stats

MDPI

Article

Smoothing in Ordinal Regression: An Application to Sensory Data

Ejike R. Ugba [1], Daniel Mörlein [2] and Jan Gertheiss [1,*]

[1] Department of Mathematics and Statistics, School of Economics and Social Sciences, Helmut Schmidt University, 22043 Hamburg, Germany; ugbae@hsu-hh.de
[2] Department of Animal Sciences, Faculty of Agricultural Sciences, Georg August University, 37077 Göttingen, Germany; daniel.moerlein@uni-goettingen.de
* Correspondence: jan.gertheiss@hsu-hh.de

Abstract: The so-called proportional odds assumption is popular in cumulative, ordinal regression. In practice, however, such an assumption is sometimes too restrictive. For instance, when modeling the perception of boar taint on an individual level, it turns out that, at least for some subjects, the effects of predictors (androstenone and skatole) vary between response categories. For more flexible modeling, we consider the use of a 'smooth-effects-on-response penalty' (SERP) as a connecting link between proportional and fully non-proportional odds models, assuming that parameters of the latter vary smoothly over response categories. The usefulness of SERP is further demonstrated through a simulation study. Besides flexible and accurate modeling, SERP also enables fitting of parameters in cases where the pure, unpenalized non-proportional odds model fails to converge.

Keywords: animal welfare; Brant test; categorical data; quality control; regularization; sensometrics

Citation: Ugba, E.R.; Mörlein, D.; Gertheiss, J. Smoothing in Ordinal Regression: An Application to Sensory Data. *Stats* **2021**, *4*, 616–633. https://doi.org/10.3390/stats4030037

Academic Editor: Dungang Liu

Received: 31 May 2021
Accepted: 15 July 2021
Published: 21 July 2021

1. Introduction

The production of entire male pigs is an alternative to surgical castration taking animal welfare concerns into account. However, the elevated levels of so-called *boar taint* may impair consumer acceptance (see, for example, the Ref. [1] and references therein). Boar taint is (presumably) caused by two malodorous volatile substances, namely: androstenone and skatole (compare, for example, the Ref. [2]). In an experimental study presented in [3], fat samples from roughly 1000 pig carcasses were collected and subjected to a thorough sensory evaluation and quantification using a panel of 10 trained assessors on a sensory score scale ranging from 0 = 'untainted' to 5 = 'strongly tainted'. The absolute frequencies of the panelist scores are shown in Figure 1. The question of interest is how the sensory evaluation is influenced by the samples' androstenone and skatole content. In this context, the Ref. [3] considered the average panel ratings as the response, which gives a quasi-continuous variable, and made standard linear modeling the approach of choice. As an alternative, panel ratings may be discretized to a binary outcome, with a typical cut-point for dichotomization (boar-tainted/no boar taint) fixed at 2; compare, for example, the Ref. [2]. With this, a binary (e.g., logit) model can be used instead of standard linear modeling. On an individual, subject-specific level, dichotomization/binary regression is thus a sensible approach as well, whereas linear modeling with a relatively small number of (ordinal) response categories may be questionable [4]. However, dichotomization still poses the problem of loss of information and choice of the threshold. Thus, for a clearer understanding of the individual panelists' rating patterns of deviant smell, we (a) consider an ordinal model utilizing as much information as possible, and (b) fit those models to each panelist separately. The latter is done because effects of androstenone and skatole can be very different between people, while it is very important to examine those subject-specific effects, for instance when selecting potential raters to identify boar-tainted carcasses at the slaughter line. That is why [3] also provided modeling on an individual level as

supplementary information (online appendix), but only considered the dichotomized data. Thus, with prediction/inference in mind, to realize an adequate predictive model of deviant smell via an ordinal regression model, we consider the following rating scale: no boar taint (0/1), low boar taint (2), medium boar taint (3) and high boar taint (4/5), where the extreme categories are collapsed (due to the small number of observations in categories 1 and 5).

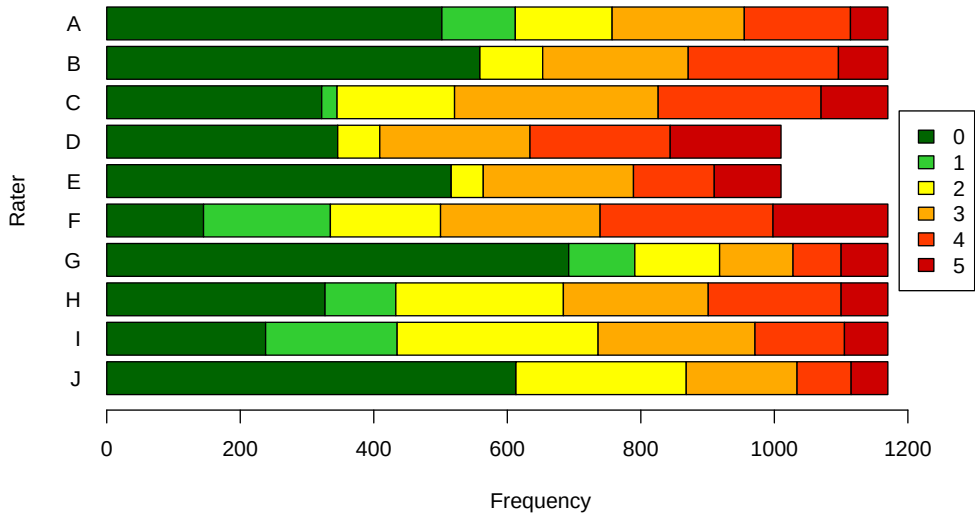

Figure 1. Summary of the individual sensory ratings for each rater/panelist; the scores of deviant smell run from 0 = 'untainted' to 5 = 'strongly tainted'.

A very popular type of ordinal regression model is the cumulative logistic model, particularly the so-called proportional odds model [5]. Using the latter, however, imposes some restrictions that may not be true for at least some of our raters. Omitting those restrictions and allowing for the most flexible cumulative logit model, on the other hand, may result in numerical problems in the fitting algorithm, high variability in the estimated effects of androstenone and skatole, and results that are hard to interpret. We hence discuss a regularized cumulative model here that is a data-driven compromise between proportional and fully non-proportional odds models, assuming that parameters of the latter vary smoothly across categories. An idea that has already been applied successfully with rating scales as predictors [6]. As will be illustrated, the model proposed maintains the flexibility of the non-proportional odds model and adapts very well to the underlying data structure, while at the same time providing competitive or better accuracy with respect to parameter estimates and prediction. Furthermore, results are typically easier to interpret. The rest of the paper is organized as follows: we begin with a short review of cumulative models for ordinal response in Section 2, and introduce our smoothing/penalty approach in Section 3. A simulation study is found in Section 4, while application to the sensory data is provided in Section 5. Section 6 concludes with a discussion.

2. Cumulative Models for Ordinal Response

Models for categorical outcome variables have been the subject of many discussions, with several approaches available in the literature for various forms of empirical applications; see, for example, the Refs. [7,8]. Assuming a categorical response variable Y, with k distinct but ordered categories, the information supplied by each response category can be incorporated in a model using the ordinal rather than the multinomial class of models [8,9]. The ordered model that is probably most frequently used is the cumulative model developed by McCullagh [5], which is not only popular from a frequentist, but also a Bayesian point of view; see, for example, the Ref. [10]. The model is often motivated by an underlying/latent, continuous variable, say $\tilde{Y}$, and a linear model of the form:

$$\tilde{Y}_i = -\boldsymbol{x}_i^\top \tilde{\boldsymbol{\delta}} + \epsilon_i, \tag{1}$$

where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a vector of covariates observed on unit $i = 1, \ldots, n$, $\tilde{\boldsymbol{\delta}}$ is the vector of corresponding regression parameters, and ϵ_i an error term with continuous distribution function F. Then, it is assumed that the observable, ordinal response Y is obtained via the threshold model

$$Y_i = r \quad \Leftrightarrow \quad \tilde{\delta}_{0,r-1} < \tilde{Y}_i < \tilde{\delta}_{0r},$$

with $-\infty = \tilde{\delta}_{00} < \tilde{\delta}_{01} < \cdots < \tilde{\delta}_{0k} = \infty$ being cut-points on the (latent) scale of $\tilde{Y}$. It then follows that

$$\begin{aligned} P(Y_i \leq r|\boldsymbol{x}_i) &= P(-\boldsymbol{x}_i^\top \tilde{\boldsymbol{\delta}} + \epsilon_i \leq \tilde{\delta}_{0r}) = P(\epsilon_i \leq \tilde{\delta}_{0r} + \boldsymbol{x}_i^\top \tilde{\boldsymbol{\delta}}) \\ &= F(\tilde{\delta}_{0r} + \boldsymbol{x}_i^\top \tilde{\boldsymbol{\delta}}). \end{aligned} \tag{2}$$

The choice of F in (2) results in different forms of the cumulative model. Normally distributed ϵ_i, for instance, leads to the so-called (cumulative) probit model. Note, however, that Gaussian ϵ_i does not mean that the latent $\tilde{Y}_i$ is also marginally normal. Neither thresholds $\tilde{\delta}_{0r}$ are assumed to be equidistant. As a consequence, skewed or bimodal (ordinal) data can also be analyzed within the framework of a cumulative/latent variable model.

Besides the probit model, the most popular cumulative model is the cumulative logit model, which is obtained from F being the logistic distribution function such that

$$\log \frac{P(Y_i \leq r|\boldsymbol{x}_i)}{P(Y_i > r|\boldsymbol{x}_i)} = \tilde{\delta}_{0r} + \boldsymbol{x}_i^\top \tilde{\boldsymbol{\delta}}, \quad r = 1, \ldots, k-1. \tag{3}$$

In case of our sensory data, however, it makes sense to rewrite model (3) in terms of

$$\log \frac{P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i \leq r|\boldsymbol{x}_i)} = \delta_{0r} + \boldsymbol{x}_i^\top \boldsymbol{\delta}, \quad r = 1, \ldots, k-1, \tag{4}$$

and $\delta_{0r} = -\tilde{\delta}_{0r}$, $\boldsymbol{\delta} = -\tilde{\boldsymbol{\delta}}$. Now model (4) gives the (log) odds of 'deviant smell', if threshold r is used for dichotomization, that is, to distinguish 'deviant' from 'normal' smell. This model (either in form (3) or (4)) is typically referred to as the proportional odds model (POM), since the effect of the covariates does not depend on the cut-point r, but is rather constant across categories. In other words, the odds ratio when increasing a specific covariate by one unit is the same for all cut-points r. For instance, as a 'textbook example', we may fit a proportional odds model to the data, as shown in Figure 1, with covariates androstenone and skatole, and a rater-specific effect in (1). Here, and throughout the paper, the two covariates were standardized after being transformed logarithmically. An interaction effect was also incorporated (as done in [3]). In a few cases (14 per rater), however, androstenone had a value of zero, which may be due to androstenone content below the detection threshold, or defective measurement. Therefore, those observations were excluded from further analysis. In the case of a consumer study with a large sample of consumers drawn at random from the population, and each consumer just evaluating a relatively small number of products, we would set the rater/consumer effect as a random effect. With a hand-picked panel of raters, however, and each panelist evaluating about 1000 products (see Figure 1), those rater-specific effects are set as fixed effects, that is, as an additional factor giving the rater.

As a next step, when comparing this model to a more complicated model where the effects of androstenone and skatole also vary with the rater (both fitted using `polr()` from the R package `MASS` [11]), the latter model is significantly better, with the p-value being virtually zero (likelihood ratio test with LR = 219.023, df = 27). This confirms the statement made earlier that the effects of androstenone/skatole can be very different between raters. One step further, it also makes sense to have parameters δ_{0r} vary with raters. Since the model with both rater-specific thresholds and effects of androstenone/skatole varying with the rater is equivalent to fitting separate models for each rater, the latter will be done throughout the paper (as those individual models are easier to interpret).

In practice, however, the proportional odds assumption made so far is also sometimes violated; compare, for example, the Ref. [12]. In general, if the effects of covariates turn out to vary (substantially) across response categories, the proportional odds assumption will produce biased results. In such a situation, a more general form of the model (4) that relaxes the proportional odds assumption may be used, although at the expense of increased model complexity. The general cumulative logit model, or rather the non-proportional odds model (NPOM), is given by

$$\log\frac{P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i \leq r|\boldsymbol{x}_i)} = \eta_{ir} = \delta_{0r} + \boldsymbol{x}_i^\top \boldsymbol{\delta}_r, \quad r = 1,\ldots,k-1, \tag{5}$$

where $\boldsymbol{\delta}_r = (\delta_{1r},\ldots,\delta_{pr})^\top$, and the restrictive global effect $\boldsymbol{\delta}$ is now replaced by a more liberal category-specific effect that accounts for every single response class/cut-point r in the model. Model (5) has the property of stochastic ordering [5], which implies that $\delta_{0,r+1} + \boldsymbol{x}_i^\top \boldsymbol{\delta}_{r+1} < \delta_{0r} + \boldsymbol{x}_i^\top \boldsymbol{\delta}_r$ holds for all $\boldsymbol{x}_i$ and all $r = 1,\ldots,k-2$, since $P(Y_i > r+1|\boldsymbol{x}_i) < P(Y_i > r|\boldsymbol{x}_i)$ must hold for all categories. However, it is often the case that such a constraint is not met during the iterative procedure typically used to fit the model, leading to unstable likelihoods with ill-conditioned parameter space. This is one reason why the simpler model (4) is often adopted in practice even when inappropriate. Alternatively, the two different forms of effects (4) and (5) may be combined in one model. In other words, assuming $\boldsymbol{x}_i$ is partitioned into $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ such that $\boldsymbol{x}_i^\top = (\boldsymbol{u}_i^\top, \boldsymbol{v}_i^\top)$, one obtains the so-called partial proportional odds model (PPOM) [13] as follows:

$$\log\frac{P(Y_i > r|\boldsymbol{x}_i)}{P(Y_i \leq r|\boldsymbol{x}_i)} = \delta_{0r} + \boldsymbol{u}_i^\top \boldsymbol{\delta} + \boldsymbol{v}_i^\top \boldsymbol{\gamma}_r, \quad r = 1,\ldots,k-1, \tag{6}$$

where $\boldsymbol{u}_i$ has a global effect $\boldsymbol{\delta}$ and $\boldsymbol{v}_i$ has a category-specific effect $\boldsymbol{\gamma}_r$. PPOM could be of help when it comes to reducing model complexity, but at the extra cost of clustering candidate covariates to a particular effect type.

To distinguish between POM and NPOM, proportional odds tests may come into play, such as likelihood ratio tests. However, with those tests being dependent on the existence of the likelihood of the general model, which is often ill-conditioned, such tests are often not feasible. One test that is independent of the likelihood of the general model is the so-called Brant test [14]. This test examines the proportionality assumption of the entire model (omnibus) alongside each of the individual variables in the model. The approach is based on viewing (5) as a combination of $k-1$ correlated binary logistic regressions. Brant shows that the separate binary logistic regression estimates $\hat{\boldsymbol{\delta}}_1,\ldots,\hat{\boldsymbol{\delta}}_{k-1}$ for $\boldsymbol{\delta}_r$, $r = 1,\ldots,k-1$, are asymptotically unbiased and follow a multivariate normal distribution. Consequently, a Wald-type test that is based on the differences in the estimated coefficients, producing a chi-square statistic, could be used. Thus, with $\boldsymbol{\delta}_r$ all equal under POM, any $\hat{\boldsymbol{\delta}}_r - \hat{\boldsymbol{\delta}}_l$, $r \neq l$ makes a possible test statistic for testing the proportional odds assumption, also componentwise. The corresponding test, however, suffers from low power and "may provide no clear indication as to the nature of the discrepancy [from the proportional odds model] detected" [14]. Therefore, Brant focused on testing $H_0 : \boldsymbol{\delta}_r = \boldsymbol{\delta}$ for all r vs. $H_1 : \boldsymbol{\delta}_r = \phi_r\boldsymbol{\delta}$, where $\phi_r > 0$ captures misspecification of the distributional form of the latent variable, in this instance, a nonlogistic link function. Indeed, when applying the Brant test to the cumulative logit model of the sensory data (see Table 1), it turned out that not all the models met the parallel slope assumption (see the highlighted p-values). The observed non-proportionality is more pronounced in the overall model than respective covariates. In summary, about half of the models under consideration failed the parallel slope assumption.

Table 1. The p-values of the Brant test of proportional odds across categories of the individual panelist ratings of deviant smell in boar samples; with usual significance codes: '***' 0.001 '**' 0.01 '*' 0.05.

	Brant Test (p-Values)			
Rater	**Overall**	**Androstenone**	**Skatole**	**Interaction**
A	0.044 *	0.732	0.022 *	0.879
B	0.246	0.421	0.841	0.239
C	0.144	0.209	0.111	0.359
D	0.008 **	0.026 *	0.375	0.288
E	0.018 *	0.043 *	0.496	0.912
F	0.992	0.969	0.861	0.817
G	0.000 ***	0.886	0.000 ***	0.546
H	0.531	0.599	0.690	0.142
I	0.000 ***	0.676	0.000 ***	0.465
J	0.787	0.414	0.311	0.923

However, the reliability of this test and similar conventional tests for validating the proportional odds assumption has often been criticized for being prone to misleading conclusions in empirical applications; see, for example, the Refs. [15,16]. Using approaches other than statistical hypothesis testing have been recommended: the Ref. [16], for instance, suggests a graphical approach for validating the proportional odds assumption, and [17] proposed modified residuals that can also be used to check the proportional odds assumption. Given our sensory data, a very intuitive graphical approach is to examine the contour plots of the estimated log-odds of being in dichotomized categories of the deviant smell model for different cut-points, r. Figures 2 and 3 show the corresponding log-odds as a function of the predictors androstenone and skatole under NPOM and POM for the seventh and eighth panelist (G, H), respectively. In addition, the dichotomized data using cut-point $r = 1, 2, 3$ are given as red/blue dots, since the cumulative logistic model (5) can be interpreted such that a binary logit model is employed on the dichotomized data using potential threshold $r = 1, 2, 3$. The log-odds shown in Figure 2 for both NPOM (top) and POM (bottom) indicate that the odds of being in the upper categories (i.e., categories of more severe boar taint) increase for increasing androstenone and skatole. With NPOM (top), however, the shape of the contour lines changes between columns, that is, thresholds r, whereas log-odds all have the same shape across cut-points for POM (bottom) by construction (as δ coefficients do not change across r, compare (4)). Having the relatively large sample size in mind ($n \approx 1000$ for each panelist), this may indicate that the proportional odds assumption is inappropriate here. For panelist H (Figure 3, top) though, contour plots change very much for different cut-points r as well, but appear rather erratic and hard to interpret, in contrast to POM results (bottom). From the latter, we get the clear picture of (log-)odds of deviant smell (i.e., upper categories) that are particularly varying in the androstenone direction (increasing for increasing androstenone), since contour lines are rather parallel to the skatole axis. Using NPOM (top), by contrast, we can hardly make such a statement. As a consequence, we may be willing to give up the flexibility of NPOM (5) in order to have a model fit that can be interpreted. In light of those findings, it would be desirable to have a tool available for moving NPOM estimates towards POM automatically, if and as far as it is supported by the data. Besides interpretation, there is another very important advantage of POM over NPOM. The former is much simpler, and the estimates' variance is typically smaller, which can lead to a smaller mean squared error even in the case of bias: the so-called *bias variance trade-off*.

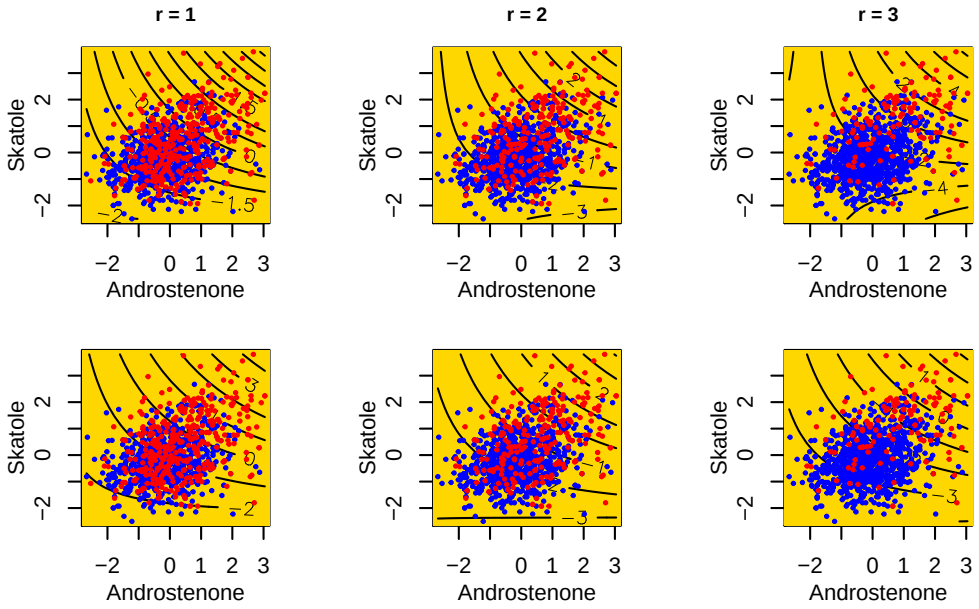

Figure 2. Fitted log-odds of sensory perception of boar taint under NPOM (**upper** row) and POM (**lower** row), having (log-transformed, standardized) androstenone and skatole, plus interaction as explanatory variables. Each column denotes the log-odds of panelist G's rating falling into the upper categories with cut-point $r = 1, 2, 3$ (column 1 to 3). The data observed are drawn as colored dots: red, if $Y_i > r$; blue, if $Y_i \leq r$.

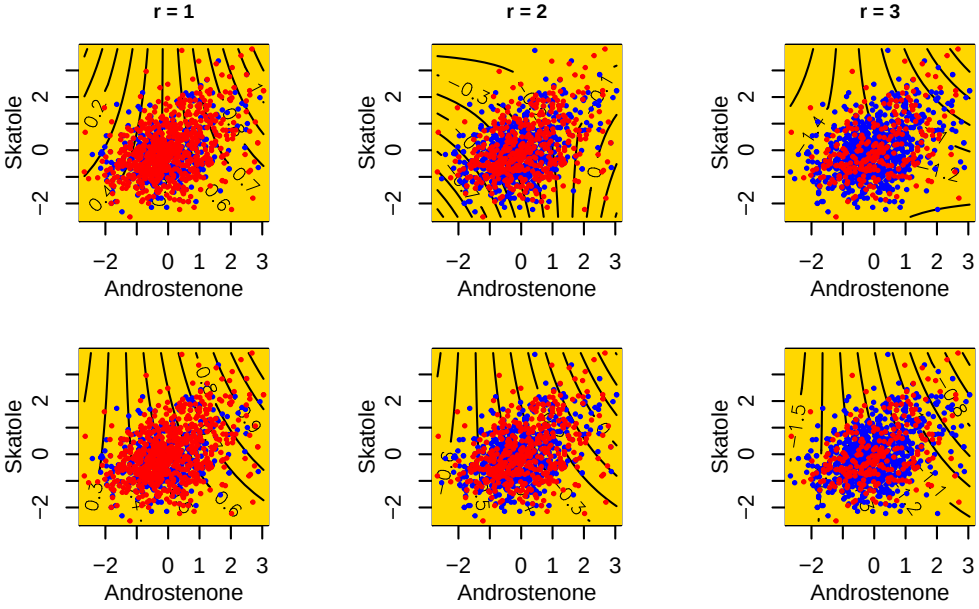

Figure 3. Fitted log-odds of sensory perception of boar taint under NPOM (**upper** row) and POM (**lower** row), having (log-transformed, standardized) androstenone and skatole, plus interaction as explanatory variables. Each column denotes the log-odds of panelist H's rating falling into the upper categories with cut-point $r = 1, 2, 3$ (column 1 to 3). The data observed are drawn as colored dots: red, if $Y_i > r$; blue, if $Y_i \leq r$.

To this end, the use of shrinkage penalties could be considered a viable means of reaching a good, data-driven compromise between the non-proportional and proportional odds model. In other words, when putting an appropriate penalty on parameters of the more general, non-proportional odds model, a trade-off may be found between bias and variance of estimated parameters. Several types of penalties have already been suggested in the literature for categorical models. On the one hand, these comprise of penalties adopted

from regularization methods for continuous models; see, for example, the Refs. [18–20]. On the other hand, there are penalties specifically designed to suit the need of categorical variables. A review and discussion of the latter is, for instance, given by [21], also sketching a smoothing penalty for ordinal regression. However, the idea is neither discussed in detail nor applied to data. In this study, we will further investigate the approach, which we call the 'smooth-effects-on-response penalty' (SERP), and use it for modeling our sensory data in a more flexible way than POM.

3. Smooth Ordinal Regression

For a smooth transition from the general model (NPOM) to the restricted model (POM), we consider the use of a specific penalty. This penalty enables the parameters of NPOM to be smoothened across response categories, resulting in a data-driven compromise between most flexible but potentially over-complex NPOM, and very popular but potentially too restrictive POM.

3.1. Smoothing Penalty

To begin with, one maximizes the following penalized log-likelihood,

$$l_{\mathrm{p}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - J_{\lambda}(\boldsymbol{\theta}), \tag{7}$$

or more specifically,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{l(\boldsymbol{\theta}) - J_{\lambda}(\boldsymbol{\theta})\} = \arg\min_{\boldsymbol{\theta}} \{-l(\boldsymbol{\theta}) + J_{\lambda}(\boldsymbol{\theta})\},$$

where $l(\boldsymbol{\theta})$, in this context, denotes the log-likelihood of the general cumulative logit model (5) and $J_{\lambda}(\boldsymbol{\theta}) = \lambda J(\boldsymbol{\theta})$ the penalty function $J(\boldsymbol{\theta})$ weighted by the tuning parameter λ. The vector $\boldsymbol{\theta}^{\top} = (\boldsymbol{\delta}_0^{\top}, \boldsymbol{\delta}^{\top})$ collects the thresholds/constants $\boldsymbol{\delta}_0^{\top} = (\delta_{01}, \ldots, \delta_{0,k-1})$ and slope parameters $\boldsymbol{\delta}^{\top} = (\delta_{11}, \ldots, \delta_{1,k-1}, \ldots, \delta_{p1}, \ldots, \delta_{p,k-1})$ from model (5). Thus, given ordered categorical outcomes $Y_i \in \{1, \ldots, k\}$, and considering that slopes $\delta_{j1}, \ldots \delta_{j,k-1}$, $j = 1, \ldots, p$, vary smoothly over the categories, the following penalty [21],

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\delta}) = \sum_{j=1}^{p} \sum_{r=1}^{k-2} (\delta_{j,r+1} - \delta_{jr})^2, \tag{8}$$

affects smoothing across response categories such that all category-specific effects associated with a covariate turn towards a common global effect. The intercepts (thresholds) are generally not penalized, but would be automatically adjusted while the other parameters are being penalized. Equation (8) in matrix form can be expressed as follows:

$$J(\boldsymbol{\delta}) = \boldsymbol{\delta}^{\top} \boldsymbol{M} \boldsymbol{\delta},$$

where, for a single predictor model, $\boldsymbol{\delta}^{\top} = (\delta_1, \ldots, \delta_{k-1})$ and $\boldsymbol{M} = \boldsymbol{D}^{\top}\boldsymbol{D}$ a $[(k-1) \times (k-1)]$ symmetric matrix, with $\boldsymbol{D}$ a $[(k-2) \times (k-1)]$ matrix of the first-order differences given by

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix}.$$

Thus, with p predictors in NPOM, the following block matrix structure is needed to enforce smoothness on all adjacent response categories associated with respective predictors of an ordered model:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{O} & & & \\ & \boldsymbol{M} & & \\ & & \ddots & \\ & & & \boldsymbol{M} \end{pmatrix},$$

where $\boldsymbol{\Omega}$ is a $[(k-1)(p+1) \times (k-1)(p+1)]$ block diagonal matrix with elements $\boldsymbol{M}$ (p times) and a matrix $\boldsymbol{O}$ consisting of zeros in the upper left corner (which makes sure that intercepts are not penalized). Then, one has

$$J(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta}.$$

Let the overall design matrix be given by $\boldsymbol{X}^\top = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)$ and let $\boldsymbol{L}_i(\boldsymbol{\theta}) = \partial h(\boldsymbol{\eta}_i)/\partial \boldsymbol{\eta}$ be the derivative of $h(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta}_i = \boldsymbol{X}_i \boldsymbol{\theta}$, where $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{i,k-1})^\top$, compare (5), and h is the inverse link, that is, the response function of the cumulative logit model. Then, the penalized score function $\boldsymbol{s}_\mathrm{p}(\boldsymbol{\theta})$ is given by

$$\boldsymbol{s}_\mathrm{p}(\boldsymbol{\theta}) = \boldsymbol{X}^\top \boldsymbol{L}(\boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})(\boldsymbol{z} - \boldsymbol{\mu}) - \lambda \boldsymbol{\Omega} \boldsymbol{\theta},$$

where $\boldsymbol{L}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{L}_i(\boldsymbol{\theta}))$ is the block-diagonal matrix of derivatives, $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{diag}(\Sigma_1(\boldsymbol{\theta}), \ldots, \Sigma_n(\boldsymbol{\theta}))$ the block-diagonal matrix of covariance matrices of k-dimensional binary response vectors $\boldsymbol{z}_i$ indicating the category of observation i, $\boldsymbol{z}^\top = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)$ the combined vector of observed values, and $\boldsymbol{\mu}^\top = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_n^\top)$ the vector of mean vectors, that is, k-dimensional class probabilities $\boldsymbol{\mu}_i = h(\boldsymbol{X}_i \boldsymbol{\theta})$. Equating the score function to zero yields the estimation equation $\boldsymbol{s}_\mathrm{p}(\boldsymbol{\theta}) = \boldsymbol{0}$, which may be solved with the following iterative routine:

$$\hat{\boldsymbol{\theta}}^{[t+1]} = \hat{\boldsymbol{\theta}}^{[t]} + \boldsymbol{F}_\mathrm{p}^{-1}(\hat{\boldsymbol{\theta}}^{[t]}) \boldsymbol{s}_\mathrm{p}(\hat{\boldsymbol{\theta}}^{[t]}),$$

where $\boldsymbol{F}_\mathrm{p}(\boldsymbol{\theta}) = \boldsymbol{F}(\boldsymbol{\theta}) + \lambda \boldsymbol{\Omega}$ is the penalized/pseudo-Fisher information matrix, $\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{X}^\top \boldsymbol{W}(\boldsymbol{\theta}) \boldsymbol{X}$ the Fisher information and $\boldsymbol{W}(\boldsymbol{\theta}) = \boldsymbol{L}(\boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \boldsymbol{L}(\boldsymbol{\theta})^\top$ the weight matrix. Assuming $\hat{\boldsymbol{\theta}}^{[0]} = (\hat{\delta}_0^{[0]\top}, \hat{\boldsymbol{\delta}}^{[0]\top})^\top$ is the vector of the initial $(k-1)(p+1)$ values in the algorithm, in our particular case (logit link), $\hat{\delta}_0^{[0]}$ is obtained from the logistic transformation (see the left-hand side of (5)) of the cumulative, relative class frequencies in the data; $\hat{\boldsymbol{\delta}}^{[0]}$, on the other hand, is a $p(k-1)$ vector of zeros. Alternatively, POM estimates may be used as starting values. Given that $\hat{\boldsymbol{\theta}}$ contains penalized estimates for $\boldsymbol{\theta}$ parameters, the approximate covariances are obtained by the sandwich matrix:

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}) \approx (\boldsymbol{F}(\hat{\boldsymbol{\theta}}) + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{F}(\hat{\boldsymbol{\theta}}) (\boldsymbol{F}(\hat{\boldsymbol{\theta}}) + \lambda \boldsymbol{\Omega})^{-1}, \tag{9}$$

where all notations are as defined earlier. For more details on the estimation procedures for cumulative models and the like, see, for example, the Refs. [8,22].

3.2. Choosing the Tuning Parameter and Measures of Performance

Since the penalty term $J_\lambda(\boldsymbol{\delta})$ in (7) depends on the tuning parameter λ, a suitable value of λ needs to be determined. Common practice is to fit the penalized model for a sequence of λ values and select the best value via a tuning routine [23,24]. A typical tuning routine entails choosing the best model based on criteria such as AIC, BIC, and so forth. Alternatively, the λ value at which the model's out-of-sample prediction error is minimal can be determined via cross-validation. There are various specifications of such prediction errors in the literature, including, classification error, squared error, minus log-likelihood error, and so forth. The pros and cons of some of these error types are, for example, reviewed in [25,26]. The squared error (or Brier score [27]) particularly captures the sum of squared distances of the observed classes and the predicted values/probabilities. We refer to this here as the mean squared prediction error (MSPE). For a multi-class model with response categories $1, \ldots, k$, it is defined as

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k}(z_{ir} - \hat{\pi}_{ir})^2, \quad \text{where } z_{ir} = \begin{cases} 1 & \text{if } y_i = r \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

is the indicator variable for each (potential) level of the dependent variable, $\hat{\pi}_{ir}$ are the predicted probabilities for the respective categories and subject i, and n is the total number of observations. We will use MSPE alongside other error metrics to determine λ and compare the performance of our proposed approach with standard approaches. Another common performance metric is the mean squared error (MSE). Given that the true parameters δ_{jr} are known (as they are in simulation studies), we can obtain the MSE of parameter estimates as follows:

$$\text{MSE} = \frac{1}{(k-1)p}\sum_{j=1}^{p}\sum_{r=1}^{k-1}(\hat{\delta}_{jr} - \delta_{jr})^2, \tag{11}$$

where k and p are the number of response categories and predictors in the model, respectively. Also, the MSE can be calculated covariate/component-wise.

Please note, in order to reduce complexity, we used a single, global penalty parameter λ in (7). In general, we could also use covariate-specific penalty parameters λ_j within the penalty term (8) in terms of $J_{\lambda_1,\ldots,\lambda_p}(\boldsymbol{\delta}) = \sum_{j=1}^{p}\lambda_j\sum_{r=1}^{k-2}(\delta_{j,r+1} - \delta_{jr})^2$. This, however, would mean that cross-validation needs to be carried out over a multi-dimensional space.

4. Numerical Experiments

Before applying SERP to the sensory data, the effect and performance of the penalty shall be investigated in simulation studies where the truth is known. Following Equation (5), the probabilities $P(Y_i > r|\boldsymbol{x}_i)$, $r = 1,\ldots,3$, were obtained with the two covariates x_{i1} and x_{i2}, including an interaction, where both variables are iid $N(0,1)$, and $i = 1,\ldots,1000$. However, on the one hand, the intercepts $\boldsymbol{\delta}_0 = (\delta_{01},\ldots,\delta_{03})^\top$ were set to be equidistant as follows: $\boldsymbol{\delta}_0 = (0.1,-1.0,-2.1)^\top$, and on the other hand, the slope parameters $\boldsymbol{\delta}_j = (\delta_{j1},\ldots,\delta_{j3})^\top$, $j = 1,2,3$, were selected to form three different simulation settings as follows:

(a) **Varying coefficients** for x_{1i} and x_{2i} where, $\boldsymbol{\delta}_1 = (0.3,0.4,0.5)^\top$, and $\boldsymbol{\delta}_2 = (0.4,0.5,0.6)^\top$;
(b) **Constant coefficients** for x_{1i} and x_{2i} where, $\boldsymbol{\delta}_1 = (0.3,0.3,0.3)^\top$, and $\boldsymbol{\delta}_2 = (0.4,0.4,0.4)^\top$ and
(c) **Varying and constant coefficients** for x_{1i} and x_{2i}, respectively, where $\boldsymbol{\delta}_1 = (0.3,0.4,0.5)^\top$, and $\boldsymbol{\delta}_2 = (0.4,0.4,0.4)^\top$.

Moreover, the interaction effects $\boldsymbol{\delta}_3 = (\delta_{31},\ldots,\delta_{33})^\top$ were also chosen to reflect the three different settings as follows: varying $\boldsymbol{\delta}_3 = (0.1,0.2,0.5)^\top$; constant $\boldsymbol{\delta}_3 = (0.1,0.1,0.1)^\top$; partly constant/varying $\boldsymbol{\delta}_3 = (0.1,0.1,0.5)^\top$. The realized $P(Y_i = r|\boldsymbol{x}_i) = P(Y_i > r-1|\boldsymbol{x}_i) - P(Y_i > r|\boldsymbol{x}_i)$ were subsequently used on a multinomial distribution to generate corresponding observed values, $y_i \in \{1,\ldots,4\}$. In general, the number of response categories, and sample and effect size(s) were chosen to be comparable to the sensory data from Sections 1, 2 and 5.

For an illustrative application of SERP on ordinal models, SERP is inflicted on a cumulative logit model built from the generated data, using a grid of $\lambda \in [0,\infty)$. As shown in Figure 4 (top), at large values of λ, all NPOM estimates level up to POM estimates; see the dashed blue horizontal line on each display. The down displays provide a second visualization of SERP's smoothing steps from NPOM's original set of estimates towards POM's estimates, represented with line strokes across ordinal levels/cut-points r. Again, we have the initial cut-point specific estimates (solid black) all shrunken to the parallel estimates (dashed blue horizontal lines) with several tuning steps of SERP in between (gray). An optimal value of λ regulating the degree of smoothing can be determined following a predefined criterion, for instance, a λ value minimizing a performance metric; see, for example, Equation (10) in Section 3. In this particular setting, the unique vertical

(dashed red) lines on the top and the non-horizontal dashed lines in the down displays of Figure 4 give the resulting set of estimates at which the 5-fold cross-validated test errors (MSPE) were minimal. Unless stated otherwise, this tuning criterion is used throughout this study for SERP-fitted models.

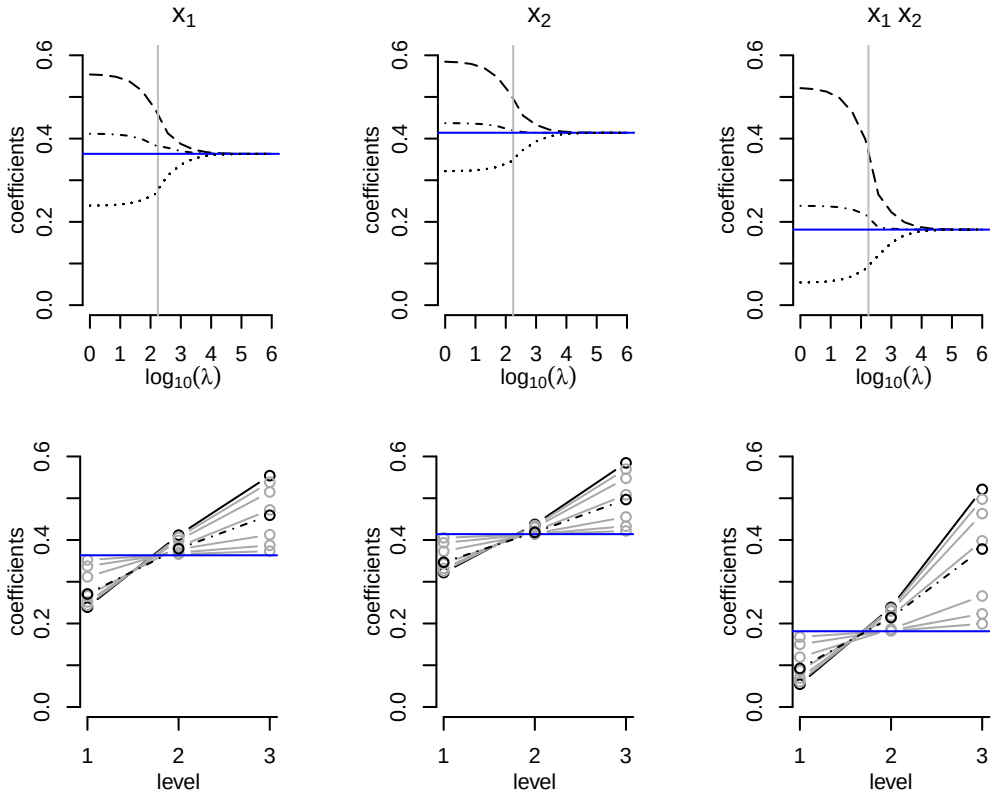

Figure 4. Estimated coefficients when using SERP under the first simulation setting, that is, (a) varying coefficients. The black lines on the top displays are the category-specific coefficient paths of $\hat{\delta}_{j1}$ (dotted), $\hat{\delta}_{j2}$ (dashed/dotted), and $\hat{\delta}_{j3}$ (dashed), $j = 1, 2, 3$, associated with the two predictors plus interaction used in the model. The solid horizontal blue and vertical gray lines denote the parallel estimates and the selected estimates based on 5-fold cross-validation, respectively. The bottom row further illustrates SERP's smoothing steps from the category-specific to the parallel estimates, the solid black, gray and blue line strokes are NPOM, SERP and POM estimates, respectively; with the dashed/dotted black lines indicating SERP estimates chosen via cross-validation.

We next investigate SERP's improvement on (N)POM (where the denotation (N)POM refers to POM and/or NPOM). Thus, following the described data-generating process, 100 replications (each of SERP and (N)POM) were obtained for the three different simulation settings. For comparison purposes, a test set error (MSPE) of each of the models was obtained for all the simulation runs. In addition to that, the MSE of estimates with respect to the true slope parameters plus interaction were also obtained. Figure 5 shows the pairwise differences across simulation runs in MSE and MSPE of (N)POM and SERP, with the horizontal dashed line in each plot indicating the mark of SERP's improvement over (N)POM. In other words, differences above the dashed lines indicate better performance of SERP than (N)POM. As observed in the first simulation setting (column 1), where the underlying coefficients vary across categories, SERP outperformed both NPOM and POM in terms of the MSE and MSPE. In the second simulation setting (column 2), where truly constant coefficients generated the data, POM expectantly performed distinctively better than NPOM. SERP, however, adapted very well and gave estimates that are as good as POM. The third simulation setting (column 3) had varying underlying coefficients for the first covariate and constant coefficients in the second. As before, SERP adapted very well

(compare the MSE for the three coefficient vectors $\delta_1, \delta_2, \delta_3$), thus producing models with better predictions on average than both NPOM and POM. It should be pointed out at this point that we used one, global λ here. Nevertheless, SERP is highly competitive to POM on (truly constant) δ_2, while performing much better on (truly varying) δ_1. If compared to NPOM, SERP is superior on δ_2 and competitive on δ_1. This indicates that coefficients are decisively shrunken towards proportionality in the first case (δ_2), while allowing for substantial non-proportionality in the latter (δ_1).

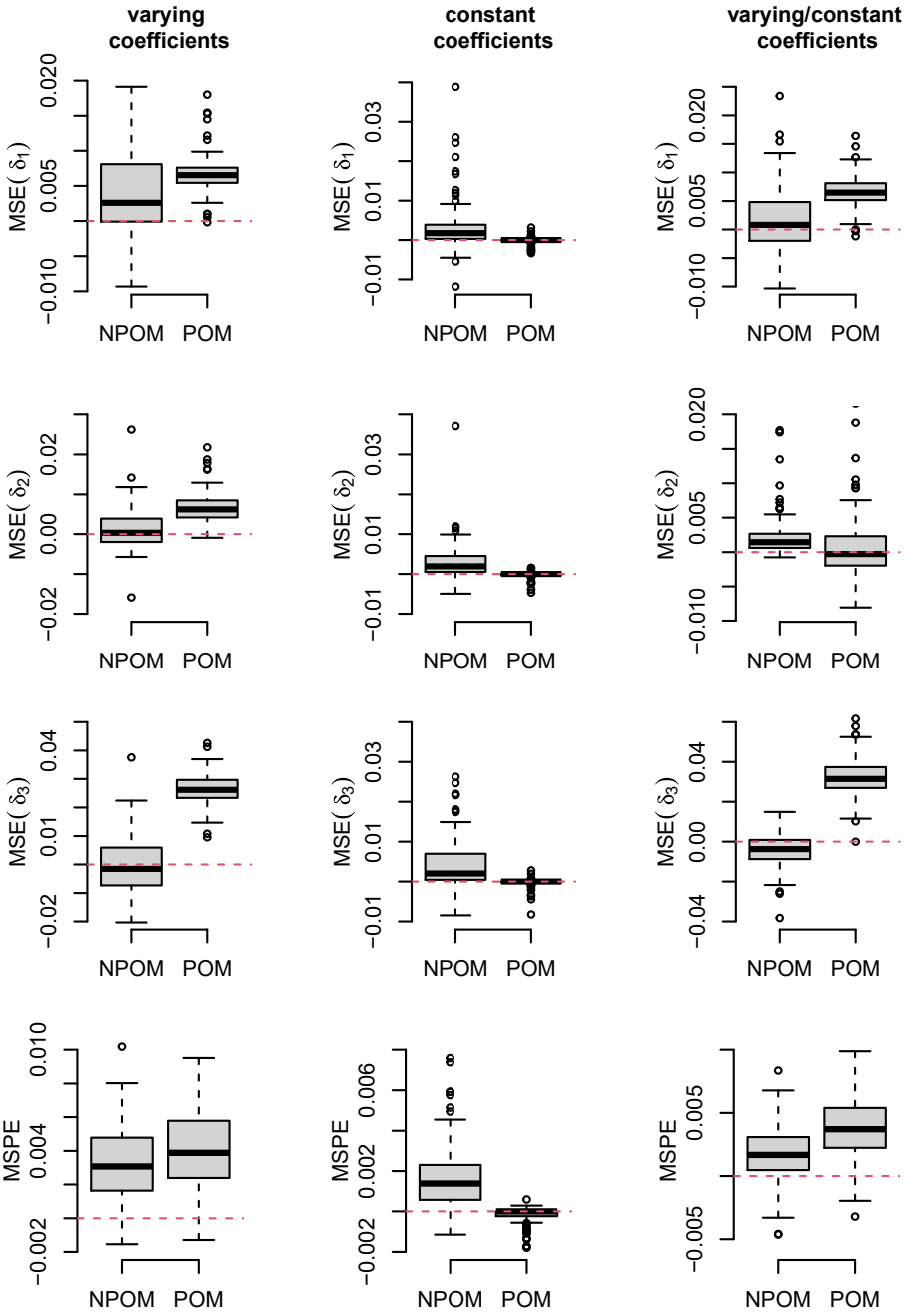

Figure 5. The pairwise differences of SERP and (N)POM across simulation runs with respect to the MSE (rows 1–3) of the slope parameter estimates and the MSPE (last row), given three different simulation settings (column-wise): (1) varying coefficients, (2) constant coefficients, and (3) varying and constant coefficients. The horizontal dashed lines indicate the mark of SERP's improvement over (N)POM, with differences above the dashed lines indicating better performance with SERP.

Finally, it should be noted that we only provided settings here that are comparable to the real, sensory data considered in Sections 1, 2, and 5 below. In general, however, it has been our experience that, if the models considered are identifiable within cross-validation, the penalty parameter chosen will yield results that are (at least) competitive to NPOM and/or POM.

5. Application to Sensory Evaluation

We continue with our real data problem from Section 1, where panelists' ratings of the degree of deviation from a normal smell were modeled using the two covariates, androstenone and skatole (each log-transformed, standardized), plus the interaction effect. For a detailed discussion of the experiment(s) to generate this data set, and a preliminary analysis of all variables of interest, the reader is referred to [3]. Our interest here is on the rater-specific ordinal models of the scores of deviant smell.

We have already introduced the cumulative logit model of the individual raters in Section 2. In order to understand the rating patterns of individual panelists and for an accurate inference, it is necessary to determine whether (and to what extent) category-specific effects or global effects are suitable for the individual models of deviant smell. Moreover, it is necessary to have the parameters in the model and the model itself be completely identifiable. SERP could, therefore, provide the means of arriving at a good set of estimates other than (N)POM's original estimates, as well as help to induce convergence where NPOM fails to converge. Hence, cumulative logit models of deviant smell were obtained for all the panelists using SERP and the two standard approaches, POM and NPOM. The obtained estimates and standard errors of SERP, with standard errors being extracted from (9), together with (N)POM are exemplarily given for panelist G and H in Tables 2 and 3, respectively. Standard errors (SE) could also be used to calculate common, approximate 95% confidence intervals in terms of 'estimate $\pm$ 2 SE'. It should be noted though, that those SE are obtained for a given smoothing parameter; compare (9). That means that we treat them as if they would have been specified a priori. Although this is commonly done in penalized regression, it ignores variation that is induced by cross-validation (or other methods used to find an appropriate λ in practice). As a consequence, those SE are typically biased downwards and lead to some under-coverage of confidence intervals. Of course, usual POM confidence intervals (in terms of SERP, obtained for $\lambda \to \infty$) are conditioned on the much stricter assumption of proportional odds, and hence are only valid if this assumption is true. With respect to point estimates, we see that in the case of panelist G, SERP shrinks NPOM estimates towards POM, but still produces a non-proportional odds model. With panelist H, by contrast, cross-validated λ yields a proportional odds model as all slope coefficients are (virtually) constant across cut-points (compare Table 3). We shall later examine the extent of SERP's improvement over (N)POM via re-sampling procedures.

Figure 6 shows the log-odds of deviant smell for panelists G, H, and I, in analogy to Figures 2 and 3. Again, we see (top row) that SERP shrinks the estimates for panelist G towards the proportional odds model, but still maintains some non-proportionality; also compare Figure 2 and the Supplementary Material, where for each panelist NPOM, POM, and SERP are shown together in one graphic. With panelist I (bottom row), we also have a smoothed version of NPOM (compare the Supplementary Material). In the case of panelist H (center row), as already observed above (Table 3), the (nearly) proportional odds model is obtained (compare Figure 3 and the Supplementary Material). More generally speaking, not only does SERP help to locate the 'best' set of coefficients, but one could also make some informed decision as to which of POM and NPOM or the combination of estimates from both models, that is, partially proportional odds (6), could be adequate in an empirical study. Another point that is nicely seen from Figure 6 and Tables 2 and 3, is that the effects of androstenone and skatole can indeed be very different between people. On the one hand, effects are much stronger for panelists G and I than for panelist H. On the other hand, panelist H senses boar taint rather for increasing androstenone than skatole, as

iso-lines look rather parallel to the y-axis in Figure 6 (middle row). For panelist G, it is a combination of both, whereas, according to panelist I, deviant smell is mainly caused by increased levels of skatole (iso-lines rather parallel to x-axis in Figure 6, bottom).

Table 2. Estimates and standard errors (SE) of regression coefficients in SERP and (N)POM for the sensory rating of panelist G, with androstenone (AN), skatole (SK) and interaction (AN:SK) as predictors of deviant smell.

	NPOM		**SERP**		**POM**	
	Estimates	**SE**	**Estimates**	**SE**	**Estimates**	**SE**
(Intercept):1	−0.774	0.141	−0.785	0.070	−0.824	0.071
(Intercept):2	−1.368	0.141	−1.363	0.078	−1.478	0.081
(Intercept):3	−2.154	0.141	−2.077	0.078	−2.317	0.104
AN:1	0.352	0.142	0.352	0.072	0.388	0.074
AN:2	0.370	0.142	0.372	0.074		
AN:3	0.401	0.142	0.401	0.074		
SK:1	0.479	0.140	0.536	0.076	0.644	0.077
SK:2	0.638	0.140	0.584	0.078		
SK:3	0.994	0.140	0.654	0.078		
AN:SK:1	0.130	0.114	0.161	0.072	0.165	0.066
AN:SK:2	0.215	0.114	0.201	0.073		
AN:SK:3	0.478	0.114	0.273	0.073		

Table 3. Estimates and standard errors (SE) of regression coefficients in SERP and (N)POM for the sensory rating of panelist H, with androstenone (AN), skatole (SK) and interaction (AN:SK) as predictors of deviant smell.

	NPOM		**SERP**		**POM**	
	Estimates	**SE**	**Estimates**	**SE**	**Estimates**	**SE**
(Intercept):1	0.522	0.065	0.526	0.064	0.526	0.064
(Intercept):2	−0.335	0.063	−0.353	0.063	−0.354	0.063
(Intercept):3	−1.233	0.075	−1.220	0.073	−1.219	0.073
AN:1	0.108	0.067	0.107	0.059	0.107	0.058
AN:2	0.134	0.066	0.108	0.059		
AN:3	0.080	0.077	0.106	0.059		
SK:1	0.004	0.069	0.028	0.061	0.030	0.061
SK:2	0.049	0.068	0.031	0.061		
SK:3	0.044	0.080	0.031	0.062		
AN:SK:1	0.026	0.057	0.019	0.049	0.019	0.049
AN:SK:2	−0.034	0.055	0.016	0.049		
AN:SK:3	0.051	0.061	0.021	0.050		

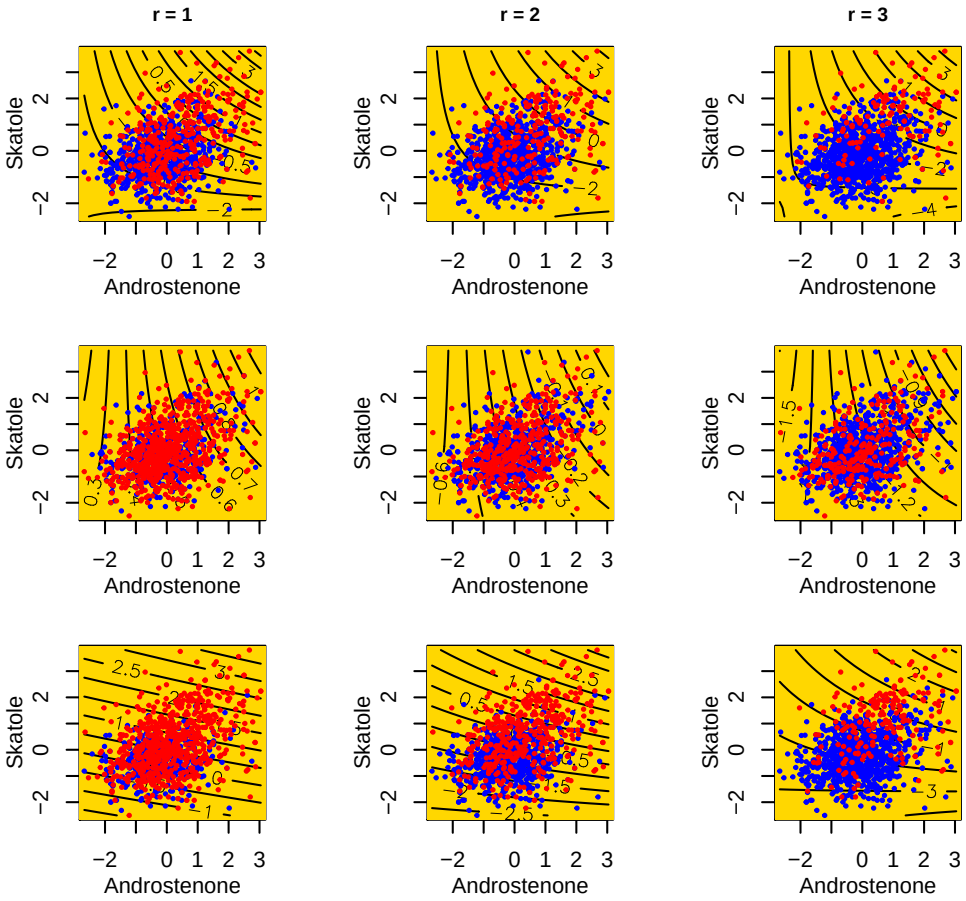

Figure 6. The result of the empirical application of SERP on the sensory data, showing the fitted log-odds of deviant smell ratings as a function of (log-transformed, standardized) androstenone and skatole for different cut-points r (the columns), and for panelists G, H and I (the rows). Data points observed (dots) are color-coded, where $Y_i > r$ (red dots) and $Y_i \leq r$ (blue dots).

Further comparison of SERP and the other methods were made with respect to the out-of-sample MSPE of the respective methods. Those were obtained for each of the panelists on a randomly chosen test set amounting to 20% of the original data set and over 100 replications of this experiment. Figure 7 captures the pairwise prediction error differences of SERP against (N)POM (with differences above zero favoring SERP). As observed, SERP shows distinct improvement over NPOM for 9 out of 10 panelists, as a greater amount of differences are seen over the dashed horizontal line for the respective panelists. Improvement over POM, however, is much less pronounced than SERP's improvement over NPOM. For panelists A, D, E, and G, at least the median of differences is (slightly) above zero. For the other panelists, though, SERP is still not worse than POM. In summary, SERP appears to be a safe choice when it comes to modeling/prediction of panelists' ratings in our application. Compared to standard approaches (POM and NPOM), results in terms of prediction accuracy are at least similar, but often better. With respect to boar taint perception, our results impressively show that people can be very different, both in terms of the effects of androstenone and skatole, as well as (non-)proportionality. This needs to be kept in mind, for instance, when hiring raters for detecting boar-tainted carcasses at the slaughter line.

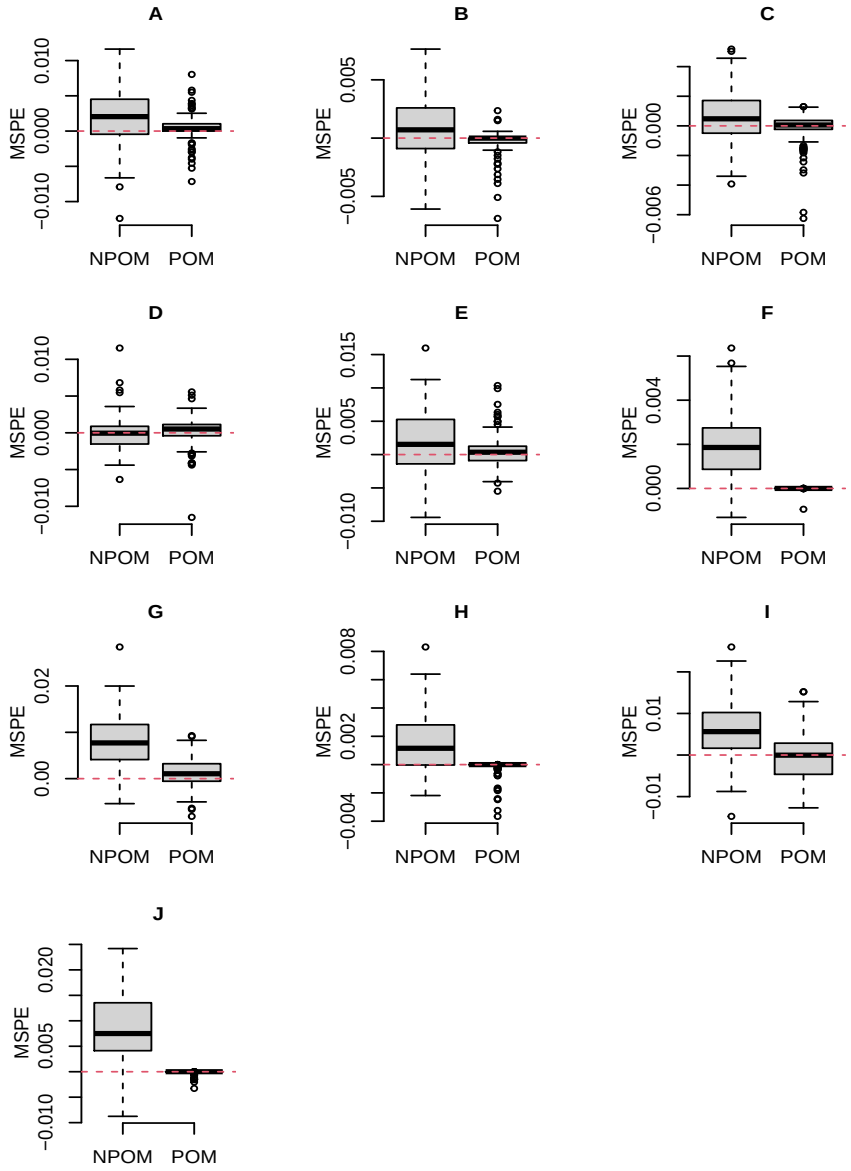

Figure 7. The result of the empirical application of SERP on the sensory data, showing the pairwise differences of the out-of-sample MSPE of SERP from (N)POM, for all panelists (**A**–**J**). The horizontal dashed lines indicate the mark of SERP improvement over NPOM and POM.

6. Discussion

Regularization has been a topic of intensive research in statistics for decades now. However, regularization methods that are specifically designed for categorical data are relatively new. Particularly, the focus has rarely been on ordinal response models. This might be due to the fact that the model most frequently used for ordinal response data (if not employing a standard linear model) is the cumulative model with global effects, particularly the proportional odds model, which can be nicely motivated via a latent variable and a corresponding linear model. Hence, most of the regularization methods proposed for ordinal response data can be seen as extensions of methods typically found in the high-dimensional linear or generalized linear model framework. For instance, the Ref. [28] considered high-dimensional genomic data and forward stagewise regression in the proportional odds model, the Ref. [29] proposed a Boosting approach for variable selection, the Ref. [30] used a sparsity-inducing Bayesian framework, whereas [20] employed a penalty approach. In the latter case, class-specific parameters were also considered, but only in a continuation ratio model, similar to [19]. Instead of a penalty term, the Ref. [31] proposed

to use pseudo-observations in order to regularize and stabilize fitting in the proportional odds model.

The proportional odds assumption, however, rules out category-specific effects. Since, with our sensory data, it is very clear that at least for some raters the proportional odds assumption is too restrictive, alternatives are needed. Simply allowing for category-specific effects that are estimated via usual maximum likelihood, however, is hardly an option because the model becomes too complex, which leads to large variance in the fitted coefficients, or even numerical issues like non-convergence. We hence investigated a penalty approach for smoothing effects across ordered categories. The new approach, called SERP, showed very encouraging results both in simulation studies and our sensory data. As observed, SERP makes it possible to find a good compromise in a completely data-driven manner between the purely non-proportional odds model and the usual but restrictive assumption of proportionality. Additionally, SERP may be used to check in a rather informal way for partial (non-)proportionality, that means, it may help to make a decision on the structure of the partial proportional odds model (6). If supported by the data, coefficients for some variables may be shrunken towards proportionality, while the coefficients for other variables still indicate non-proportionality.

A penalty very similar to SERP has been proposed by [32] in the bivariate ordered logistic model. In this framework, the authors also proposed a partial likelihood ratio test (PLRT) which works with penalized likelihood. As an alternative to the Brant test, a corresponding PLRT could also be used with our data and methodology to distinguish between raters where the proportional odds assumption might be acceptable, and those where it is significantly violated. In our case, however, the Brant test turned out to be superior in both size and power. For instance, the PLRT was very sensitive to the smoothing parameter λ. Particularly, it produced unsatisfactory results for large λ (where the results derived by [32] do not hold). When deciding between global and category-specific effects after penalized fitting employing SERP, we hence prefer making the decision in a rather informal way on the basis of the estimated penalty parameter and regression coefficients.

Besides (non-)proportional odds models as considered in this paper, various other models and methods have been proposed for handling ordinal data. For instance, the stereotype model [33], probabilistic index models [34–36], and rank-based models [37,38], just to name a few. In the present paper, however, we focused on cumulative models in combination with a specific type (8) of quadratic, first-order smoothing penalty. An Ref. [39] implementation of the proposed method is provided as add-on package `serp` [40], available from open access repositories CRAN and Github.

Supplementary Materials: Contour plots of the odds (on log scale) of sensory perception of boar taint under NPOM, SERP and POM for all panelists in analogy to Figures 2, 3 and 6 are available online at https://www.mdpi.com/article/10.3390/stats4030037/s1, Figures S1–S10.

Author Contributions: Conceptualization, J.G. and D.M.; methodology, J.G. and E.R.U.; software, E.R.U.; validation, E.R.U. and J.G.; data analysis, E.R.U. and J.G.; data curation, E.R.U., J.G. and D.M.; writing—original draft preparation, E.R.U.; writing—review and editing, J.G. and D.M.; supervision, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by Deutsche Forschungsgemeinschaft grant number GE2353/2-1. The sensory data was collected as part of the STRAT-E-GER project, which was supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program.

Institutional Review Board Statement: Not applicable, as this research only involved secondary analysis of existing data, collected for the purposes of a prior study [3].

Informed Consent Statement: Not applicable, as this research only involved secondary analysis of existing data, collected for the purposes of a prior study [3].

Data Availability Statement: The sensory data [3] analyzed in this paper is available from zenodo. Additionally, a second real data example (analysis of wine dataset) using the proposed method is provided in R package `serp` [40].

Acknowledgments: The authors want to thank the editor and two anonymous reviewers for their very constructive comments that helped to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Trautmann, J.; Gertheiss, J.; Wicke, M.; Mörlein, D. How olfactory acuity affects the sensory assessment of boar fat: A proposal for quantification. *Meat Sci.* **2014**, *98*, 255–262. [CrossRef]
2. Meier-Dinkel, L.; Gertheiss, J.; Müller, S.; Wesoly, R.; Mörlein, D. Evaluating the performance of sensory quality control: The case of boar taint. *Meat Sci.* **2015**, *100*, 73–84. [CrossRef]
3. Mörlein, D.; Trautmann, J.; Gertheiss, J.; Meier-Dinkel, L.; Fischer, J.; Eynck, H.J.; Heres, L.; Looft, C.; Tholen, E. Interaction of skatole and androstenone in the olfactory perception of boar taint. *J. Agric. Food Chem.* **2016**, *64*, 4556–4565. [CrossRef]
4. Larrabee, B.; Scott, H.M.; Bello, N.M. Ordinary least squares regression of ordered categorical data: Inferential implications for practice. *J. Agric. Biol. Environ. Stat.* **2014**, *19*, 373–386. [CrossRef]
5. McCullagh, P. Regression models for ordinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1980**, *42*, 109–142. [CrossRef]
6. Tutz, G.; Gertheiss, J. Rating scales as predictors–the old question of scale level and some answers. *Psychometrika* **2014**, *79*, 357–376. [CrossRef]
7. Agresti, A. *Categorical Data Analysis*, 2nd ed.; John Wiley and Sons: New York, NY, USA, 2002.
8. Tutz, G. *Regression for Categorical Data*; University Press: Cambridge, UK, 2011.
9. Tutz, G. Ordinal regression: A review and a taxonomy of models. *WIREs Comput. Stat.* **2021**, e1545. [CrossRef]
10. Sha, N.; Dechi, B.O. A Bayes inference for ordinal response with latent variable approach. *Stats* **2019**, *2*, 23. [CrossRef]
11. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; ISBN 0-387-95457-0.
12. Irvine, K.M.; Rodhouse, T.J.; Keren, I.N. Extending ordinal regression with a latent zero-augmented Beta distribution. *J. Agric. Biol. Environ. Stat.* **2016**, *21*, 619–640. [CrossRef]
13. Peterson, B.; Harrell, F.E. Partial proportional odds models for ordinal response variables. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1990**, *39*, 205–217. [CrossRef]
14. Brant, R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **1990**, *46*, 1171–1178. [CrossRef]
15. Bender, R.; Grouven, U. Using binary logistic regression models for ordinal data with non-proportional odds. *J. Clin. Epidemiol.* **1998**, *51*, 809–816. [CrossRef]
16. Harrell, F.E. *Regression Modeling Strategies*; Springer: New York, NY, USA, 2001.
17. Liu, D.; Zhang, H. Residuals and diagnostics for ordinal regression models: A surrogate approach. *J. Am. Stat. Assoc.* **2018**, *113*, 845–854. [CrossRef] [PubMed]
18. Le Cessie, S.; Van Houwelingen, J.C. Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1992**, *41*, 191–201. [CrossRef]
19. Archer, K.; Williams, A. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Stat. Med.* **2012**, *31*, 1464–1474. [CrossRef] [PubMed]
20. Tran, T.; Phung, D.; Luo, W.; Venkatesh, S. Stabilized Sparse Ordinal Regression for Medical Risk Stratification. *Knowl. Inf. Syst.* **2015**, *43*, 555–582. [CrossRef]
21. Tutz, G.; Gertheiss, J. Regularized regression for categorical data. *Stat. Model.* **2016**, *16*, 161–200. [CrossRef]
22. Fahrmeir, L.; Tutz, G. *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed.; Springer: New York, NY, USA, 2001.
23. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
24. Sun, W.; Wang, J.; Fang, Y. Consistent selection of tuning parameters via variable selection stability. *J. Mach. Learn. Res.* **2013**, *14*, 3419–3440.
25. Van Houwelingen, J.C.; Le Cessie, S. Predictive value of statistical models. *Stat. Med.* **1990**, *9*, 1303–1325. [CrossRef]
26. Efron, B. How biased is the apparent error rate of a prediction rule? *J. Am. Stat. Assoc.* **1986**, *81*, 461– 470. [CrossRef]
27. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather. Rev.* **1950**, *78*, 1–3. [CrossRef]
28. Hou, J.; Archer, K.J. Regularization method for predicting an ordinal response using longitudinal high-dimensional genomic data. *Stat. Appl. Genet. Mol. Biol.* **2015**, *14*, 93–111. [CrossRef]
29. Zahid, F.M.; Tutz, G. Proportional odds models with high-dimensional data structure. *Int. Stat. Rev.* **2013**, *81*, 388–406. [CrossRef]
30. Satake, E.; Majima, K.; Aoki, S.C.; Kamitani, Y. Sparse Ordinal Logistic Regression and Its Application to Brain Decoding. *Front. Neuroinform.* **2018**, *12*, 51. [CrossRef]
31. Zahid, F.M.; Ramzan, S.; Heumann, C. Regularized proportional odds models. *J. Stat. Comput. Simul.* **2015**, *85*, 251–268. [CrossRef]

32. Enea, M.; Lovison, G. A penalized approach for the bivariate ordered logistic model with applications to social and medical data. *Stat. Model.* **2019**, *19*, 467–500. [CrossRef]
33. Anderson, J.A. Regression and ordered categorical variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1984**, *46*, 1–30. [CrossRef]
34. Thas, O.; De Neve, J.; Clement, L.; Ottoy, J.P. Probabilistic index models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2012**, *74*, 623–671. [CrossRef]
35. De Neve, J.; Thas, O. A regression framework for rank tests based on the probabilistic index model. *J. Am. Stat. Assoc.* **2015**, *110*, 1276–1283. [CrossRef]
36. De Schryver, M.; De Neve, J. A tutorial on probabilistic index models: Regression models for the effect size P(Y1<Y2). *Psychol. Methods* **2019**, *24*, 403–418. [PubMed]
37. Akritas, M.G.; Brunner, E. Nonparametric models for ANOVA and ANCOVA: A review. In *Recent Advances and Trends in Nonparametric Statistics*; Akritas, M.G., Politis, D.N., Eds.; Elsevier: Amsterdam, The Netherlands, 2003; pp. 79–91.
38. Chatterjee, D.; Bandyopadhyay, U. Testing in nonparametric ANCOVA model based on ridit reliability functional. *Ann. Inst. Stat. Math.* **2019**, *71*, 327–364. [CrossRef]
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
40. Ugba, E.R. serp: Smooth Effects on Response Penalty for CLM. R Package Version 0.1.8. 2021. Available online: https://CRAN.R-project.org/package=serp (accessed on 20 July 2021)