{(i)} \mid \theta_{j-1}^{(i)}, \boldsymbol{\beta}_j^{(i)}, \boldsymbol{\beta}_{j-1}^{(i)})} \times \frac{q_u(u_1, \dots, u_{J-2})}{1} \\ & \left. \times \left( \frac{1}{J-1} \right) \left( \frac{2}{J-1} \right)^{J-2} \right]. \end{aligned} \quad (33)$$

We then proceed to update the cut-points,  $\boldsymbol{\theta}^{(i)}$ , in the same way as described in Section 3.2.

We note that this general RJ-MCMC algorithm can be adapted in various ways simply by altering the move probabilities. For example, we can remove the variable selection steps and just allow the model to move between the PO and NPO structures for each variable by setting  $p_{\text{jump}} = 0$ . Similarly, we can also fix all parameters to have either a PO or NPO structure, both with or without variable selection, by adjusting  $p_{\text{jump}}$  and  $p_{\text{move}}$  accordingly.

### 4.3 Tuning

In some cases there may be some identifiability issues between regression parameters and their corresponding inclusion indicators when implementing variable selection routines using the framework described above. For example, there may be almost identical likelihoods when a parameter is removed (set to zero) and when a parameter is present but has a value close to zero (e.g. O'Hara and Sillanpää, 2009). If vague priors are used for the regression parameters and inclusion indicators, then these parameters may be unidentifiable. One way to control this is to use a more informative prior, such as one guided by the data or training runs of the model. However, as O'Hara and Sillanpää

(2009) note, there is a danger that these approaches will contravene the philosophical construct that the prior distribution should represent one's beliefs about the parameters *before* obtaining any data.

A potential way to tackle this problem in this case is to introduce a hyperprior governing the variance component of the priors for the regression parameters,  $\beta$ . This could be done in various ways, but for PO structures we set

$$\beta_k \sim N(0, \sigma_{k\beta}^2) \text{ where } \sigma_{k\beta} \sim U(0, \xi) \quad (34)$$

and  $\xi$  is the maximum *a priori* range for  $\sigma_{k\beta}$  (O'Hara and Sillanpää, 2009), and for NPO structures we set

$$\beta_{kj} \sim N(0, \sigma_{kj\beta}^2) \text{ where } \sigma_{kj\beta} \sim U(0, \xi). \quad (35)$$

This adds a further complexity to the model since it introduces additional parameters to sample during the dimension-jumping steps. For example, a  $0 \rightarrow \text{PO}$  move would now consist of moving from  $0 \rightarrow (\beta_k, \sigma_{k\beta})$ , likewise a  $\text{PO} \rightarrow \text{NPO}$  jump would consist of moving from  $(\beta_k, \sigma_{k\beta}) \rightarrow (\beta_{k1}, \sigma_{k1\beta}, \dots, \beta_{k(J-1)}, \sigma_{k(J-1)\beta})$  and so on. To do this we update each  $\beta$  parameter and its corresponding  $\sigma_\beta$  parameter at the same time, using independent proposal distributions.

A slight complexity is that the standard deviations must be positive. Hence for a  $0 \rightarrow \text{PO}$  or  $0 \rightarrow \text{NPO}$  move (or the reverse moves), we use the same bijective functions as are described in Section 4.1, except that the dummy variables for the standard deviations are i.i.d. samples from a  $U(0, \xi)$  distribution. The acceptance probabilities are adjusted accordingly. For a  $\text{PO} \rightarrow \text{NPO}$ , we use a slightly different bijective function for proposing the standard deviations than for the regression parameters. Here we propose values for  $u_1, \dots, u_{J-2}$  as i.i.d.  $U(\max[0, \sigma_k - \tau_\sigma], \min[\sigma_k + \tau_\sigma, \xi])$  variables, and then define

$$\begin{aligned} g_{\text{PO} \rightarrow \text{NPO}}(\sigma_k, u_1, \dots, u_{J-2}) &= (\sigma_k, \sigma_k + u_1, \dots, \sigma_k + u_{J-2}) \\ &= (\sigma_{k1}, \dots, \sigma_{k(J-1)}). \end{aligned} \quad (36)$$

To make the reverse move we do not have to propose any new values, and simply use the inverse function

$$\begin{aligned} g_{\text{NPO} \rightarrow \text{PO}}(\sigma_{k1}, \dots, \sigma_{k(J-1)}) &= (\sigma_{k1}, \sigma_{k2} - \sigma_{k1}, \dots, \sigma_{k(J-1)} - \sigma_{k1}) \\ &= (\sigma_k, u_1, \dots, u_{J-2}). \end{aligned} \quad (37)$$

The acceptance probabilities are updated accordingly, but the additional proposals of the standard deviation terms do not change the Jacobian terms in (32) or (33).

#### 4.4 Including categorical explanatory variables with $> 2$ levels, and interaction effects

When comparing nested models including interaction effects, it is usual to specify that interactions can only be included as long as the corresponding main effect terms are

also included, and that higher-order interaction terms are included only if all lower-order terms are included (Krzanowski, 1998). These constraints can be incorporated into the routines described in Section 4 by altering the move probabilities. For example, consider the possible moves for a main effect variable,  $X_k$ , currently included in the model (with a PO structure). If there were no interaction effects, then we propose to exclude the variable with probability  $p_{\text{jump}}$ . If we are modelling interaction effects, then we would instead propose to exclude the variable with probability  $p'_{\text{jump}}$ , where

$$p'_{\text{jump}} = \begin{cases} 0 & \text{if any interaction effect relating to } X_k \text{ is present,} \\ p_{\text{jump}} & \text{otherwise.} \end{cases} \quad (38)$$

Likewise, to add interaction effects we need to check that all associated main effects and lower-order interaction effects are present first. This ensures that we only drop or add variables in the correct manner.

Explanatory variables with  $> 2$  categories require more than one dummy variable to model (see Section 3.4). In this case, when proposing to add or remove a variable of this form, we must ensure that all associated dummy variables are added or removed simultaneously. We propose to add a variable of this nature with probability  $p_{\text{jump}}$ , and then for each associated dummy variable  $X_k$ , we independently propose whether these will have PO or NPO structures on addition, with probabilities  $p_{\text{PO}}$  or  $p_{\text{NPO}}$  respectively. The acceptance probabilities are amended accordingly, with the Jacobian term being just a product of the corresponding Jacobian terms for each of the dummy variables. The reverse process proceeds in a similar manner.

## 5 Applications

All the following routines were coded in C and R (R Core Team, 2012) and are available in an R package called `BayesOrd`, which in turn uses the `coda` (Plummer et al., 2006) and `multicore` (Urbanek, 2011) packages to produce output and run multiple chains in parallel. All results are reported to 2 significant figures (s.f.). The development version of this package is available at <https://github.com/tjmckinley/BayesOrd>. Following Link and Eaton (2012), we do not thin our MCMC chains once the burn-in has been discarded.

### 5.1 Simulation study

To test the performance of our algorithms, we simulated different data sets assuming

- (a) each variable has a PO structure;
- (b) each variable has an NPO structure; and
- (c) a mixture of PO and NPO variables are used.

For each scenario we simulated  $n_{\text{sim}} = 100$  data sets, each containing  $n = 1000$  samples. Each sample corresponds to measurements on the response variable ( $Y$ ) and 7 explanatory variables (5 binary,  $X_1, \dots, X_5$ , and 2 discrete,  $X_6$  and  $X_7$ ). The response is an ordinal variable with three levels.

Each simulation proceeds as follows:

1. In scenario (a), set each explanatory variable,  $X_k$ , to have a PO structure. In scenario (b) set each structure to NPO, and in scenario (c) sample the structure for each  $X_k$  from a Bernoulli distribution with probability 0.5.
2. For each categorical variable  $X_{ik}$  ( $i = 1, \dots, n; k = 1, \dots, 5$ ), sample its value (0 or 1) from a Bernoulli distribution with probability 0.5. (Ensure that there are at least 5% of samples in each group by resampling if required.)
3. For each discrete  $X_{ik}$  ( $i = 1, \dots, n; k = 6, 7$ ), sample data points as  $X_{ik} = |X'_{ik}|$ , where  $X'_{ik} \sim N(0, \sigma_k^2)$ . Here,  $\sigma_k = |\sigma'_k|$  and  $\sigma'_k \sim N(0, 5^2)$ .
4. Sample the regression parameters  $\beta_{kj} \sim N(0, 5^2)$ , where  $j = 1, 2$  corresponds to the length of the response. For each  $k$  corresponding to a PO structure, set  $\beta_{kj} = \beta_{k1} \forall j$ .
5. Sample the first threshold parameter,  $\theta_1 \sim N(-1, 0.1^2)$ , and then simulate the second threshold parameter,  $\theta_2$  conditional on  $\theta_1$ , the simulated data  $\mathbf{X}$  and the regression parameters  $\boldsymbol{\beta}$ , ensuring that the stochastic ordering conditions (16) hold. To do this we can use (18) to define a lower bound for  $\theta_2$ , and then add some positive random noise (we chose the absolute value from a  $N(0, 0.1^2)$  distribution). (Note that in the case of scenario (a) we only need to simulate such that  $\theta_1 < \theta_2$ , since the stochastic ordering conditions always hold.)
6. Finally, sample values of the response variable,  $Y_i$ , from a multinomial distribution with probability vector defined using (23). (Ensure that there are at least 5% of the samples in each category of the response, else re-simulate.)

Once the data sets were simulated, we proceeded to fit PO and NPO models in both Bayesian and maximum likelihood (ML) frameworks. The Bayesian models were fitted using the routines developed in this manuscript and implemented in the `BayesOrd` package. The maximum likelihood PO models were fitted using the `polr` function in the `MASS` (Venables and Ripley, 2002) package in R, and the ML NPO models were fitted using binary logistic regressions, as described in e.g. Bender and Grouven (1998). We also fitted a Bayesian PPO model, using the reversible-jump routines described earlier to choose between the competing structures for each variable. For the MCMC routines, we used 200,000 updates, with the first 10,000 discarded as burn-in.

To summarise the results we examine the distributions for the squared error between the true value of the regression parameters and the ML estimate or posterior mean accordingly. Table 1 summarises these results. Focussing first on the results from scenario (a), we can see that as expected, the PO models perform well, with the ML and Bayesian estimates showing a similar degree-of-accuracy. The NPO and PPO models also perform well, suggesting that although they are overparameterised, given enough data they can produce robust inference on the parameters.

For data sets simulated using scenario (b), the PO models now fit poorly, but the NPO and PPO models once again perform well. Similar patterns are observed for the

simulations based on scenario (c), and once again the Bayesian NPO and PPO models perform very well in comparison to the other approaches. There are occasional poor estimates of the parameters used in the simulations (seen by the high 97.5% credible intervals in Table 1). These could be caused either by a lack-of-fit (most likely when these values are very high), or more frequently when the data are a sample from the extremes of the expected sampling distribution. In any case, the Bayesian methods seem more robust to these outliers, particularly compared to the extreme ML NPO mismatches. We postulate that this is likely due to the fact that the Bayesian methods contain all information in the likelihood, as opposed to the ML NPO method which must treat groups independently.

As a simple exploration of the utility of the Bayesian PPO model for discriminating between PO and NPO structures, we apply a threshold such that any variable with  $PPA_{PO} > 0.5$  is classified as having a PO structure. In this case we have  $7 \times 100$  possible predictions for each simulation scenario. In the case of scenario (a), only  $3/700 = 0.43\%$  are incorrectly specified as having an NPO structure. In the case of scenario (b),  $229/700 = 33\%$  are misclassified as having a PO structure. For scenario (c),  $124/700 = 18\%$  are misclassified, of which  $95/700 = 14\%$  are NPO variables misclassified as PO variables, and  $29/700 = 4\%$  are PO variables misclassified as NPO variables. This shows a good predictive power, bearing in mind that the Bayesian model choice framework intrinsically favours more parsimonious models, and as such the majority of misclassifications were NPO variables being reduced to PO variables, such as we might expect if the differences between the regression parameters for each level of the response are small. Of course these ‘misclassifications’ may be directly due to quirks in the data as a result of random sampling, and to this end Table 1 suggests that the Bayesian NPO and PPO estimates are robust compared to other methods, even accounting for any misclassification in the actual structure used for the simulations. We reiterate that these model fits were performed blind, without a prerequisite descriptive analysis that might shed some light on our *a priori* expectations of variable structures. In practice we would take more care with our preliminary model exploration and our model diagnostics, but with this in mind we think the methods perform well.

## 5.2 Longitudinal study of individual-level risk factors affecting body condition score in a dog population in Zenzele, South Africa

These data form part of a wider study to examine the impact of immunological and demographic factors on canine rabies vaccination coverage. This study was conducted in four locations: Braamfischerville and Zenzele in Gauteng province, South Africa; and Antiga and Kelusa in Bali province, Indonesia. Full details of the study, and a comprehensive analysis of all the data collected from each of the sites is provided in Morters et al. (2014).

To illustrate the methodology, we focus attention on one particular data set from Zenzele, exploring individual-level risk factors associated with body condition score in a population of dogs. The data set consists of 2746 entries, for 738 dogs, with each dog examined between 1 and 17 times across the period 3rd March 2008–8th April 2011.

| Scenario | Algorithm      | 2.5% | Median | 97.5% |
|----------|----------------|------|--------|-------|
| (a) PO   | ML (PO)        | 0.00 | 0.06   | 3.5   |
|          | Bayesian (PO)  | 0.00 | 0.06   | 1.8   |
|          | ML (NPO)       | 0.00 | 0.10   | 7.8   |
|          | Bayesian (NPO) | 0.00 | 0.07   | 3.1   |
|          | Bayesian (PPO) | 0.00 | 0.06   | 1.9   |
| (b) NPO  | ML (PO)        | 0.02 | 5.2    | 85    |
|          | Bayesian (PO)  | 0.02 | 6.3    | 86    |
|          | ML (NPO)       | 0.00 | 0.32   | 1821  |
|          | Bayesian (NPO) | 0.00 | 0.37   | 34    |
|          | Bayesian (PPO) | 0.00 | 0.48   | 37    |
| (c) PPO  | ML (PO)        | 0.00 | 2.5    | 61    |
|          | Bayesian (PO)  | 0.00 | 3.4    | 66    |
|          | ML (NPO)       | 0.00 | 0.24   | 321   |
|          | Bayesian (NPO) | 0.00 | 0.18   | 18    |
|          | Bayesian (PPO) | 0.00 | 0.16   | 20    |

Table 1: Summaries of squared error between estimated and true values, for data sets generated using three different scenarios (defined in the main text). Within each panel,  $n_{\text{sim}} = 100$  simulated data sets are generated, each of size  $n = 1000$  samples. Each panel is further stratified by the type of model (PO, NPO, PPO) and the fitting mechanism (ML or Bayesian).

Body condition score (BCS) was assessed using a nine-point scoring system (German and Holden, 2006), with each dog being scored by two assessors simultaneously. The system assigns a score of 1–9, with 1 being very underweight, 5 being normal, and 9 being obese. To maintain a reasonable sample size in each group, we amalgamated the extreme scores, resulting in 5 BCS groups: 1–2, 3, 4, 5 and 6–9. Eight explanatory variables were collected: gender (male/female), OPL (oestrus-pregnancy-lactation; coded as normal/lactating/pregnant), number of dogs in the sample unit (discrete between 1–9), age (0–6 months, 7–12 months, 13–36 months and >36 months), sterilisation (true/false), confinement (true/false), owner reported clinical signs in the previous 7 days (none/minor/major-short duration/major-medium duration/major-long duration) and clinical signs observed by enumerator during interview (none/minor/major). In the interests of comparison, we fitted two separate models, the first assuming that the *maximum* BCS between the two assessors was correct, and the second assuming the *minimum* was correct. Summaries of the data are provided in Table 2, and distributions by BCS are shown in Figure 1.

To account for the repeated measurements, an individual dog-level term,  $\psi_{D_i}$ , was introduced, with prior distribution

$$\psi_{D_i} \sim N(0, \sigma_\psi^2), \quad (39)$$

where  $D_i$  denotes the specific dog corresponding to observation  $i$  ( $D_i = 1, \dots, 738$ ), and  $\sigma_\psi^2$  has a vague gamma hyperprior with shape and rate parameters 0.01 and 0.01 respectively (i.e. mean=1 and variance=100). At each iteration 30% of the  $\psi_{D_i}$  terms

| Variable   | Level       | Count / summary  |
|--|-------------|--|
| BCS  | 1–2         | 123  |
|  | 3           | 462  |
|  | 4           | 927  |
|  | 5           | 858  |
|  | 6–9         | 376  |
| Gender   | Female      | 1468   |
|  | Male        | 1278   |
| OPL  | Normal      | 2483   |
|  | Lactating   | 160  |
|  | Pregnant    | 103  |
| # dogs in SU                                     |             | <i>Min:</i> 1<br><i>Lower quartile:</i> 1<br><i>Median:</i> 1<br><i>Mean:</i> 1.8<br><i>Upper quartile:</i> 2<br><i>Max:</i> 9 |
| Age  | 0–6m        | 294  |
|  | 7–12m       | 452  |
|  | 13–36m      | 587  |
|  | > 36m       | 1413   |
| Sterilisation                                    | No          | 2679   |
|  | Yes         | 67   |
| Confinement                                      | No          | 1955   |
|  | Yes         | 791  |
| Owner reported clinical signs in previous 7 days | None        | 2300   |
|  | Minor       | 165  |
|  | Major/short | 65   |
|  | Major/med.  | 135  |
| Clinical signs observed by enumerator            | Major/long  | 81   |
|  | None        | 1983   |
|  | Minor       | 497  |
|  | Major       | 266  |

Table 2: Marginal summaries of the data (assuming *maximum* BCS). The final column contains counts unless otherwise stated. For comparison, the BCS counts when choosing the *minimum* BCS are 191, 652, 1082, 617 and 204 respectively.

were updated in turn at random, using a uniform random walk proposal with the maximum proposal jump given by  $\tau_\psi$ . Likewise the variance  $\sigma_\psi^2$  was also updated in the same manner with the maximum proposal jump given by  $\tau_{\sigma_\psi}$ .

To complete the Bayesian specification we set the prior variance for the cut-points,  $\sigma_\theta^2 = 1$ , and the maximum *a priori* value for the standard deviations of the regression parameter priors,  $\xi = 20$  (following O’Hara and Sillanpää, 2009). The proposal param-

eters were:  $\tau_\beta = 1$ ,  $\tau_\theta = 1$ ,  $\sigma_{P\beta}^2 = 1$ ,  $\tau_\sigma = 1$ ,  $\tau_\psi = 1$  and  $\tau_{\sigma_\psi} = 1$ . Two chains were run, and after a short training run of 1,000 iterations, from which initial values for the main chains were generated, we ran 500,000 iterations with the first 50,000 discarded as burn-in. To produce the fitted plots (Figure 1) we took 2,000 samples from the posterior. Full trace and density plots are given in Supplementary Materials.

The PPAs for each variable, averaged across the competing models are shown in Table 3. For clarity, values  $< 1 \times 10^{-1}$  are rounded to zero. More precise results are shown in Supplementary Table S1. Using conventional rules-of-thumb for interpreting these values (see e.g. Viallefont et al., 2001), if a variable has a PPA of inclusion of  $< 0.5$ , then we consider that there is negligible evidence to support this variable being associated with the response. PPAs of inclusion of 0.5–0.75 are considered weak evidence, 0.75–0.95 positive evidence, 0.95–0.99 strong evidence and  $> 0.99$  very strong evidence.

| Variable   | Level       | Max. BCS |      |      | Min. BCS |      |      |
|--|-------------|----------|------|------|----------|------|------|
|  |             | PO       | NPO  | Exc. | PO       | NPO  | Exc. |
| Gender   | F           |          |      |      |          |      |      |
|  | M           | 0.52     | 0    | 0.48 | 0.72     | 0    | 0.28 |
| OPL  | Norm.       |          |      |      |          |      |      |
|  | Lac.        | 1        | 0    | 0    | 1        | 0    | 0    |
|  | Preg.       | 1        | 0    | 0    | 1        | 0    | 0    |
| # dogs in SU                                     |             | 0.025    | 0    | 0.98 | 0.038    | 0    | 0.96 |
| Age  | 0–6m        |          |      |      |          |      |      |
|  | 7–12m       | 1        | 0    | 0    | 1        | 0    | 0    |
|  | 13–36m      | 1        | 0    | 0    | 1        | 0    | 0    |
|  | >36m        | 0        | 1    | 0    | 0.82     | 0.18 | 0    |
| Sterilised                                       | N           |          |      |      |          |      |      |
|  | Y           | 0.5      | 0    | 0.5  | 0.43     | 0    | 0.57 |
| Confined   | N           |          |      |      |          |      |      |
|  | Y           | 0.98     | 0.02 | 0    | 1        | 0    | 0    |
| Owner reported clinical signs in previous 7 days | None        |          |      |      |          |      |      |
|  | Minor       | 0        | 0    | 1    | 0        | 0    | 1    |
|  | Major/short | 0        | 0    | 1    | 0        | 0    | 1    |
|  | Major/med.  | 0        | 0    | 1    | 0        | 0    | 1    |
|  | Major/long  | 0        | 0    | 1    | 0        | 0    | 1    |
| Clinical signs observed by enumerator            | None        |          |      |      |          |      |      |
|  | Minor       | 1        | 0    | 0    | 1        | 0    | 0    |
|  | Major       | 1        | 0    | 0    | 1        | 0    | 0    |

Table 3: Posterior probabilities of association for different variables, averaged across all models.

In this case we can see that there is consistency in the variables identified as being important from both analyses (i.e. using the maximum and minimum BCS scores as the response). In this case we identify gender as showing weak evidence of an association; and OPL, age, confinement and enumerator observed clinical signs as showing very strong evidence of an association.

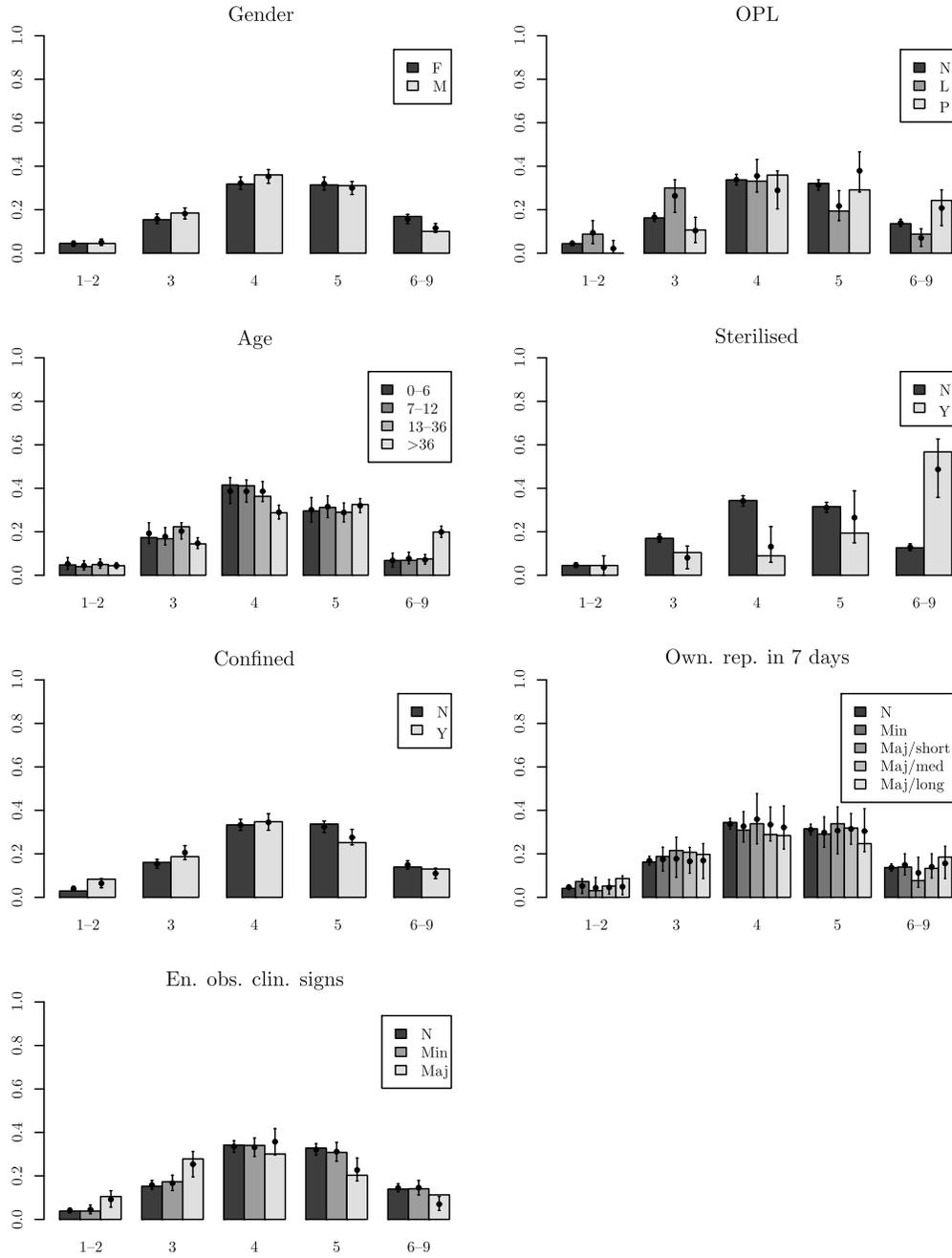


Figure 1: Marginal posterior predictive distributions for the explanatory variables, against the observed data. Bars represent the data, the points are the marginal predictive means and the error bars are the 95% prediction intervals.

For those variables with PPAs  $> 0.5$ , we can see that in almost all of these cases the PO structure is preferred, which is reflected in the model averaged log cumulative odds ratios shown in Table 4. For clarity, where the posterior means and SDs are the same across the levels (to 2 s.f.—i.e. the variable has an effective PO structure), we show only a single result. For all intents and purposes the only variable that shows any possible non-negligible support for an NPO structure is the  $>36$  month age class. Posterior predictive distributions for the observed values can be obtained, and the marginal means and 95% prediction intervals for the categorical explanatory variables are shown in Figure 1.

The overall patterns using both the minimum BCS and maximum BCS as response are the same, and so in the following discussion we will focus on the estimates obtained from using the maximum BCS only. When assessing the posteriors, it is possible to produce conditional inference based on a given model, or produce a posterior based on a weighted mixture of the posteriors from each of the models being averaged over (see e.g. Kass and Raftery, 1995; Viallefont et al., 2001; O’Hara and Sillanpää, 2009). In the case of variables that have a non-zero posterior probability of exclusion, the latter approach will shrink these estimates towards zero.

With this in mind, males are on average 1.2 times more likely to have a *lower* BCS than females. However, lactating females are, on average, 3.0 times more likely to have a *lower* BCS than equivalent males and non-pregnant females. Pregnant females on the other hand, are 1.6 times more likely to be have a *higher* BCS. For this variable, there is mixed support for the inclusion of gender with a PO structure, and exclusion altogether. Therefore, in this case these posterior estimates will have been shrunk towards zero relative to the conditional posterior given inclusion. This effect will be minimal for the other variables discussed below which each have a high probability of inclusion.

The effect of age is interesting; relative to the 0–6 month category, dogs aged between 7–12 months are 1.1 times more likely to have a *higher* BCS; dogs aged between 13–36 months are 1.2 times more likely to have a *lower* BCS, and as adults ( $>36$  months) they are between 1–4.5 times more likely to have a *higher* BCS, depending on the category level (since the adult age class has very strong evidence of an NPO structure). A likely explanation is that this pattern reflects normal morphological variation—generally, as dogs become older their activity levels will decrease, resulting in a general increase in BCS.

An interesting finding in this analysis is that other than the  $> 36$  month age category, all other variables had very strong support for a PO structure (conditional on inclusion). One of the key motivations for the study that generated these data was to examine the hypothesis that these canine populations are regulated by environmental resource constraints (as they would be in wild populations). If this hypothesis is true, then consistent with empirical evidence in other species and ecological theory, the thin dogs should generally be the ones with highest energy requirements (particularly lactating and growing dogs). The marginal distributions shown in Figure 1 provide qualitative evidence against this hypothesis, since whilst on average there was a tendency for lactating dogs to be thinner than non-lactating dogs, overall the body condition distribution for lactating dogs shows that most dogs are in reasonable body condition,

| Variable<br>(baseline level)                          | Level of<br>variable | Level of<br>response | Maximum BCS |      | Minimum BCS |      |
|---|----------------------|----------------------|-------------|------|-------------|------|
|   |                      |                      | Mean        | SD   | Mean        | SD   |
| Gender (F)  | M                    | 1                    | -0.18       | 0.20 | -0.28       | 0.21 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
| OPL (normal)  | Lac.                 | 1                    | -1.1        | 0.18 | -1.3        | 0.18 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
|   | Preg.                | 1                    | 0.46        | 0.25 | 0.47        | 0.22 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
| Age (0–6m)  | 7–12m                | 1                    | 0.12        | 0.15 | 0.35        | 0.16 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
|   | 13–36m               | 1                    | -0.19       | 0.15 | 0.09        | 0.15 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
|   | >36m                 | 1                    | -0.0065     | 0.22 | 0.64        | 0.28 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
| Confined (N)  | Y                    | 1                    | -0.56       | 0.14 | -0.58       | 0.12 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
| Clinical signs<br>observed by<br>enumerator<br>(None) | Minor                | 1                    | -0.14       | 0.11 | -0.14       | 0.11 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |
|   | Major                | 1                    | -1          | 0.15 | -1.3        | 0.15 |
|   |                      | 2                    |             |      |             |      |
|   |                      | 3                    |             |      |             |      |
|   |                      | 4                    |             |      |             |      |

Table 4: Model averaged posterior means and standard deviations for the log cumulative odds ratios. Only those variables with non-negligible association to response (i.e. a PPA of inclusion of  $>0.5$ ) are shown. For clarity, those variables that have the same means and SDs for each level of the response (to 2 s.f.) are shown as a single entry.

with fewer dogs in the extremes. Crucially there are underweight lactating dogs and underweight non-lactating dogs, and there are overweight lactating dogs and overweight non-lactating dogs—consistent with variable food availability most likely from an owner, rather than from the environment (e.g. scavenging). The same is true for young dogs.

A similar argument could be made by examining the evidence for PO versus NPO structures for these key variables (particularly OPL and age). Under the hypothesis of environmental constraints limiting population size, then we might expect lactating and young dogs to be more likely to exhibit an NPO structure, with decreasing negative log-odds ratios with increasing BCS. We do not observe this here. There is strong evidence of an NPO structure for the  $> 36$  month age class, though this is again consistent with the population being ‘managed’, rather than acting as a wild population.

Similar results are obtained for all four study regions. This information has important implications for designing optimal vaccination strategies against rabies in these populations. For full details of the study, and a comprehensive discussion about all the collected evidence, see Morters et al. (2014).

Confinement is associated with a *lower* BCS, with confined dogs being 1.7 times more likely to have a lower BCS than unconfined dogs. Although confinement, as defined in this study, was highly variable (with regards to the length of time dogs were confined and the frequency that they were released), in general it was observed that dogs that were tied up were often neglected. See Morters et al. (2014) for a full discussion on these issues.

Finally, the clinical signs variables cover a wide range of possible conditions. These were classified into ‘minor’ (considered unlikely to cause weight loss, such as localised skin lesions and lameness) and ‘major’ (considered likely to cause weight loss, such as vomiting and lethargy). This variable serves as an indicator of the general health of the dog, and it can be seen that as expected, dogs that show evidence of an ongoing medical condition (that is likely to cause weight loss), are more likely to have *lower* BCS values than their healthy counterparts: 1.2 times more likely for minor ailments and 2.7 times more likely for major ailments.

## 6 Discussion

We have introduced a method for fitting cumulative link ordinal regression models that does not require *a priori* assumptions regarding PO or NPO structures to model the relationship between the response and explanatory variables. For categorical explanatory variables we show how stochastic ordering can be ensured in the case of NPO models, and provide a pragmatic approach to ensuring that stochastic ordering holds for continuous or discrete covariates within the range of the observed data. In addition these approaches can be extended to incorporate variable selection within a Bayesian framework, allowing posterior probabilities of association to be produced for competing models. It is straightforward to include individual-level terms to account for repeated measures, and Bayesian model averaging can be used to provide weighted PPA estimates for the parameters that account for model uncertainty. We have illustrated the methods on a large-scale real-life data set.

The method uses reversible-jump MCMC to jump between models of differing dimensionality. However, implementation difficulties can exist with this method, particularly when jumping between models where the dimensionality is quite different. We found that the simple proposal mechanisms used throughout the paper worked well for this application and others we have tried. Nonetheless it is likely that specific situations may require more additional tuning (as with any MCMC method). For example, if some of the intervals specified by the stochastic ordering conditions (16) are small, and the proposal size,  $\tau_\beta$  is too large, then for NPO structures this may result in a large proportion of proposed values for the  $\beta_{kj}$  parameters being rejected as a result of breaking the prior conditions on stochastic ordering. An alternative would be to sample from some form of truncated distribution, though due to the nature of the constraints, this is not trivial.

Another interesting alternative would be to use some form of *shrinkage* model, where the model is defined as

$$\log\left(\frac{\gamma_{ij}}{1 - \gamma_{ij}}\right) = \theta_j - \boldsymbol{\eta}_j^T \mathbf{X}_i, \quad (40)$$

where  $i = 1, \dots, n$  and  $j = 1, \dots, J - 1$ . The conditional prior distributions for the  $\eta_{kj}$  parameters are centred around the corresponding  $\beta_k$  with a small prior variance. The  $\beta_k$  parameters can be given the same prior distribution as before. In this variation the model does not change dimensionality, and so no reversible-jump step is required. The  $\eta_{kj}$  parameters then correspond to the degree to which the parameter estimates deviate away from the proportional odds structure. This idea could also be expanded to incorporate variable selection in various ways (see e.g. O’Hara and Sillanpää, 2009).

Using single-component updates with simple random-walk proposals can also produce Markov chains that are highly autocorrelated, and thus require a large number of iterations and a lot of thinning. Adaptive proposal mechanisms (Haario et al., 2001; Roberts and Rosenthal, 2009) exist for standard (i.e. non-transdimensional) MCMC, that can automatically tune the proposal distributions to produce much more efficient chains in terms of both convergence and mixing. However, it is not currently understood whether these sorts of approaches hold for transdimensional routines, and this is a key area of ongoing research for those who are developing these methods (Hastie and Green, 2012). For the kinds of examples shown in this paper the runtimes required to produce a reasonable number of pseudo-independent samples are not prohibitive, and so we do not worry about this aspect here. It is not the purpose of this paper to provide a catch-all routine that works well in every situation, but rather to provide a flexible method that can be adapted to deal with different situations as required.

We occasionally noticed some identifiability issues when fitting NPO models, predominantly between categorical explanatory variables with low counts in some of the groups, and the cut-off parameters. This can be tackled in two main ways: firstly, the variables can be recategorised to ensure that there is a minimum number of individuals in each group. Secondly, we can start the MCMC routines using more informative initial values. Appealing to the Occam’s Razor principal, in this paper we decided to generate initial values by producing a short training run, using a PO model that includes all of the explanatory variables (but ignoring the repeated measures). We then ran the full

model using the parameter values from the final iteration of the training run as initial values. A similar approach would be to generate maximum likelihood estimates for the simple PO model and use these as initial values instead.

An example of the utility of these routines is that stochastic ordering can be ensured for continuous/discrete covariates within the range of the observed data. It is theoretically possible to ensure these conditions hold for any finite range of values, if an upper or lower bound was known from sources of information other than the observed data. In any case, if it is of interest to extrapolate beyond the range of the data, then it is possible to use the posterior samples to explore the range of covariate values over which the stochastic conditions will hold—essentially building a posterior distribution for the range of valid values. This could also be used as a form of sensitivity analysis to the model assumptions based on the model fit.

It is also worth noting that although we have illustrated these methods using a logistic link function, the methods are applicable to any monotonically increasing link function (though of course the interpretation of the regression parameters will no longer be in terms of the cumulative odds).

## Supplementary Material

Supplementary Materials: Bayesian model choice in cumulative link ordinal regression models: an application in a longitudinal study of risk factors affecting body condition score in a dog population in Zenzele, South Africa (DOI: [10.1214/14-BA884SUPP](https://doi.org/10.1214/14-BA884SUPP); .zip).

## References

- Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley, 2nd edition. [2](#), [7](#), [8](#)
- Akaike, H. (1974). “A new look at statistical model identification.” *IEEE Transactions on Automatic Control*, AU-19: 195–223. [3](#)
- Albert, J. and Chib, S. (1997). “Bayesian methods for cumulative, sequential and two-step ordinal data regression models.” Technical report. [4](#), [8](#)
- Albert, J. H. and Chib, S. (1993). “Bayesian analysis of binary and polychotomous response data.” *Journal of the American Statistical Association*, 88(422): 669–679. [1](#), [3](#), [7](#)
- Ananth, C. V. and Kleinbaum, D. G. (1997). “Regression models for ordinal responses: A review of methods and applications.” *International Journal of Epidemiology*, 26(6): 1323–1333. [2](#)
- Bender, R. and Grouven, U. (1998). “Using binary logistic regression models for ordinal data with non-proportional odds.” *Journal of Clinical Epidemiology*, 51(10): 809–816. [2](#), [8](#), [16](#)
- Brant, R. (1990). “Assessing proportionality in the proportional odds model for ordinal logistic regression.” *Biometrics*, 46(4): 1171–1178. [2](#)

- Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). “Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 65(1): 3–55. 13
- Chib, S. (1995). “Marginal likelihood from the Gibbs output.” *Journal of the American Statistical Association*, 90(432): 1313–1321. 4
- Chu, W. and Ghahramani, Z. (2005). “Gaussian processes for ordinal regression.” *Journal of Machine Learning Research*, 6: 1–48. 3, 4
- Cole, S. R., Allison, P. D., and Ananth, C. V. (2004). “Estimation of cumulative odds ratios.” *Annals of Epidemiology*, 14: 172–178. 2
- Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley. 7, 8
- Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). “On Bayesian model and variable selection using MCMC.” *Statistics and Computing*, 12: 27–36. 4, 6
- Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, 2nd edition. 3
- Fahrmeier, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer. 8
- Feinberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Springer. 2
- Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. CRC Press, 2nd edition. 5
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd edition. 5
- German, A. J. and Holden, S. L. (2006). “Subjective estimation of body condition can predict body fat mass as well as condition scoring with an established 9-point scale.” In *BSAVA Congress 2006 Scientific Proceedings*, 508. BSAVA Publications. 18
- Gibbons, R. D. and Hedeker, D. (1997). “Random effects probit and logistic regression models for three-level data.” *Biometrics*, 53: 1527–1537. 3
- Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996). *Markov Chain Monte Carlo In Practice*. Chapman and Hall. 5
- Green, P. J. (1995). “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.” *Biometrika*, 82(4): 711–732. 4, 5
- Haario, H., Saksman, E., and Tamminen, J. (2001). “An adaptive Metropolis algorithm.” *Bernoulli*, 7(2): 223–242. 25
- Hartzel, J., Agresti, A., and Caffo, B. (2001). “Multinomial logit random effects models.” *Statistical Modelling*, 1: 81–102. 3
- Hastie, D. I. and Green, P. J. (2012). “Model choice using reversible jump Markov chain Monte Carlo.” *Statistica Neerlandica*, 66(3): 309–338. 5, 25
- Hastings, W. (1970). “Monte Carlo sampling methods using Markov chains and their applications.” *Biometrika*, 57: 97–109. 5

- Hedeker, D. (2003). “A mixed-effects multinomial logistic regression model.” *Statistics in Medicine*, 22: 1433–1446. 3
- Hedeker, D. and Gibbons, R. D. (1994). “A random-effects ordinal regression model for multilevel analysis.” *Biometrics*, 50: 933–944. 3
- Holmes, C. C. and Held, L. (2006). “Bayesian auxiliary variable models for binary and multinomial regression.” *Bayesian Analysis*, 1(1): 145–168. 3, 4
- Ishwaran, H. (2000). “Univariate and multivariate ordinal cumulative link regression with covariate specific cutpoints.” *The Canadian Journal of Statistics*, 28(4): 715–730. 8
- Ishwaran, H. and Gatsonis, C. A. (2000). “A general class of hierarchical ordinal regression models with applications to correlated ROC analysis.” *The Canadian Journal of Statistics*, 28(4): 731–750. 3
- Jeffreys, H. (1935). “Some tests of significance, treated by the theory of probability.” *Proceedings of the Cambridge Philosophy Society*, 31: 203–222. 5
- (1961). *The Theory of Probability*. Oxford, 3rd edition. 5
- Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer-Verlag, New York. 7
- Kass, R. E. and Raftery, A. E. (1995). “Bayes Factors.” *Journal of the American Statistical Association*, 90(430): 773–795. 3, 4, 5, 22
- Krzanowski, W. J. (1998). *An Introduction to Statistical Modelling*. Arnold. 15
- Lall, R., Campbell, M. J., Walters, S. J., Morgan, K., and MRC CFAS Co-operative (2002). “A review of ordinal regression models applied on health-related quality of life assessments.” *Statistical Methods in Medical Research*, 11: 49–67. 2
- Lang, J. B. (1999). “Bayesian ordinal and binary regression models with a parametric family of mixture links.” *Computational Statistics and Data Analysis*, 31: 59–87. 3
- Leon-Novelo, L. G., Zhou, X., Bekele, B. N., and Müller, P. (2010). “Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories.” *Biometrics*, 66: 966–974. 3
- Liang, H., Wu, H., and Zou, G. (2008). “A note on conditional AIC for linear mixed-effects models.” *Biometrika*, 95(3): 773–778. 3
- Link, W. A. and Eaton, M. J. (2012). “On thinning of chains in MCMC.” *Methods in Ecology and Evolution*, 3: 112–115. 15
- Liu, L. C. and Hedeker, D. (2006). “A mixed-effects regression model for longitudinal multivariate ordinal data.” *Biometrics*, 62: 261–268. 3
- McCullagh, P. (1980). “Regression models for ordinal data (with discussion).” *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2): 109–142. 1, 6, 7
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). “Equations of state calculations by fast computing machine.” *Journal of Chemical Physics*, 21: 1087–1091. 5

- Morters, M., McKinley, T., Restif, O., Conlan, A., Cleaveland, S., Hampson, K., Whay, B., Damriyasa, I. M., and Wood, J. (2014). “The demography of free-roaming dog populations and applications to disease and population control.” *to appear in Journal of Applied Ecology*. 2, 3, 17, 24
- Mwalili, S. M., Lesaffre, E., and Declerck, D. (2005). “A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study.” *Applied Statistics*, 54: 77–93. 4
- O’Brien, S. M. and Dunson, D. B. (2004). “Bayesian multivariate logistic regression.” *Biometrics*, 60: 739–746. 4
- O’Hara, R. B. and Sillanpää, M. J. (2009). “A review of Bayesian variable selection methods: what, how and which.” *Bayesian Analysis*, 4(1): 85–118. 3, 4, 5, 13, 14, 19, 22, 25
- Paquet, U., Holden, S., and Naish-Guzman, A. (2005). “Bayesian hierarchical ordinal regression.” In Duch, W., Oja, E., and Zadrozny, S. (eds.), *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, 267–272. Springer-Verlag Berlin Heidelberg. 4
- Peterson, B. and Harrell, F. E. (1990). “Partial proportional odds models for ordinal response variables.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(2): 205–217. 2, 10
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). “CODA: Convergence Diagnosis and Output Analysis for MCMC.” *R News*, 6(1): 7–11.  
URL <http://CRAN.R-project.org/doc/Rnews/> 15
- R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.  
URL <http://www.R-project.org/> 15
- Richardson, S. and Green, P. (1997). “On Bayesian analysis of mixtures with an unknown number of components (with discussion).” *Journal of the Royal Statistical Society. Series B (Methodological)*, 59: 731–792. 6
- Roberts, G. O. and Rosenthal, J. S. (2009). “Examples of adaptive MCMC.” *Journal of Computational and Graphical Statistics*, 18(2): 349–367. 25
- Tutz, G. and Scholz, T. (2003). “Ordinal regression modelling between proportional odds and non-proportional odds.” Technical report, Institute of Statistics, University of Munich. 2, 4, 7, 8
- Urbanek, S. (2011). *multicore: Parallel processing of R code on machines with multiple cores or CPUs*. R package version 0.1-7.  
URL <http://CRAN.R-project.org/package=multicore> 15
- Vaida, F. and Blanchard, S. (2005). “Conditional Akaike information for mixed-effects models.” *Biometrika*, 92(2): 351–370. 3
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, fourth edition. ISBN 0-387-95457-0.  
URL <http://www.stats.ox.ac.uk/pub/MASS4> 16

- Viallefont, V., Raftery, A. E., and Richardson, S. (2001). “Variable selection and Bayesian model averaging in case-control studies.” *Statistics in Medicine*, 20: 3215–3230. [3](#), [20](#), [22](#)
- Waagepetersen, R. and Sorensen, D. (2001). “A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping.” *International Statistical Review*, 69(1): 49–61. [5](#)
- Webb, E. L. and Forster, J. J. (2008). “Bayesian model determination for multivariate ordinal and binary data.” *Computational Statistics and Data Analysis*, 52: 2632–2649. [3](#), [4](#)
- Yi, N., Banerjee, S., Pomp, D., and Yandell, B. S. (2007). “Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits.” *Genetics*, 176: 1855–1864. [3](#)

**Acknowledgments**

TJM is supported by Biotechnology and Biological Sciences Research Council grant number BB/I012192/1. MM is supported by a grant from the International Fund for Animal Welfare (IFAW) and the World Society for the Protection of Animals (WSPA), with additional support from the Charles Slater Fund and the Jowett Fund. JW is supported by the Alborada Trust and the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security and the Fogarty International Centre. Thanks to Andrew Conlan and Richard Dybowski for useful discussions, and to the anonymous referees whose comments and suggestions helped greatly improve this manuscript.