Statistical Modelling 2016; **16(3)**: 249–260

Rejoinder: Regularized regression for categorical data

Gerhard Tutz[1] and Jan Gertheiss[2]
[1]Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany
[2]Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

1 Introduction

First of all, we want to thank all the discussants for their very thoughtful comments, additional illustrations, simulation studies, data analyses and much more. We agree with Alan Agresti that regularization for categorical data is in an early stage and still much has to be done. In particular, we confined ourselves to regression modelling. Other areas as, for example, measurement of association and analysis of contingency tables might call for quite different solutions. We are happy to see that the article stimulated some research and the discussants' contributions brought some progress and new ideas. Once again, we want to thank the discussants for their inspiring and encouraging comments. In this reply, we will only address some of the many points that were raised. We will start with a joint response to the comments made by Alan Agresti and Shepard/Liu, as they partly discussed similar issues.

2 Discussion by A. Agresti, B. Shepherd and Q. Liu

Alan Agresti argues that simple ordinal response models are very useful to obtain information on first-order effects. This is certainly true, and in cumulative models one might consider category-specific effects as secondary effects that mainly describe departures from the overall effects. However, for sequential models, which are an important class of ordinal response models, category-specific effects signal that transitions between categories, given a specific category has been reached, vary across categories. If categories refer to time, one models time-varying effects of covariates, which might give a much more detailed and more appropriate picture than a model that assumes that the transition between categories has the same strength everywhere. Then effects

Address for correspondence: Gerhard Tutz, Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany.
E-mail: gerhard.tutz@stat.uni-muenchen.de

 10.1177/1471082X16652780

have a simple interpretation and one does not have to sacrifice optimality of fit for ease of interpretation, as mentioned by Bryan Shepherd and Qi Liu. In general, we think that, if enough data is available, it is a reasonable way to start with a more complex model in combination with a penalty, and let the data decide which model to choose.

Several strategies can be used to reduce the complexity of ordinal predictors. Monotone scores can be assigned to the categories but have the drawback that it is not obvious how to choose scores and, moreover, one uses a scale level that is not supported by the data. The chosen scores are typically assumed to be measurements on an interval scale. In research communities like psychology that are more sensitive to scale levels, the underlying assumption might be considered as rather strong. Nevertheless, scores are frequently used with good reasons. Also Bryan Shepherd and Qi Liu seem to prefer the assignment of scores to categories. They propose to assign the numbers 1 to k and then fit splines to reduce the number of parameters. If the objective is to just obtain smooth effects, the resulting curves will typically be very similar to the ones obtained by penalizing squared distances of adjacent parameters because the latter is just a special case of first-order P-splines, as we discussed in Section 3.2.1 of the main article. As with higher order splines and the usual roughness penalty on the second derivative, we can also use penalties on dummy coefficients penalizing deviations from linearity (Gertheiss and Oehrlein, 2011). Using this approach, we might even say that we let the data tell us which scores to use in a linear model: are the numbers $1, 2, \ldots$ a good choice or do we need some other quantification? If, on the contrary, the objective is not smoothing but one wants to investigate if categories differ in their effect on the response, the assignment of scores is rather damaging. Thus, it depends on the objective of the data analysis if the assignment of scores is sensible.

An alternative strategy mentioned by Alan Agresti and others is to impose an ordering constraint on category parameters. This is certainly attractive if monotonicity of effects is known. However, as noted by Shepherd and Liu, the results using isotonic regression will generally be different from the results when using difference penalties. In particular, difference penalties can in some sense exploit the monotonicity, too, but they do not fail if this assumption is wrong. From our point of view, although Shepard/Liu raise doubts, difference penalties work also if relationships are not monotone, because they do not assume monotonicity. For instance, Sweeney *et al.* (2016) compare the mixed model-based restricted likelihood ratio test (RLRT) using the ordinal smoothing penalty (see Gertheiss, 2014 and Section 3.2.2 of the main article) to various tests assuming monotonicity (Lin *et al.*, 2007, 2014; Pramana *et al.*, 2010) or linearity, and standard ANOVA. In summary, they found that 'in the cases of linear and monotone functions, the tests which make these assumptions may perform better than RLRT, but usually not by a large margin. On the other hand, when these assumptions are not true, RLRT is distinctly better. Furthermore, the only other test (considered) not making structural assumptions, standard ANOVA, is mostly outperformed by RLRT.'

Of course, estimated parameters will not only differ between isotonic regression and penalty-based approaches, but also within penalty methods, such as quadratic and L_1 difference penalties. This is nicely seen, for example, from the Wisconsin

breast cancer dataset considered in the comment by Chiquet *et al.,* and discussed in Section 4 of this Rejoinder. With this data, the assumption of monotonicity appears very reasonable, and in particular the ordinal fusion penalty produces very interesting results by clustering a lot of predictor levels; see Section 4 for details. Concerning the comparison to isotonic regression, instead of using the pool adjacent violators algorithm, one might use methods that allow to include various explanatory variables in a flexible form in combination with regularization (see, for example, Leitenstorfer and Tutz, 2007; Tutz and Leitenstorfer, 2007).

Shepherd and Liu also investigated another interesting concept, the use of difference penalties for continuous data. They found that the relationship between predictor and response was well captured, but at the cost of too many parameters. Nevertheless, it might be useful if one has a discrete but metrically scaled predictor instead of a truly continuous, for example, normally distributed predictor. In that case, we can even use the information available on distances on the predictor scale, by employing weights within the penalty that (inversely) depend on the distance of two parameters. Furthermore, the number of parameters is automatically reduced when using L_1 penalties. Since typically many parameters will be fused, a step function will be obtained; compare, for example, the so-called fused lasso signal approximator (Hoefling, 2010).

Shepherd and Liu also comment on the regularized fixed effects model. It is certainly legitimate to believe that the distribution of random effects is strictly continuous, and therefore all random effects are at least slightly different. One strength of the regularized fixed effects model is that it automatically identifies clusters of random effects that can be considered as identical; however, the more important strength is that the model is not based on the assumption that explanatory variables and random effects are uncorrelated. Although this assumption may hold in many biometrical applications, it is often doubtful in the social sciences.

3 Discussion by P. Bühlmann and R. Dezeure

Peter Bühlmann and Ruben Dezeure discussed inference tools for high-dimensional categorical covariates that go beyond simple point estimates. Those tools, in particular confidence intervals and p-values, have not been considered in detail in our article. Therefore, the comment of Bühlmann and Dezeure is very important and highly welcome. Bühlmann and Dezeure adapt recent results obtained for high-dimensional generalized regression to categorical predictors with lasso-type penalties. In particular, they consider a setting with nominal covariates and the usual dummy coding with reference category and use the R package `hdi` (Meier *et al.*, 2014; Dezeure *et al.*, 2015) to obtain confidence intervals for the dummy coefficients. A simulation study nicely shows that this approach indeed allows to draw conclusions about the underlying model structure.

An important point to discuss here is how this strategy can be used for ordinal covariates. (a) If covariates are ordinal, a slight but useful modification could be

not to use the common dummy coding, but split coding which explicitly focuses on differences of adjacent parameters (see also the response to the comment by Chiquet *et al.* below). Employing the approach of Bühlmann and Dezeure on the recoded design matrix would directly yield confidence intervals for those differences, which are typically the interesting ones with ordinal predictors (see also Walter *et al.*, 1987). We also want to mention here again that also the quadratic smoothing penalty can be used for inference, at least if the number of predictors involved is not too large. The key is to reformulate the penalty within a mixed models framework, compare Section 3.2 of the main article. Once this has been done, the entire mixed models methodology for statistical inference is available, including statistical tests and confidence intervals. For instance, the null-hypothesis known from ANOVA that the expected response does not differ between factor levels can be tested (see Gertheiss, 2014 for details). (b) With ordinal predictors, as discussed in the main article, there are various types of penalties available, such as smoothing only, smoothing and selection or fusion, and the researcher may wonder which one to choose (see also the comment by Chenlei Leng). Bühlmann and Dezeure argue that by 'visual inspection of confidence intervals, one could determine for which of the ordinal categorical variables some smoothing or clustering of categories would be expected to be beneficial'. This seems to be an exploratory way if one has no idea at all what penalty to use. However, we want to emphasize that from our point of view, the choice of the penalty should depend on the specific nature of the application and, most importantly, on the objective of the data analysis. For instance, there are cases where it does not make sense to assume that the predictor's influence on the response can be described by a step function (see also our response to the comment by Chenlei Leng below). Typically, however, this cannot be decided by statistical methods only, or the statistician alone, but together with collaborators who are familiar with the specific field of application and can tell us which structure behind the data they want to investigate.

Another important point discussed by Bühlmann and Dezeure is the use of tree-based approaches, such as random forests (Breiman, 2001), for categorical covariates. Compared to the parametric models discussed in our article, trees have the advantage that different kinds of covariates, such as metric and categorical ones, can be mixed without the need to use different penalty parameters or think about appropriate weights. Moreover, potential interactions are taken into account. A potential drawback of simple trees might be that they tend to be unstable against small variations in the data and the prediction accuracy is typically not very good. A much better choice are random forests, which are often among the most powerful methods for prediction. However, compared to parametric statistical models, they rather act like a 'black box'. Therefore, results are often hard to interpret, and they should only be used when the main focus is on prediction (compare also Bühlmann and Dezeure).

For comparison with penalty-based methods, we consider the food data and the model for ordinal predictors with smoothing and selection penalty, see Figure 1 in the main article. Figure 1 (left) in this Rejoinder shows the fitted values when using the penalty approach versus results obtained with the R package `randomForest` (Liaw and Wiener, 2002) on the entire dataset. It is seen that most values are very similar. To investigate prediction accuracy, we randomly draw a test set of 100 observations; on

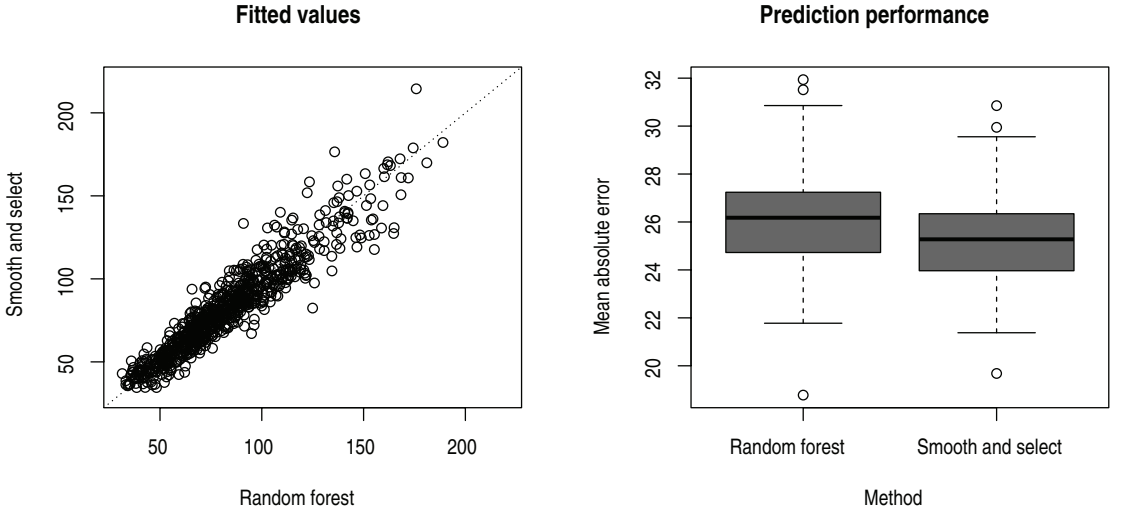

Figure 1 Fitted values for the food data with smoothing and selection penalty vs. random forest (left), prediction performance in terms of the mean absolute prediction error on randomly chosen test sets (right).

the remaining (approximately 700) observations the methods are trained. Figure 1 (right) shows prediction performance for the two methods in terms of the mean absolute error across 100 random splits into training and test data. It seems that the regularized linear model even produces slightly better predictions than the random forest on the food data, suggesting that no substantial interactions were missed by the linear model.

4 Discussion by J. Chiquet, Y. Grandvalet and G. Rigaill

Julien Chiquet, Yves Grandvalet and Guillem Rigaill raise an important point, namely the combination of a coding scheme and a penalty. They state that 'choosing an appropriate coding is at least as important as choosing the right penalty'. We agree that there are definitely coding/penalty combinations that work better than others, but it is hard to say whether one aspect might be more important than the other, because there is an interplay between penalty and coding and we always have to consider them jointly. For instance, some coding in combination with some specific penalty can be equivalent to some other coding/penalty choice. More specifically, let us have a look at ordinal predictors and the usual dummy coding with reference category in combination with the quadratic smoothing penalty (4) from Section 3.1.1 in the main article. On the one hand, the model being estimated with this approach is invariant against the choice of the reference category, because the value of the penalty does not change with the reference category, and neither does the likelihood. On the other hand, simple dummy coding with the smoothing penalty is also equivalent to

split coding, that is, a coding scheme parameterizing differences of adjacent categories (Walter *et al.*, 1987; Tutz, 2012), combined with a simple ridge penalty on those parameters; compare Gertheiss and Tutz (2009) and Gertheiss and Oehrlein (2011). Similar statements hold for the smoothing and selection penalty (Gertheiss *et al.*, 2011) and the ordinal fusion penalty (8) from Section 3.1.3 in the main article. Consequently, it is often not a matter of coding *or* penalty, but choosing the right penalty for a specific coding.

The analyses presented by Chiquet *et al.* nicely show that the naive combination of usual dummy coding with reference category and standard penalties is often a bad idea. More specifically, a simple ridge or (group) lasso penalty on dummy-coded variables may only be useful in some special cases; for instance, when the reference category is special in some sense, such as a control level, and differences to this level are of primary interest. Typically, however, other penalties are much more sensible for categorical variables, which has been one of the central points in our discussion paper.

We also showed, and want to point out here again, that dummy coding with reference category can indeed be used in regularized regression, as long as an appropriate penalty is chosen. An important reason for using a reference category is interpretation. Chiquet *et al.* mention that the use of a reference category is somewhat an heritage from the non-regularized setup, where it 'stabilizes the optimization process'. We rather think the reason is identifiability, not stabilization. If one has multiple predictors but as many parameters as categories per covariate, the parameters are not identifiable. By using regularization methods one may obtain estimates, but these are not identifiable either. To obtain identifiable parameters, still a side constraint is needed, which brings us back to interpretation, which is linked to the side constraints.

The concrete choice of coding/penalty combination should be guided mainly by interpretation and the purpose of the data analysis. This is also illustrated by the Wisconsin breast cancer dataset. For these data, Chiquet *et al.* favour the coop(erative)-lasso (Chiquet *et al.*, 2013), which is another interesting penalty also mentioned in the main article but not discussed in much detail. It is particularly useful under the assumption that there is a monotonic relationship between predictor and response. In case of the Wisconsin breast cancer dataset, one might rightfully assume that the relationship is monotonic. Consequently, the coop-lasso performs well. However, the penalty methods that are compared include only one method that actually uses the ordering of categories, namely the coop-lasso. When considering some of the other penalties from the main article that exploit the predictors' ordinal scale level, we see that also those penalties can do a very good job here, in particular the ordinal fusion penalty. For illustration, Figure 2 shows the results for four out of the nine predictors when using dummy coding with the first level as the reference in combination with the quadratic smoothing penalty (dotted), or the ordinal fusion penalty with standard (solid) or adaptive (dashed) weights. The plots are taken from Gertheiss *et al.* (2013), where also the results for the other five predictors can be found. We see that the quadratic penalty, which does not use any assumptions about monotonicity, produces estimates that appear less plausible. The fusion penalty, however, gives very interesting and monotonically increasing step functions (although not exploiting monotonicity assumptions either). Those functions can tell us whether some levels may be fused.

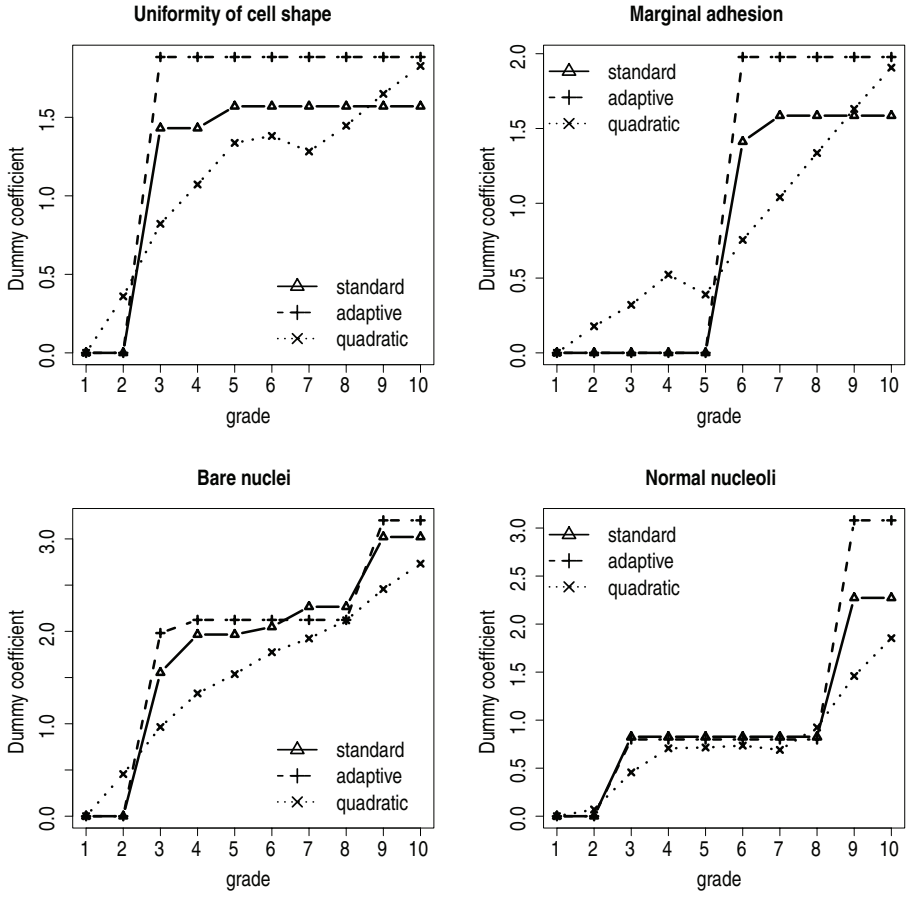

Figure 2 Exemplary results for some predictors from the Wisconsin Breast Cancer dataset with standard/adaptive fusion and quadratic smoothing penalty.

This might reveal how categories are used by the rater. Furthermore, if one goal of the analysis is to check whether the 1–10 grading scheme might be reduced to a simpler one with less levels, the fusion penalty is the one to choose. We hope that also Alan Agresti appreciates this example since he was rather sceptical about fusion penalties.

5 Discussion by C. J. Flynn, C. M. Hurvich and J. S. Simonoff

Flynn, Hurvich and Simonoff address a general problem of regularized estimators as the lasso. Although various loss bounds have been derived that support the use of the lasso for a deterministic choice of the regularization parameter, in practice, the tuning parameter is chosen data dependently with good reasons. As pointed out in the work of Flynn *et al.* (2014), the loss of the lasso when using data-dependent

tuning parameters and without knowing which variables have non-zero coefficients as compared to the loss obtained for the true sparse model is much larger than suggested by oracle inequalities. They demonstrate the effect for categorical predictors in a small simulation study. It is seen that in the presence of unnecessary predictors the model that includes all possible predictors yields much larger losses than the (unknown) true sparse model. The encouraging result is that the use of a more structured penalty, namely the ordinal group lasso instead of the simple lasso or the group lasso, yields losses that deteriorate definitely lesser. It would also be interesting to see the results for the fusion penalties discussed in the main article considering X either as nominal or ordinal. But it can be assumed that the results would be worse than for the ordinal group lasso, because in the true underlying model no categories are clustered. In general, we can only agree with the general message that carefully reasoned assumptions about underlying structures are helpful to obtain better estimates. Without assumptions, one is quite flexible but performance may suffer substantially. In our article, we tried to give an overview of penalties, and thus, in some sense, potential assumptions that may help when modelling categorical data. Another very important and popular way to take some prior belief about statistical models into account is Bayesian methods. We only sketched this in our article but fortunately Helga Wagner and Daniela Pauger considered Bayesian approaches (for categorical predictors) in more detail in their comment (see also Section 7 of this Rejoinder).

6 Discussion by C. Leng

Chenlei Leng raised several important questions, both with respect to regularization in general, and more concretely concerning the food spending data. In what follows, we will try to find answers to a least some of his questions.

(1) *What model to use?* As a general, but rather abstract, rule, the model should be flexible enough to take the specific characteristics of the data into account, but simple enough to make interpretation possible. In other (Einstein's) words, it should be 'as simple as possible, but not simpler'. In the case of a regression model with a large number of categorical predictors, for instance, we think that one would typically start with a main effects model, as interactions are difficult to interpret. As an example, consider the food data from the main article or the 'international classification of functioning, disability and health' (ICF) data from the work of Gertheiss *et al.* (2011). The ICF consists of various items with ordinal scale that can be used by health professionals to document the health and functioning of patients. To evaluate preselected, disease-specific sets of items, the so-called ICF core sets, a subjective, well-established measure of the patients' general well-being is regressed on those core sets. The ICF core set for chronic widespread pain, for instance, consists of 67 categorical variables, with 5 or 9 levels each (see Gertheiss *et al.* (2011) for details), making a main effects model the logical choice. With multi-categorical data, however, we would even consider nine covariates as available for the Wisconsin breast cancer data from Section 4 a 'large' number of predictors, due to the relatively large number

of parameters involved. Also in this case, we think a model beyond the main effects model will be hard to interpret. But of course this is not a strict rule and decisions need to be made with respect to the specific application. For example, there can be situations where a certain variable may act as an effect-modifying factor in a varying coefficient model; compare, for example, Gertheiss and Tutz (2012) and Hastie and Tibshirani (1993). Furthermore, the answer to the question about interpretability also depends on the personal perspective and preferences. With ordinal response, for example, as Alan Agresti pointed out, the proportional odds model may be preferred over a model with category-specific parameters because it gives 'first-order effects [that] are often informative for overall summaries, explaining the most important dimension of an effect'. Sometimes, however, the researcher only has to think about the model to start with, and use penalties to reduce a complicated model in a data-driven way to facilitate interpretation; which brings us to the next question.

(2) *What penalty to use?* As already pointed out above (see our response to the comment by Bühlmann and Dezeure), we belief the choice of the penalty should mainly depend on the specific nature of the application and the objective of the data analysis. For instance, if the researcher is mainly interested in differences between categories of categorical predictors and wondering whether some of those categories could be fused, a fusion penalty is obviously the one to choose. If predictors are ordinal and it can be assumed that there is a smooth effect, the quadratic smoothing penalty is preferable of course. In particular, when the number of predictors is large, the latter should be combined with selection. With the ICF data mentioned above, for instance, it hardly makes sense to assume that the influence of the ICF categories on the response is appropriately described by a step function. We would rather assume a smooth shape. Furthermore, it is intended to further reduce the ICF core sets. Therefore, the smoothing and selection penalty has been chosen here. Also with the food data, smoothing ordinal predictors makes sense, but if the researcher is particularly interested in differences between levels, the fusion penalty would be the better choice. This, however, depends on the objective of the analysis and the researcher's specific interests. Hence it is hard to give definite answers here. With nominal covariates having a relatively large number of categories, we think the fusion penalty is often a good choice, because the question which categories can/should be distinguished is a very typical one in this setting; compare, for example, Bondell and Reich (2009) and Post and Bondell (2013).

(3) *What criterion to use?* This question has different aspects, (a) should additional weights within the penalty be used and (b) how to choose the penalty parameter(s)? When talking about standard versus adaptive weights, the latter ones should only be used when the sample size is large enough such that the initial estimates determining the weights are sufficiently accurate. Although adaptive penalties with 'oracle properties' can also be problematic from a theoretical point of view (see, e.g., Pötscher and Schneider, 2009), it has been our experience that they often produce good results in practice when the sample size is large. When both ordinal and nominal predictors, or predictors with a different number of levels are included in the model/penalty, we need to choose weights that prevent us from penalizing certain terms more strongly than others (if it is not intended to do so). Finding an answer to (b) is difficult. It seems

that each researcher has personal preferences here, certainly with good reasons. Besides the approaches mentioned by Chenlei Leng, one could also use a hyperprior for the penalty parameters in a fully Bayesian framework or use (restricted) maximum likelihood when reformulating quadratic penalties as a mixed model; see also the comment by Helga Wagner and Daniela Pauger, and our response below.

7 Discussion by H. Wagner and D. Pauger

Flynn *et al.* already mentioned the 'advantages that carefully-reasoned appropriate assumptions about statistical structures' can offer (see above). The use of penalties can be seen as the frequentist way to incorporate, typically rather mild, prior assumptions in statistical modelling. Within a Bayesian framework, this would be done via appropriate prior distributions. In some cases, there is even a one-to-one connection to penalty methods. In our main article, however, we largely neglected Bayesian methods. Therefore, we are very grateful that Helga Wagner and Daniela Pauger considered those approaches in more detail.

Wagner and Pauger focus on effect fusion, which is based on parameters $\theta_{j,rs} = \beta_{jr} - \beta_{js}$, that is, pairwise differences of (dummy) coefficients. For ordinal predictors, it makes sense to consider differences of adjacent coefficients only, as discussed in detail in our main article. In a Bayesian framework, a prior distribution is chosen for the θ-parameters, which is very similar to the concept of using penalties. With ordinal predictors and independent, mean zero normal priors with a given variance, the Bayesian approach is even completely equivalent to the quadratic smoothing penalty (4) from the main article if the penalty parameter is fixed at the right value (see also Gertheiss and Tutz, 2009). However, there are also differences between Bayesian regularization and our penalties. In a fully Bayesian framework, for instance, a hyperprior is put on the variance/penalty parameters, and those parameters are thus estimated jointly with the regression coefficients of interest. This is often seen as a major advantage of Bayesian methods as 'no cross-validation is needed [...] and [...] also uncertainty on the regularization parameters can be assessed' (Wagner and Pauger). However, also the hyperprior typically depends on some parameters. This is illustrated by Wagner and Pauger who considered the rent data from Gertheiss and Tutz (2010) and used a normal prior for the θs, an inverse gamma prior distribution for the corresponding variance parameters, and varied the inverse gamma's scale parameter. The resulting coefficient paths for the β-coefficients show how those depend on the choice of the hyperprior's scale parameters. So also with Bayesian methods the question how to choose those parameters remains.

An interesting intermediate between Bayesian methods and the use of penalties, in particular for ridge-type penalties, is the use of mixed models. For instance, the quadratic smoothing penalty for ordinal predictors can also be formulated in a mixed models framework where θ-parameters from above are specified as random effects; compare Gertheiss and Oehrlein (2011) and Section 3.2 of the main article. In this case, the variance parameters of the random effects can be estimated by maximum

likelihood or restricted maximum likelihood, without the need for cross-validation. In addition, the entire mixed models machinery for statistical inference is available, including statistical tests and confidence intervals (compare the response to the comment by Bühlmann and Dezeure).

References

Bondell HD and Reich BJ (2009) Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, **65**, 169–77.

Breiman L (2001) Random forests. *Machine Learning*, **45**, 5–32.

Chiquet J, Grandvalet Y and Charbonnier C (2013) Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics*, **6**, 795–830.

Dezeure R, Bühlmann P, Meier L, and Meinshausen N (2015) High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, **30**, 533–58.

Flynn CJ, Hurvich CM and Simonoff JS (2014) On the sensitivity of the lasso to the number of predictor variables. arXiv preprint arXiv:1403.4544.

Gertheiss J (2014) Anova for factors with ordered levels. *Journal of Agricultural, Biological, and Environmental Statistics*, **19**, 258–77.

Gertheiss J, Hogger S, Oberhauser C and Tutz G (2011) Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **60**, 377–96.

Gertheiss J and Oehrlein F (2011) Testing relevance and linearity of ordinal predictors. *Electronic Journal of Statistics*, **5**, 1935–59.

Gertheiss J, Stelz V and Tutz G (2013) Regularization and model selection with categorical covariates. In Lausen B, Van den Poel D and Ultsch A eds., *Algorithms from and for Nature and Life*, pages 215–222. Berlin/Heidelberg: Springer.

Gertheiss J and Tutz G (2009) Penalized regression with ordinal predictors. *International Statistical Review*, **77**, 345–65.

Gertheiss J and Tutz G (2010) Sparse modeling of categorial explanatory variables. *Annals of Applied Statistics*, **4**, 2150–80.

Gertheiss J and Tutz G (2012) Regularization and model selection with categorial effect modifiers. *Statistica Sinica*, **22**, 957–82.

Hastie T and Tibshirani R (1993) Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**, 757–96.

Hoefling H (2010) A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, **19**, 984–1006.

Leitenstorfer F and Tutz G (2007) Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, **8**, 654–73.

Liaw A and Wiener M (2002) Classification and regression by randomforest. *R News*, **2**, 18–22.

Lin D, Pramana S, Verbeke T and Shkedy Z (2014) IsoGene: Order-Restricted Inference for Microarray Experiments. R package version 1.0-23.

Lin D, Shkedy Z, Yekutieli D, Burzykowski T, Göhlmann HW, Bondt AD, Perera T, Geerts T and Bijnens L (2007) Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resampling-based inference. *Statistical Applications in Genetics and Molecular Biology*, **6**, article 26.

Meier L, Meinshausen N, and Dezeure R (2014) hdi: High-Dimensional Inference. R package version 0.1–2.

Post J and Bondell H (2013) Factor selection and structural identification in the interaction ANOVA model. *Biometrics*, **69**, 70–9.

Pötscher BM and Schneider U (2009) On the distribution of the adaptive lasso estimator.

Journal of Statistical Planning and Inference, **139**, 2775–90.

Pramana S, Lin D, Haldermans P, Shkedy Z, Verbeke T, De Bondt A, Talloen W, Göhlmann H and Bijnens L (2010) IsoGene: An R package for analyzing dose-response studies in microarray experiments. *R Journal*, **2/1**, 5–12.

Sweeney E, Crainiceanu C and Gertheiss J (2016) Testing differentially expressed genes in dose-response studies and with ordinal phenotypes. *Statistical Applications in Genetics and Molecular Biology*, in press.

Tutz G (2012) *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Tutz G and Leitenstorfer F (2007) Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics*, **16**, 165–88.

Walter SD, Feinstein AR, and Wells CK (1987) Coding ordinal independent variables in multiple regression analysis. *American Journal of Epidemiology*, **125**, 319–23.