Statistical Modelling 2016; **16(3)**: 205–211

Discussion on 'regularized regression for categorical data (Tutz and Gertheiss)'

Peter Bühlmann[1] and Ruben Dezeure[1]
[1]Seminar for Statistics, Department of Mathematics, ETH Zürich, Switzerland

Key words: categorical data, combining different penalization types, confidence intervals, high-dimensional generalized regression, ordinal categorical predictors, *p*-values, uncertainty quantification

We congratulate Gerhard Tutz and Jan Gertheiss, referred in the sequel as 'TG', for an interesting and inspiring article on the important topic of regression for categorical data. Much has happened over the 15 years, including earlier contributions from the authors. The current overview and expository article is the most welcome addition to the literature.

1 Some further thoughts on the article

TG have written a masterpiece on modern regression with categorical covariables. Special attention is given to the case with ordinal categorical predictors: TG explain a variety of possibilities to build in additional information in terms of sparsity, smoothness and 'clusteredness'. From a fundamental perspective, the case with ordinal categories is complex and difficult. Using a vague analogy to nonparametric curve estimation, the issue here is a kind of varying regularization problem to adapt well to the rough and flat parts of an underlying true function (besides the role of selection). 'Clever' regularization or penalization is certainly a powerful vehicle for obtaining accurate point estimates of the underlying generalized regression coefficients, or for model parameters beyond regression as discussed also in TG in Sections 5 and 6.

1.1 Different penalization types for different predictors

A natural question in practice will be the choice of the penalization type which might be different for different covariables and which of the categorical variables should be subject to a penalty term. Regarding the second question, a sparsity-type penalization

Address for correspondence: Peter Bühlmann, Seminar for Statistics, Department of Mathematics, HG G 17, Rämistrasse 101, CH-8092 Zürich, Switzerland.
E-mail: buhlmann@stat.math.ethz.ch

 10.1177/1471082X16642566

should automatically *select* the variables to be strongly clustered, strongly smoothed or thresholded to zero and, thus, such sparsity inducing schemes often work well for the purpose of selection (on the level of individual or groups of parameters) (van de Geer *et al.*, 2014). Cross-validation is a natural candidate for addressing the first question (aside from its use to choose the amount of regularization), but it might become computationally very cumbersome to run it a large multitude of times for evaluating many combinations. Here, a combination means that some subset of ordinal covariables are subject to a clustering penalty (Section 3.1.3 in TG), some are subject to a smoothing penalty (Section 3.1.1 in TG) and some are subject to a sparsity only penalty (Section 3.1.2 in TG). We will illustrate in Section 2 an exploratory way, by inspecting confidence intervals, from which one might extract some reasonable information for a combination of the qualitatively different penalty terms acting on different covariables.

1.2 Tree methods and random forests

Section 7.3 in TG mentions tree-based approaches. They have the advantage that they work in a natural way for mixed variables, where some of the covariables are continuous, some categorical with nominal and ordinal categories. Approaches based on penalization become intrinsically more complicated with mixed data, requiring a careful choice of various penalty terms and their corresponding regularization parameters (see also above regarding the difficulty to choose different penalization types). If the task would be prediction only, random forests (Breiman, 2001) is often a very powerful method which (essentially) does not require to choose any tuning parameter(s): Reasons for its success in general include that the algorithm also takes into account interactions between variables and that it can deal in a scale-free manner with mixed variables. We wonder how well it would perform for prediction in the dataset on 'Spending for Food' analyzed in TG. Based on adaptations of random forests, we have obtained in other works some interesting results for predictive survival modelling (Hothorn *et al.*, 2006), for predictive imputation of missing data (Stekhoven and Bühlmann, 2012) but also for variable importance in mixed undirected graphical models (Fellinghauer *et al.*, 2013). The disadvantage of random forests 'technology' is that it does not yield statistical point estimation or inferential statements in a well-defined statistical model and notions of measuring 'variable importance', as advocated in Breiman (2001) or its modified version in (Strobl *et al.*, 2008), are still very algorithmic with no statistical guarantees in terms of p-values or confidence intervals.

2 Some outlook: Confidence intervals and *p*-values

We would like to outline here the additional perspective of uncertainty quantification. Recently, some progress has been made to construct confidence intervals and statistical hypothesis tests in potentially high-dimensional generalized regression models (Bühlmann, 2013; Zhang and Zhang, 2014; van de Geer *et al.*, 2014; Javanmard and Montanari, 2014; Dezeure *et al.*, 2015; van de Geer and

Stucky, 2016). The key idea is to 'de-bias' or 'de-sparsify' the sparse Lasso estimator (Tibshirani, 1996). Starting from the Lasso, the de-sparsifying operation leads to a non-sparse but regular estimator whose low-dimensional components have an asymptotic Gaussian distribution.

For example, consider a linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon.$$

with fixed design $n \times p$ matrix $\mathbf{X}$ and a Gaussian error term ε with i.i.d. components having mean zero and variance σ_ε^2 (where p denotes here the number of parameters). The situation might be very high-dimensional with $p > n$, but the parameter β^0 is assumed to be sparse. Assuming such sparsity conditions and a restricted eigenvalue assumption on $\mathbf{X}$ (cf. Bühlmann and van de Geer, 2011), we have:

$$\sqrt{n}(\hat{\beta} - \beta^0) = W + \Delta,$$
$$W \sim \mathcal{N}_p(0, \sigma_\varepsilon^2 \Omega), \ \max_{j=1,\ldots,p} |\Delta_j| = o_P(1) \ (p \geq n \to \infty).$$

where Ω is a known covariance matrix depending on the design matrix $\mathbf{X}$ (in analogy to ordinary least squares where Ω would be equal to $(\mathbf{X}^T\mathbf{X}/n)^{-1}$). The remarkable feature of this result is that: (a) the de-sparsified estimator has $1/\sqrt{n}$ convergence rate, despite the very high-dimensional setting and unlike for estimating the regression function where the convergence rate is only $\sqrt{\log(p)/n}$ (cf. Bühlmann and van de Geer, 2011) and (b) the limiting distribution is known up to the unknown noise variance which can be estimated also in the high-dimensional setting. Thus, we can construct confidence intervals for single parameters β_j^0, and we can test statistical hypotheses for groups of parameters. Regarding the latter, a group hypotheses is of the form

$$H_{0,G};\ \beta_j^0 = 0 \text{ for all } j \in G.$$

where $G \subseteq \{1, \ldots p\}$. If G is very large, we have to rely on the test-statistic $\max_{j \in G} |\hat{\beta}_j / \text{s.e.}(\hat{\beta}_j)|$, and if G is small (say, $|G| \leq 10$), we can also use the sum-statistic $\sum_{j \in G} \hat{\beta}_j$: the corresponding asymptotic limiting distributions can be calculated or efficiently simulated in case of the max-type statistics. In addition, multiple testing correction can be done efficiently since the covariance structure, proportional to Ω, is known. All of this (but not the sum-type statistics) is implemented in the `R`-package `hdi` for generalized linear models (Meier *et al.*, 2014; Dezeure *et al.*, 2015).

The methodology outlined above can be used for categorical (nominal) predictor variables, at least in the simple form. In the notation of TG, consider the model formulation equation (1) for a linear model as in equation (3): the regression parameters for the jth covariable are $\beta_{jr},\ r = 1, \ldots, k_j$ (using the constraint from TG with $\beta_{j0} = 0$), for $j = 1, \ldots, p$. The R-package `hdi` then leads to confidence intervals for the underlying true parameters $\beta_{jr}^0,\ r = 1, \ldots k_j,\ j = 1, \ldots p$. We can also test whether

the j categorical covariable has a significant effect on the response Y by considering the null hypothesis

$$H_{0,G_j}: \ \beta_{jr} = 0 \text{ for all } r = 1, \ldots, k_j.$$

Testing this hypothesis is related to the group wise selection discussed in Section 3.1.2 in TG, but now with a statistical uncertainty measure in terms of a p-value instead of a point estimator only.

We illustrate the confidence intervals of the single regression parameters for a simulated dataset. The suggested approach allows also to informally infer whether the coefficients from ordered categorical covariables are similar for neighbouring levels within the same categorical variable, by inspecting the confidence intervals. The results are displayed in Figure 1. If the confidence intervals would heavily overlap for the coefficients from neighbouring levels within ordinal categorical variables, clustering of the corresponding categories would be natural; if the overlap of neighbouring confidence intervals would be less pronounced, smoothing of the levels would seem reasonable and if the confidence intervals would not overlap at all, there should be no penalty or regularization for the levels within the categorical variable. In particular, with such a '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. (Trying all combinations and optimizing a cross-validation score could become quickly computationally infeasible). We infer from Figure 1 that there are four significant categorical variables, namely Variables 1, 3, 15 and 31 and these are indeed the only true variables having an effect. The plot also suggests the following scenarios: Variable 1 has two clusters of categorical values (which is true), Variable 3 has unstructured categorical values (which is true), Variable 15 has two clusters of categorical values (which is true) and that there is an indication that Variable 31 has a smooth structure for its categorical values (which is true). Thus, the plot in Figure 1 is indeed rather informative about the true underlying structure. More statements can be made: the significance of the categorical variables when adjusted for controlling the familywise error rate in multiple testing, using the Bonferroni-Holm procedure, leads to the four significant variables (as mentioned already) and to no false positive findings; without the multiplicity adjustment, 9 out of 554 confidence intervals for coefficients whose true values are zero do not cover the true value zero and hence would lead to false positives.

The model underlying the one simulated dataset of sample size $n = 800$ with 100 categorical variables is constructed as follows.

(M) Consider $X \sim \mathcal{N}_{100}(0, \Sigma)$ with Toeplitz matrix $\Sigma_{ij} = 0.9^{|i-j|}$. Generate n i.i.d. samples from $\mathcal{N}_{100}(0, \Sigma)$. The active set of categorical variables with non-zero effect for the response is $S^0 = \{1, 3, 15, 31\}$. We categorize each variable $X^{(j)}$ ($j = 1, \ldots, 100$) according to the empirical quantile (over the $n = 800$ realizations) of $X^{(j)}$ with U_j categories: $U_1 = 5$, $U_3 = 6$, $U_j = 10$ ($j = 15, 31$), all other U_j's from Uniform from $\{2, \ldots, 12\}$. This leads to a dummy encoding design matrix

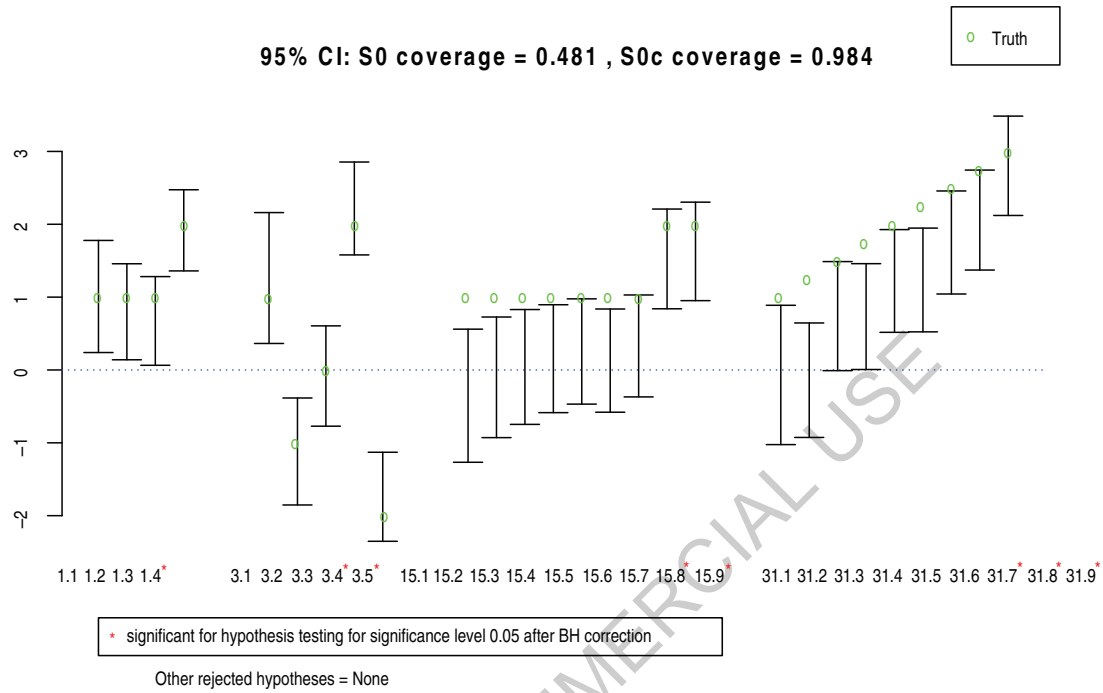

Figure 1 Confidence intervals for the non-zero regression coefficients based on the de-sparsified Lasso for high-dimensional inference with `R`-package `hdi`. The underlying data is a single realization from model (M) described below with sample size $n = 800$ and 100 categorical variables with 581 parameters in total. Shown are the confidence intervals for the parameters from the active set $S^0 = \{1, 3, 15, 31\}$ of the categorical variables. The complement $(S^0)^c = \{1, \ldots, 100\} \setminus S^0$ encodes the categorical variables which have no effect on the response. The true parameters were covered by the nominal 95% confidence intervals in 48.1% of the 27 intervals corresponding to the active variables in S^0 (often only slightly missing to cover the true parameter) and in 98.4% of the 554 intervals corresponding to the non-active variables in $(S^0)^c$. When using a Bonferroni-Holm correction for controlling the familywise error rate in multiple testing, we detect all the four categorical variables from S^0 and no false positive finding is made (i.e., no significant variable from $(S^0)^c$). **Source:** Authors' own.

$\mathbf{X}$ with $\sum_{j=1}^{100}(U_j - 1)$ columns. Consider the regression coefficients:

clustered categorical values:
$\beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 1,\ \beta_{1,4} = 2,$
$\beta_{15,1} = \beta_{15,2} = \ldots = \beta_{15,7} = 1,\ \beta_{15,8} = \beta_{15,9} = 2,$
smoothed categorical values:
$\beta_{31,1} = 1,\ \beta_{31,2} = 1.25,\ \beta_{31,3} = 1.5, \ldots, \beta_{31,9} = 1 + 8 \cdot 0.25,$
unstructured coefficients:
$\beta_{3,1} = 1,\ \beta_{3,2} = -1,\ \beta_{3,3} = 0,\ \beta_{3,4} = 2,\ \beta_{3,5} = -2.$

Finally, take the error term as $n = 800$ realizations from $\varepsilon \sim \mathcal{N}(0, 1)$.

We have not explored how to obtain confidence statements based on regularization for smoothing or clustering the category levels within an ordinal categorical variable (whose importance has been presented in TG for point estimation). Following the same idea as for the de-sparsified estimator outlined above (see also van de Geer *et al.*, 2014), one might be able to gain statistical power by constructing confidence statements based on a version of a regularized estimator which encourages not only sparsity (as the Lasso), but also to smooth or cluster ordinal categories. To our knowledge, this has not been pursued so far.

3 Conclusions

Tutz and Gertheiss have provided a very insightful and careful overview of methodology developed in the last 10–15 years for regularized regression for categorical data. Their focus is on point estimation, which is the first step in statistical inference. We have illustrated that recent techniques for constructing confidence statements in high-dimensional generalized regression have the potential to be useful also for regression with categorical data. For a simulated dataset with sample size $n = 800$, 100 categorical variables and 581 parameters in total, we obtain very reasonable results even without smoothing or clustering categorical values: this can serve as a step for inferring which categorical variables (or parts of their levels) should be subject to smoothing or clustering as described in TG. Obtaining more powerful confidence statements by also using smoothing or clustering of ordinal categories might be possible. Achieving this, particularly in the high-dimensional setting, is an open problem.

References

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

Bühlmann P (2013) Statistical significance in high-dimensional linear models. *Bernoulli*, **19**, 1212–42.

Bühlmann P and van de Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg: Springer-Verlag.

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.

Fellinghauer B, Bühlmann P, Ryffel M, von Rhein M and Reinhardt J (2013) Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, **64**, 132–52.

Hothorn T, Bühlmann P, Dudoit S, Molinaro A and van der Laan M (2006) Survival ensembles. *Biostatistics*, **7**, 355–73.

Javanmard A and Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, **15**, 2869–909.

Meier L, Meinshausen N and Dezeure R (2014) *Hdi: High-dimensional Inference*. http://CRAN.R-project.org/package=hdi. R package version 0.1-2.

Stekhoven D and Bühlmann P (2012) Missforest—Nonparametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–18.

Strobl C, Boulesteix A-L, Kneib T, Augustin T and Zeileis A (2008) Conditional variable importance for Random Forests. *BMC Bioinformatics*, **9**, 307.

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–88.

van de Geer S, Bühlmann P, Ritov Y and Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, **42**, 1166–202.

van de Geer S and Stucky B (2016) χ^2-confidence sets in high-dimensional regression. In Statistical Analysis for High-Dimensional Data. The Abel Symposium 2014 (eds. Frigessi A, Bühlmann P, Glad IK, Langaas M, Richardson S and Vannucci M), pp. 279–306. Springer International Publishing

Zhang C-H and Zhang S (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, **76**, 217–42.