The Journal of Open Source Software

ordPens: An R package for Selection, Smoothing and Principal Components Analysis for Ordinal Variables

Aisouda Hoshiyar[1]

1 School of Economics and Social Sciences, Helmut Schmidt University, Hamburg, Germany

DOI: 10.21105/joss.03828

Software

- Review
- Repository
- Archive

Editor: Øystein Sørensen

Reviewers:

- @fartist
- @FranjoIM
- @mingzehuang

Submitted: 12 October 2021
Published: 06 December 2021

Summary

Ordinal data are a common case in applied statistics. In order to incorporate the ordinal scale level, among other things, regularization techniques are often suggested in the literature (Tutz & Gertheiss, 2014, 2016). In particular, penalization approaches for smoothing and selection when dealing with Likert-type data – which are by no means restricted to Likert scale – are commonly proposed. `ordPens` is a package in the R programming language (R Core Team, 2021) and provides several penalty approaches for ordinal predictors in regression models and ordinal variables for principal component analysis (PCA).
In the regression context, smoothing is obtained by introducing a penalty term and a tuning parameter controling the amount of penalty. Adding the penalty term to the likelihood function yields the penalized likelihood, which is then maximized. Different types of penalization can be considered, depending on whether to achieve smoothing, selection or clustering of variables. Smoothing only can be done by penalizing the sum of squared differences of adjacent coefficients for a given variable, subject to proper ordering. A modified group lasso based on a difference penalty can be used for selection. Clustering/fusion of categories can be achieved by the fused lasso penalizing absolute differences by using the L_1-norm.

Statement of Need

As suggested by Tutz & Gertheiss (2014) and Tutz & Gertheiss (2016), selection, and/or smoothing/fusing of ordinally scaled independent variables shall be done using a modified group lasso or generalized ridge penalty when dealing with ordinally scaled predictors in regression analysis. The penalized log-likelihood to be maximized takes the form $\boldsymbol{l}_p(\boldsymbol{\beta}) = \boldsymbol{l}(\boldsymbol{\beta}) - \lambda \boldsymbol{J}(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ corresponding to the vector of regression parameters, λ representing the smoothing parameter and $\boldsymbol{J}(\cdot)$ being the penalty function.
`ordPens` (Gertheiss & Hoshiyar, 2021) offers various tools for data analysis of ordinally scaled data. The package attacks the afore mentioned tasks and offers penalized regression for smoothing, selection and fusion. Specifically, the function `ordSmooth()` for smoothing only incorporates the generalized ridge penalty

$$\boldsymbol{J}(\boldsymbol{\beta}) = \sum_{s=1}^{p} \boldsymbol{\beta}_s^T \boldsymbol{D}_{d,s}^T \boldsymbol{D}_{d,s} \boldsymbol{\beta}_s,$$

with $\boldsymbol{D}_{d,s}$ being the matrix generating differences of order d and $\boldsymbol{\beta}_s^T = (\beta_{s1}, ..., \beta_{sk_s})$ being the parameter vector linked to the sth (dummy-coded) predictor with categories $1, ..., k_s$. The `ordSelect()` function performs smoothing and selection by adopting a modified group lasso penalty based on differences of the form

$$\boldsymbol{J}(\boldsymbol{\beta}) = \sum_{s=1}^{p} \sqrt{k_s} \sqrt{\boldsymbol{\beta}_s^T \boldsymbol{D}_{d,s}^T \boldsymbol{D}_{d,s} \boldsymbol{\beta}_s}.$$

Clustering of categories is done by the function `ordFusion()`, which uses a fused lasso penalty based on differences of first order:

$$\boldsymbol{J}(\boldsymbol{\beta}) = \sum_{s=1}^{p} \sum_{j=2}^{k_s} |\beta_{sj} - \beta_{s,j-1}|.$$

For more information on the original group lasso (for nominal predictors and grouped variables in general), see Meier et al. (2008) and Yuan & Lin (2006). For details on the fused lasso, see Tibshirani et al. (2005). In the case of smoothing only, the package includes auxiliary functions such that `mgcv::gam()` (Wood, 2008, 2017) can be used for fitting generalized linear and additive models with first- and second-order ordinal smoothing penalty as well as built-in smoothing parameter selection. Also, `mgcv` tools for further statistical inference can be used, see Gertheiss et al. (2021) for details. Furthermore, testing for differences in the means, known as analysis of variance (ANOVA), is provided for ordered factors by the function `ordAOV()` penalizing (squared) differences of adjacent means. Testing for differentially expressed genes, when analyzing microarrays of gene expression data, is incorporated by the function `ordGene()`. Technical details can be viewed from Gertheiss (2014) and Sweeney et al. (2016), respectively.

If, in contrast, dimension reduction is desired in an unsupervised way, principal components analysis can be applied to ordinal data as well. However, those data are usually either treated as numeric implying linear relationships between the variables at hand, or non-linear PCA is applied where the obtained coefficients are sometimes hard to interpret. Note that in IBM SPSS Statistics (Version 25.0), for instance, there is an option available for smoothing quantifications by the use of spline functions, which, however, limits the type of functions that can be fitted when using a small number of knots and a suitable choice may be challenging for the (inexperienced) user. On the other hand, as splines are defined on interval scale whereas ordinal variables can only take some discrete values, the usage of spline functions may be seen as unnecessarily complex for scaling ordinal data. To incorporate the ordinal scale level, the concept of penalization can also be adapted here, as suggested in Hoshiyar et al. (2021). Penalized non-linear principal components analysis for ordinal variables is incorporated in the function `ordPCA()` using a second-order difference penalty. In addition, the function provides performance evaluation and selection of an optimal penalty parameter using k-fold cross-validation. Also, the option of both non-monotone effects and incorporating constraints enforcing monotonicity is provided. Penalized non-linear PCA therefore serves as an intermediate between the standard methods typically used so far (see above). The new approach offers both better interpretability as well as better performance on validation data.

A topic of future research would be the analysis of dependencies within a (high dimensional) set of ordinal variables by graphical models. A further typical approach when dealing with ordinal data is motivated by assuming a latent continuous variable linked to the ordinal variable via thresholds. The proportional odds model, which is also motivated as a latent variable approach, in combination with the ordinal penalty could be also of interest for future research. Another interesting field is found in Huang et al. (2021), who analyze (mixed) ordinal dependencies using a latent Gaussian copula model based on rank correlations. Assuming a latent continuous variable, however, may not always be desirable by the data analyst. The methods implemented in `ordPens` (up to version 1.0.0) therefore do not underly the latent variable assumption.

Availability

The R package `ordPens` is publicly available on CRAN and Github, where issues can be opened. `ordPens` is licensed under the GPL-2 General Public License. Documentation and examples are contained in the package manual, which can be found on CRAN.

To install ordPens, simply run:

```
install.packages("ordPens")
```

For penalized regression and ordinal ANOVA see also `vignette("ordPens", package = "ordPens")`. Penalized non-linear PCA is also documented in detail and can be accessed via `vignette("ordPCA", package = "ordPens")`.

ordPens in action

This example illustrates penalized non-linear PCA on the so-called brief ICF core set on Chronic Widespread Pain (CWP) consisting of 26 ordinally scaled variables. Details on the data can be found in Gertheiss et al. (2011); or by typing `?ICFCoreSetCWP`. Analysis of the so-called comprehensive core set for CWP, consisting of 67 ICF variables, is found in Hoshiyar et al. (2021). Note that the ordPCA vignette also analyzes the comprehensive core set. Figure 1, generated by the following code, illustrates the estimated coefficients of selected variables for different values of the penalty parameter λ along with cross-validation results.

```
library(ordPens)

# load ICF data & code adequately
data(ICFCoreSetCWP)
H <- ICFCoreSetCWP[,1:67] + matrix(c(rep(1,50), rep(5,16), 1),
                                   nrow(ICFCoreSetCWP), 67, byrow = TRUE)

# select brief core set variables
brief <- c("b130","b134","b140","b147","b152","b1602","b280","b455",
           "b730","b760","d175","d230","d240","d430","d450","d640",
           "d760","d770","d850","d920","e1101","e310","e355","e410",
           "e420","e570")

H <- H[, brief]
xnames <- names(H)

# ordinal penalized PCA
icf_pca1 <- ordPCA(H, p = 2, lambda = c(5, 0.5, 0.001), qstart = NULL,
                   crit = 1e-7, maxit = 100, Ks = c(rep(5,20),rep(9,6)),
                   constr = c(rep(TRUE,20), rep(FALSE,6)), CV = FALSE,
                   k = 5)

# 5-fold cross-validation
lambda <- 10^seq(4, -4, by = -0.1)
set.seed(1234)
cvResult <- ordPCA(H, p = 2, lambda = lambda, Ks = c(rep(5,20),rep(9,6)),
                   constr = c(rep(TRUE,20), rep(FALSE,6)),
                   CV = TRUE, k = 5, CVfit = FALSE)

# plotting results for selected variables
par(mfrow = c(2,3))
for(i in which(xnames=="b280"|xnames=="d450"|xnames=="e1101"|
               xnames=="e410")){
```

```
  plot(icf_pca1$qs[[i]][,3], type = "b", xlab = "category", col = 1,
       ylab = "quantification", main = xnames[i], bty = "n", xaxt = "n",
       ylim = range(icf_pca1$qs[[i]]))
  lines(icf_pca1$qs[[i]][,2], type="b", col=2, lty=2, pch=2, lwd=2)
  lines(icf_pca1$qs[[i]][,1], type="b", col=3, lty=3, pch=3, lwd=2)
  axis(1, at = 1:length(icf_pca1$qs[[i]][,1]))
}

plot(log10(lambda), apply(cvResult$VAFtrain, 2, mean), type = "l",
     xlab = expression(log[10](lambda)), ylab = "VAF", cex.axis = 1.2,
     main = "training data", cex.lab = 1.2)

plot(log10(lambda), apply(cvResult$VAFtest, 2, mean), type = "l",
     xlab = expression(log[10](lambda)), ylab = "VAF", cex.axis = 1.2,
     main = "validation data", cex.lab = 1.2)
abline(v = log10(lambda)[which.max(apply(cvResult$VAFtest,2,mean))],
       lty=2)
```

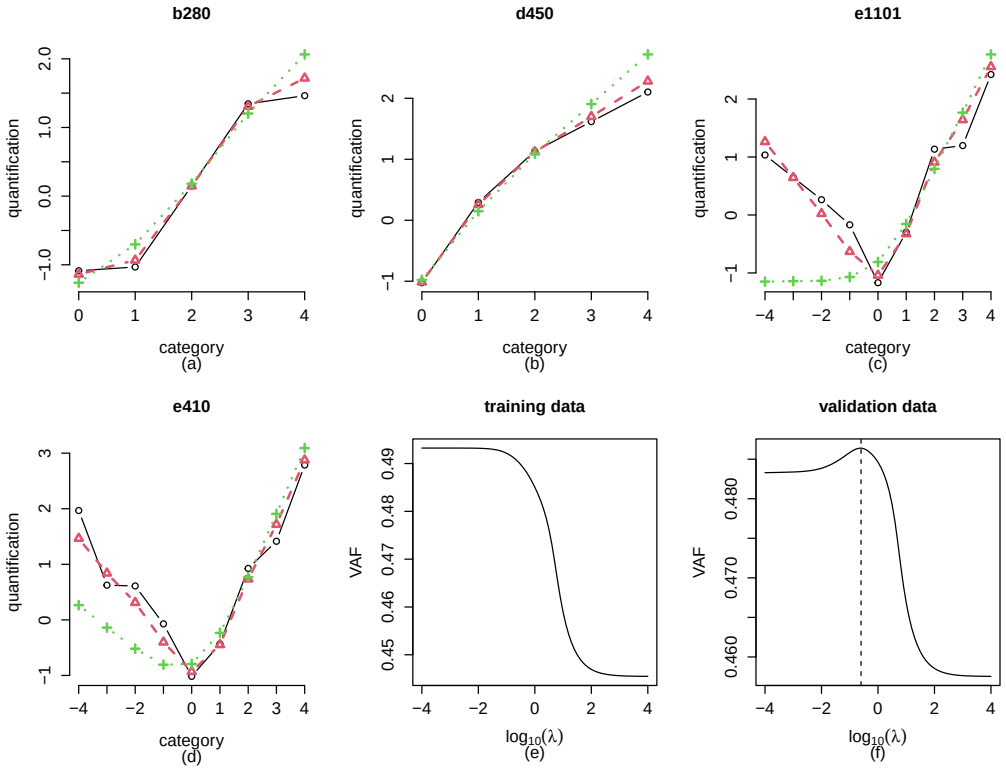

Figure 1: Category quantifications/scores for $\lambda \to 0$ (solid black), $\lambda = 0.5$ (dashed red), $\lambda = 5$ (dotted green) (a)–(d); VAF by the first 5 principal components: (e) training data, (f) validation data with optimal λ (dashed line).

Acknowledgements

This work was supported in part by Deutsche Forschungsgemeinschaft (DFG) under Grant GE2353/2-1.

I thank Jan Gertheiss and Fabian Scheipl for their contributions to the software package. Jan Gertheiss created initial package versions 0.1-1 up to 0.3-1 and helped discussing the manuscript. Fabian Scheipl implemented the ordinal smoothing penalty for use within `mgcv`.

References

Gertheiss, J. (2014). ANOVA for factors with ordered levels. *Journal of Agricultural, Biological, and Environmental Statistics*, *19*(2), 258–277. https://doi.org/10.1007/s13253-014-0170-5

Gertheiss, J., Hogger, S., Oberhauser, C., & Tutz, G. (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society C*, *60*, 377–395. https://doi.org/10.1111/j.1467-9876.2010.00753.x

Gertheiss, J., & Hoshiyar, A. (2021). *ordPens: Selection, fusion, smoothing and principal components analysis for ordinal variables*. https://CRAN.R-project.org/package=ordPens

Gertheiss, J., Scheipl, F., Lauer, T., & Ehrhardt, H. (2021). *Statistical inference for ordinal predictors in generalized linear and additive models with application to bronchopulmonary dysplasia*. https://arxiv.org/abs/2102.01946

Hoshiyar, A., Kiers, H. A. L., & Gertheiss, J. (2021). *Penalized non-linear principal components analysis for ordinal variables with an application to international classification of functioning core sets*. https://arxiv.org/abs/2110.02805

Huang, M., Müller, C. L., & Gaynanova, I. (2021). Latentcor: An R package for estimating latent correlations from mixed data types. *Journal of Open Source Software*, *6*(65), 3634. https://doi.org/10.21105/joss.03634

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 53–71. https://doi.org/10.1111/j.1467-9868.2007.00627.x

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sweeney, E., Crainiceanu, C., & Gertheiss, J. (2016). Testing differentially expressed genes in dose-response studies and with ordinal phenotypes. *Statistical Applications in Genetics and Molecular Biology*, *15*(3), 213–235. https://doi.org/10.1515/sagmb-2015-0091

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

Tutz, G., & Gertheiss, J. (2014). Rating scales as predictors - the old question of scale level and some answers. *Psychometrika*, *79*(3), 357–376. https://doi.org/10.1007/S11336-013-9343-3

Tutz, G., & Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, *16*(3), 161–200. https://doi.org/10.1177/1471082X16642560

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society B*, *70*, 495–518. https://doi.org/10.1111/j.1467-9868.2007.00646.x

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press. ISBN: 9781498728331

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x