

RESEARCH NOTE

Open Access



Statistical inference for ordinal predictors in generalized additive models with application to Bronchopulmonary Dysplasia

Jan Gertheiss^{1*} , Fabian Scheipl², Tina Lauer^{3,4} and Harald Ehrhardt^{3,4}

Abstract

Objective: Discrete but ordered covariates are quite common in applied statistics, and some regularized fitting procedures have been proposed for proper handling of ordinal predictors in statistical models. Motivated by a study from neonatal medicine on Bronchopulmonary Dysplasia (BPD), we show how quadratic penalties on adjacent dummy coefficients of ordinal factors proposed in the literature can be incorporated in the framework of generalized additive models, making tools for statistical inference developed there available for ordinal predictors as well.

Results: The approach presented allows to exploit the scale level of ordinally scaled factors in a sound statistical framework. Furthermore, several ordinal factors can be considered jointly without the need to collapse levels even if the number of observations per level is small. By doing so, results obtained earlier on the BPD data analyzed could be confirmed.

Keywords: Chronic lung disease, Logit model, Ordinal data, Regularization, Smoothing penalty

Introduction

Bronchopulmonary Dysplasia (BPD) is a chronic lung disease often found in preterm infants with lungs not fully developed. Disturbance of lung development and severity of BPD is caused by various peri- and postnatal factors including prematurity itself, as well as pre- and postnatal infections [1]. BPD is measured on ordinal scale with grades 0, 1, 2, 3, but often dichotomized as 0: ‘no/mild BPD’ and 1: ‘moderate/severe BPD’. One goal of the study reported here is to investigate whether the time after birth some specific bacteria were found for the first time in the children’s upper airways has an effect on BPD. Initially, $n = 102$ preterm infants with a birth weight < 1000 g and gestational age $\leq 32 + 0$ weeks were analyzed within a retrospective cohort study at the tertiary

perinatal center of Justus-Liebig-University Giessen (Germany) between January 2014 and June 2017. Two infants, however, had to be excluded from further analyses at some point due to missing information on some bacterial colonization. Earlier analyses already showed that the later bacteria were found, the lower the risk of developing BPD [2]. However, it is not fully understood yet which specific bacteria have an effect, and in which way. Therefore we will draw special attention to the time period/week (after birth) three types of bacteria—gram negative/positive and pathogenic—were found for the first time in the upper airway of the respective child. Although ‘time’ is supposed to be a continuous variable, information is only available in a discretized way here, because samples were only obtained once a week. Furthermore, the last category ‘week > 6 ’ is open/censored. If an observation is falling in the last category, we only know that until week six the respective germ had not been found yet. So, the corresponding covariate may only be considered as categorical but ordinal.

*Correspondence: jan.gertheiss@hsu-hh.de

¹ School of Economics and Social Sciences, Helmut Schmidt University/University of the Federal Armed Forces, Hamburg, Germany
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Besides the information on bacteria, some additional risk factors need to be taken into account, such as the weight and sex of the child, the number of days antibiotics and steroids were given, or information on multiples. For doing so, a logit model with categorical/ordinal predictor ‘bacteria/week’ and additional, potentially confounding covariates may be fit. Typically, a categorical predictor is included as dummy-coded factor, ignoring, however, the information on the categories’ ordering (if so). In the case presented, additional problems are caused by the fact that some categories/levels only have a few observations, and sometimes all those observations are falling in the same response category. Consequently, some coefficients in a logit model fit via usual maximum likelihood tend towards $\pm\infty$.

For preventing numerical problems like inflating regression coefficients, penalization can be a viable solution [3]. Furthermore, a penalty term can be used for exploiting/respecting a covariate’s ordinal scale level. Following [4–7], for instance, a difference/smoothing penalty might be put on adjacent dummy coefficients of the ordinal factor when fitting the model. An approach that has already been applied successfully in medical research; see, e.g., [8, 9]. In the BPD application, however, the question naturally arises how to test for significance of the ordinal predictor in the penalized setting. In a linear model with normal errors, this can be done using a (restricted) likelihood ratio test [10–13], after rewriting the ordinal penalty as a mixed model [14–16]. However, the corresponding test is not available for generalized linear models, such as the logit model considered here. In this note, we will illustrate how technology developed for generalized additive models [17, 18] can be used to fit generalized linear and additive models with ordinal smoothing penalty, and conduct further statistical inference.

Main text

Methods

Given a response y with distribution from a simple exponential family, and a set of covariates x_1, \dots, x_p , a generalized additive model [17] has the form:

$$\eta = \alpha + f_1(x_1) + \dots + f_p(x_p), \quad \mu = h(\eta), \tag{1}$$

where μ is the (conditional) mean of y given the covariates, h is a (known) response function, and η is comparable to the linear predictor in generalized linear models [19, 20]. The difference to a generalized linear model is that non-linear functions f_j , $j = 1, \dots, p$, are allowed in η , but still the structure of η is additive. Of course, if f_j are restricted to be linear, a generalized linear model is obtained as a special case. In a (generalized) additive model, however, it is usually only assumed that

functions f_j are reasonably smooth; and one way to fit such models, as for instance implemented in the popular R package `mgcv` [18, 21], is to specify a set of basis functions for each predictor and to employ an appropriate, quadratic smoothing penalty on the corresponding basis coefficients. That means, we assume that

$$f_j(x) = \sum_{r=1}^{q_j} \beta_{jr} B_{jr}(x), \tag{2}$$

with $B_{j1}(x), \dots, B_{jq_j}(x)$ being a reasonably rich set of basis functions chosen for function f_j , and $\beta_{j1}, \dots, \beta_{jq_j}$ are the corresponding basis coefficients. When fitting those basis coefficients to the data, a penalty term $J_j(\beta_j)$ is typically added for each covariate x_j , penalizing wiggly basis coefficients and thus wiggly functions f_j . The strength of the penalty and hence the amount of smoothing is controlled through a tuning parameter, often denoted by λ_j .

Now suppose you have a categorical predictor x_j with levels $1, \dots, k_j$. Then, there is a somewhat natural basis: the basis of (dummy) functions ($l = 1, \dots, k_j$)

$$B_{jl}(x_j) = \begin{cases} 1 & \text{if } x_j = l, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Since we know that x_j can only take values $1, \dots, k_j$, we do not need to think about the type and number of basis functions, placing of knots, etc., as we usually do with continuous covariates. If now a (quadratic) first-order difference penalty

$$J_j(\beta_j) = \sum_{l=2}^{k_j} (\beta_{jl} - \beta_{j,l-1})^2, \tag{4}$$

is put on the basis/dummy coefficients $\beta_j = (\beta_{j1}, \dots, \beta_{jk_j})^\top$, this gives exactly the smoothing penalty as mentioned above [4–7]. Alternatively, the second-order penalty

$$\begin{aligned} J_j(\beta_j) &= \sum_{l=2}^{k_j-1} ((\beta_{j,l+1} - \beta_{jl}) - (\beta_{jl} - \beta_{j,l-1}))^2 \\ &= \sum_{l=2}^{k_j-1} (\beta_{j,l+1} - 2\beta_{jl} + \beta_{j,l-1})^2, \end{aligned} \tag{5}$$

can be used [14]. One of the benefits of considering ordinal predictors along with quadratic difference penalties in the framework of generalized additive models is that after implementing basis (3) in the appropriate way, `gam()` from `mgcv` can be used directly to fit a generalized linear/additive model with ordinal predictor(s) as needed for the BPD data. Besides pure model fitting, however, this provides us with additional tools; in

particular, built-in estimation of the penalty/smoothing parameter(s) via (restricted) maximum likelihood ((RE) ML), further statistical inference, such as confidence intervals, and checking significance of smooth terms. Those tools utilize the mixed model and Bayesian interpretation of quadratic smoothing penalties on basis coefficients such as (4) and (5); compare [18, 22, 23] for details.

Add-on functions implementing the ordinal basis for use within mgcv have been made publicly available through R package ordPens [24]. After installing and loading ordPens, the gam() function from mgcv can be used with smooth terms s(..., bs = "ordinal", m = 1) or s(..., bs = "ordinal", m = 2) for the first- and second-order penalty, respectively. See the ordPens manual (R function ordSmooth()) for details and examples. To investigate whether the p-values of Wald-type tests with respect to smooth terms as provided by summary.gam() are reliable if using the ordinal basis/smoothing penalty, we used the confounder model, i.e., the model with information on bacteria removed, to estimate BPD probabilities. That means, the null hypothesis that the effect of (ordinal, bacteria-specific predictor) x is zero, is true by construction in this hypothetical model, because fitted BPD probabilities do not depend on x , given the other covariates. Using those probabilities, we simulated 'new' BPD response data, fit the model with smooth ordinal x added (and smoothing parameter estimated by REML), and stored the p-value of x . For x , we used information on gram negative/positive and pathogenic bacteria, respectively. As

noted above, the corresponding ordinal factor gives the week colonization by the respective type of (oral) bacteria was detected for the first time. For each type of x , this was repeated 1000 times.

Results

Figure 1 shows QQ-plots of the p-values observed on the simulated data employing the first- and second-order penalty, respectively. Since the distribution of smaller p-values is particularly relevant when testing with usual $\alpha \leq 0.1$, we restrict plotting to that area. It is seen that p-values obtained when employing the first-order penalty (red) are typically too small. Problems with the first-order penalty can be explained by the fact that the null space of the corresponding smooth term has dimension zero (compare the mgcv manual). In other words, the null hypothesis in the framework of mixed models (which is used for estimation here), a zero variance component, is on the boundary of the parameter space, which means that standard theory does not apply [10, 11]. Results for the second-order penalty (blue), by contrast, look very encouraging. Consequently, we will only report results on the actually observed BPD data for the second-order penalty below. In earlier analyses [2], separate models were fit for each type of bacteria, and the first two weeks were collapsed to make (unpenalized) model fitting with dummy-coded, ordinal factors feasible. Thanks to the penalties presented here, we are now able to include all three predictors jointly while using all information with the resolution available.

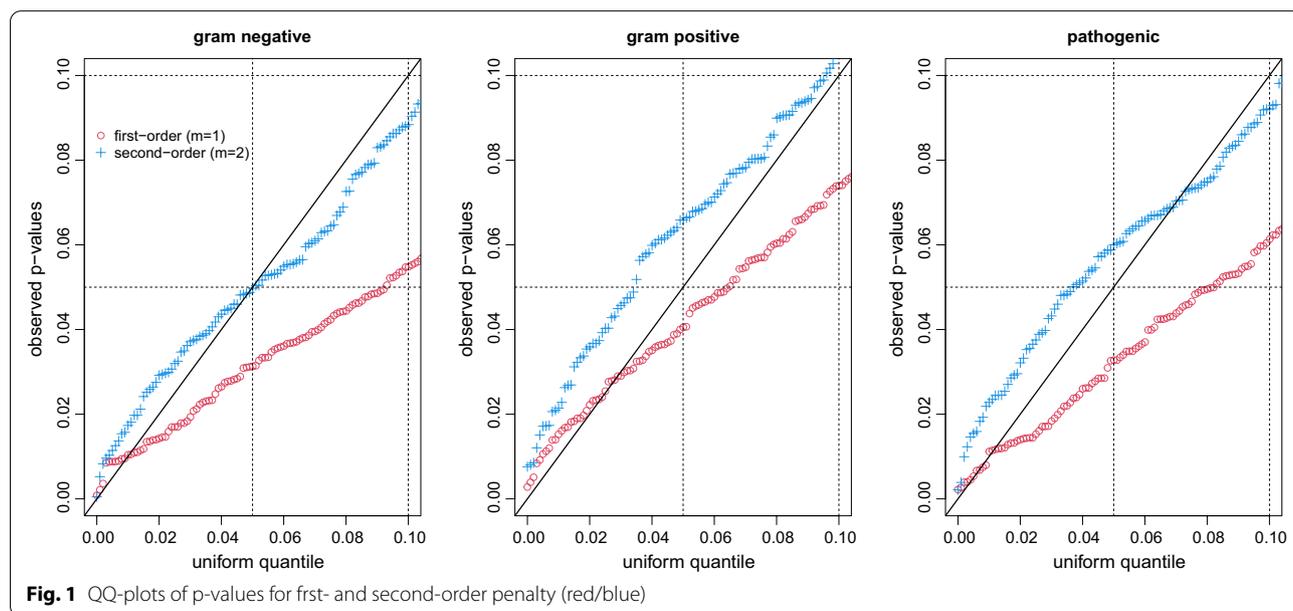


Table 1 Results for parametric and smooth terms in the full and reduced model when using the second-order ordinal smoothing penalty

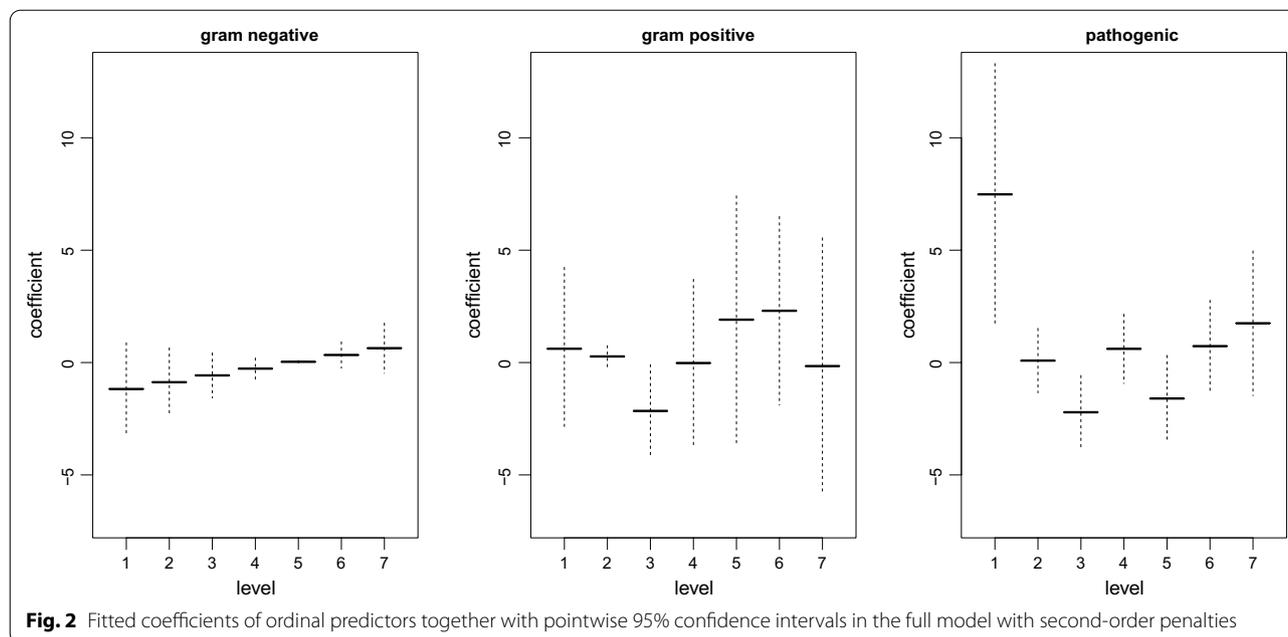
Full model				
Parametric terms				
Covariate	Estimate	Std. error	z-value	p-value
(Intercept)	6.214	2.630	2.363	0.018
Weight (g)	− 0.013	0.004	− 3.381	<0.001
SGA sym.	1.909	1.359	1.405	0.160
Sex (male)	3.022	1.114	2.712	0.007
Multiples	1.524	0.744	2.048	0.041
Steroids	− 0.241	0.090	− 2.684	0.007
Antibiotics	0.079	0.090	0.874	0.382
Smooth terms				
Predictor	edf	Ref.df	Chi.sq	p-value
Gram negative	1.000	1.000	1.307	0.253
Gram positive	3.708	4.393	5.227	0.264
Pathogenic	5.258	5.831	13.711	0.030
Reduced model				
Parametric terms				
Covariate	Estimate	Std. error	z-value	p-value
(Intercept)	6.426	2.170	2.961	0.003
Weight (g)	− 0.012	0.003	− 3.859	< 0.001
SGA sym.	1.991	1.290	1.544	0.123
Sex (male)	2.107	0.834	2.527	0.012
Multiples	1.054	0.528	1.995	0.046
Steroids	− 0.174	0.072	− 2.432	0.015
Antibiotics	0.079	0.078	1.013	0.311
Smooth terms				
Predictor	edf	Ref.df	Chi.sq	p-value
Gram negative	−	−	−	−
Gram positive	−	−	−	−
Pathogenic	4.973	5.696	13.573	0.027

Table 1 (top) shows the results for the parametric terms if using the second-order penalty (5) for the smooth terms. In particular, it is seen/confirmed that low birth weight is a risk factor for BPD, and also male infants and multiples have an increased risk of developing BPD. Antenatal steroids, by contrast, may decrease the risk. Results for the different types of bacterial colonization, which are included as ordinal predictors with smooth effects, are also given in Table 1 and Fig. 2. We see that the only significant effect is detected for pathogenic bacteria. The fitted function (Fig. 2, right) gives the impression that early detection is associated with increased risk of BPD. Statistical uncertainty, however, is very large (due to the small number of samples with week/level 1) as indicated by the confidence interval. When excluding

information on gram negative and positive bacteria from the model, results for the remaining terms (Table 1, bottom) as well as fitted functions/coefficients for pathogenic bacteria (not shown) look very similar as before. In summary, our results using the ordinal smoothing approach are in line with earlier analyses [2], but allow for considering all three ordinal predictors (gram negative/positive, pathogenic bacteria) jointly, without the need to collapse levels.

Limitations

With respect to the application/BPD data, the main limitation is the small number of samples in week 1 for pathogenic bacteria. Since the shape of the function in Fig. 2 (right) depends on the coefficient for week/level



1, this shape should not be over-interpreted here. From a technical point of view, if using the second-order penalty (5), a problem can occur with confidence intervals in terms of under-coverage if the fitted coefficient function is close to being linear (compare Fig. 2, left). This problem is also found for (generalized) additive models with continuous covariates, and the suggested fix is to change the target of inference to the smooth term plus the overall model intercept [18, 25]. Furthermore, our implementation does not include extensions like ordinal smoothing spline *isotonic* regression [7]. Finally, it should be noted that all statements made and conclusions drawn in this article refer to statistical inference if smoothing parameters are estimated by REML (or ML). When using GCV (which is the default in `mgcv!`), results should be treated with caution.

Abbreviations

BPD: Bronchopulmonary Dysplasia; (RE)ML: (Restricted) maximum likelihood; GCV: Generalized cross-validation.

Acknowledgements

The authors would like to thank the Editor and an anonymous reviewer for their constructive feedback that helped us to improve the paper.

Author's contributions

JG conducted the data analysis/numerical experiments and wrote the manuscript. FS implemented the method for use within `mgcv`. TL and HE collected the data and were involved in the discussion of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported in part by Deutsche Forschungsgemeinschaft (DFG) under Grant GE2353/2-1.

Availability of data and materials

The software presented together with examples is publicly available on CRAN through open source R add-on package `ordPens` [24]. The data [2] that support the findings of this study are available on reasonable request from H.E. (harald.ehrhardt@paediat.med.uni-giessen.de). An earlier yet extended version of this article providing further technical details and simulation studies is available at <https://arxiv.org/abs/2102.01946>.

Declarations

Ethics approval and consent to participate

The results presented only refer to a secondary data analysis. The original study [2] had been conducted following the rules of the Declaration of Helsinki of 1975, revised in 2013. The retrospective analysis was approved by the ethics committee of the Justus-Liebig-University Giessen (Az 97/14).

Consent for publication

Not applicable (secondary data analysis).

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Economics and Social Sciences, Helmut Schmidt University/University of the Federal Armed Forces, Hamburg, Germany. ²Department of Statistics, Ludwig Maximilians University, Munich, Germany. ³Department of General Pediatrics and Neonatology, Justus Liebig University, Giessen, Germany. ⁴German Center for Lung Research (DZL), Universities of Giessen and Marburg Lung Center (UGMLC), Giessen, Germany.

Received: 18 January 2022 Accepted: 8 March 2022

Published online: 22 March 2022

References

- Gronbach J, Shahzad T, Radajewski S, Chao C-M, Bellusci S, Morty RE, Reichertzer T, Ehrhardt H. The potentials and caveats of mesenchymal stromal cell-based therapies in the preterm infant. *Stem Cells Int.* 2018;2018:9652897.
- Lauer T, Behnke J, Oehmke F, Bäcker J, Gentil K, Chakraborty T, Schlöter M, Gertheiss J, Ehrhardt H. Bacterial colonization within the first six weeks of life and pulmonary outcome in preterm infants < 1000g. *J Clin Med.* 2020;9:2240.
- Hoerl AE, Kennard RW. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics.* 1970;12:55–67.
- Gertheiss J, Tutz G. Penalized regression with ordinal predictors. *Int Statist Rev.* 2009;77:345–65.
- Tutz G, Gertheiss J. Rating scales as predictors—the old question of scale level and some answers. *Psychometrika.* 2014;79:357–736.
- Tutz G, Gertheiss J. Regularized regression for categorical data (with discussion). *Statis Model.* 2016;16:161–200.
- Helwig NH. Regression with ordered predictors via ordinal smoothing splines. *Front Appl Math Statist.* 2017;3:15.
- Cieza A, Oberhauser C, Bickenbach J, Chatterji S, Stucki G. Towards a minimal generic set of domains of functioning and health. *BMC Public Health.* 2014;14:218.
- Glass SM, Ross SE. Modified functional movement screening as a predictor of tactical performance potential in recreationally active adults. *Int J Sports Phys Ther.* 2015;10:612–21.
- Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. *J R Statist Soc B.* 2004;66:165–85.
- Crainiceanu CM, Ruppert D, Claeskens G, Wand MP. Exact likelihood ratio tests for penalized splines. *Biometrika.* 2005;77:91–103.
- Greven S, Crainiceanu CM, Küchenhoff H, Peters A. Restricted likelihood ratio testing for zero variance components in linear mixed models. *J Comput Graph Statist.* 2008;17:870–91.
- Scheipl F, Greven S, Küchenhoff H. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput Statist Data Anal.* 2008;52:3283–99.
- Gertheiss J, Oehrlin F. Testing relevance and linearity of ordinal predictors. *Electron J Statist.* 2011;5:1935–59.
- Gertheiss J. Anova for factors with ordered levels. *J Agricult Biol Environ Statist.* 2014;19:258–77.
- Sweeney E, Crainiceanu C, Gertheiss J. Testing differentially expressed genes in dose–response studies and with ordinal phenotypes. *Statis Appl Genet Mol Biol.* 2016;15:213–35.
- Hastie T, Tibshirani R. Generalized additive models. London: Chapman & Hall; 1990.
- Wood SN. Generalized additive models: an introduction with R. 2nd ed. Boca Raton: CRC Press; 2017.
- Nelder JA, Wedderburn RWM. Generalized linear models. *J R Statist Soc A.* 1972;135:370–84.
- McCullagh P, Nelder JA. Generalized linear models. 2nd ed. New York: Chapman & Hall; 1989.
- R Core Team: R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Statist Soc B.* 2011;73:3–36.
- Wood SN. On p-values for smooth components of an extended generalized additive model. *Biometrika.* 2013;100:221–8.
- Gertheiss, J., Hoshiyar, A.: ordPens: selection, fusion, smoothing and principal components analysis for ordinal variables. (2021). R package version 1.0.0. <https://CRAN.R-project.org/package=ordPens>.
- Marra G, Wood SN. Coverage properties of confidence intervals for generalized additive model components. *Scand J Statist.* 2012;39:53–74.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

