Complexity reduction via deselection for boosting distributional copula regression

Annika Strömer[1], Nadja Klein[2], Christian Staerk[1], Hannah Klinkhammer[1], Andreas Mayr[1]

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn, Bonn, Germany
[2] Chair of Uncertainty Quantification and Statistical Learning, Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Dortmund, Germany

E-mail for correspondence: `stroemer@imbie.uni-bonn.de`

Abstract: Boosting distributional copula regression is a flexible tool to jointly model multivariate outcomes, in which all parameters of the joint distribution can be related to covariates via additive predictors. Estimation via model-based boosting allows to fit these complex models also to high-dimensional data ($p > n$). Additionally, boosting can incorporate data-driven variable selection simultaneously for all parameters of the marginal distributions as well as for the association parameter of the copula. However, as known from univariate (distributional) regression models, the boosting algorithm tends to select too many variables, particularly for low-dimensional settings ($p < n$). To counteract this behavior, we adapt a recent deselection approach for statistical boosting to multivariate copula regression models to deselect base-learners with only a negligible impact on the overall performance of the model. We illustrate our approach by jointly modelling LDL and HDL cholesterol based on large UK Biobank genotype data.

Keywords: Model-based boosting; Variable selection; GAMLSS; Copula regression.

1 Introduction

In distributional copula regression, potentially different marginal response distributions can be combined by an appropriate copula function that defines the dependency structure between the outcomes for multivariate modelling. Within the framework of generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005), all parameters

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

of the distributional copula regression model (i.e. the distribution parameters of the marginals and the dependency parameter) are modelled by an additive predictor incorporating different effect types for the covariates (Klein and Kneib, 2016). In combination with component-wise gradient boosting, we can incorporate data-driven variable selection for potentially high-dimensional data, which is controlled by the number of boosting iterations (Mayr et al., 2012). However, despite these advantages, the boosting algorithm still tends to select too many variables (including ones which are non-informative or have a very low signal), which occurs particularly for low-dimensional settings. In these situations, we can observe a slow overfitting behavior, which results in a later stopping of the algorithm and therefore a larger set of included base-learners that might have only minor importance. As a result, we are faced with an unnecessary large model, that might be performing good for prediction but is difficult to interpret.

2 Deselection of base-learners

We address this issue by adapting the deselection approach by Strömer et al. (2022) for boosting distributional copula regression. The pragmatic and simple idea is to start with a classical boosted model tuned by cross-validation or resampling techniques to determine the optimal stopping iteration m_{stop} to achieve high prediction accuracy. Then, the base-learners and variables that were selected but only have a minor impact on the model are identified and are deselected. Afterwards, the model is boosted again only with the remaining ones. The importance of a base-learner j in the deselection approach is measured via the risk reduction after m_{stop} iterations:

$$R_j = \sum_{m=1}^{m_{\text{stop}}} I(j = j^{*[m]})(r^{[m-1]} - r^{[m]}), \quad j = 1, \ldots, \sum p_k,$$

where I denotes the indicator function and $j^{*[m]}$ is the selected base-learner in iteration m. Furthermore, $r^{[m-1]} - r^{[m]}$ represents the risk reduction in iteration m, for risks $r^{[m]}$ and $r^{[m-1]}$ at iterations m and $m-1$. Note that in the case of distributional copula regression, all distribution parameters are considered together and each parameter $\theta_k, k = 1, \ldots, K$ may depend on a different number of variables p_k. For a given threshold $\tau \in (0, 1)$, we deselect base-learner j if

$$R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]}),$$

where $r^{[0]} - r^{[m_{\text{stop}}]}$ represents the total risk reduction and R_j denotes the attributable risk reduction of base-learner j. In other words, only base-learners which contribution R_j to the total risk reduction is larger than the relative τ threshold (e.g., 1%, Strömer et al, 2022) will remain in the model after the deselection step.

3 Simulations for comparison with competitors

We conducted a simulation study (based on a similar set-up as in Hans et al., 2023) to investigate and compare the variable selection properties, the predictive performance and the computation time of the classical boosting algorithm with the adapted deselection approach. As additional competitors, we also considered stability selection (Meinshausen and Bühlmann, 2010) and probing (Thomas et al., 2017) to benchmark our results. For a low-dimensional setting, the classical boosted model selected many non-informative variables for every distribution parameter. All approaches effectively reduced the number of false positives. Probing and stability selection did not select all informative variables in each simulation run, whereas the deselection approach maintained all informative variables in the model. In a high-dimensional setting, fewer non-informative variables were included in the boosted models. The approaches performed similar as in the low-dimensional setting and reduced the number of selected non-informative variables almost completely.
A comparison of the negative log-likelihood for the low-dimensional and high-dimensional setting showed that stability selection and deselection resulted in a slightly better predictive performance than the classical boosted model. Probing, on the other hand, led to a lower predictive performance. In terms of computation time, probing is the fastest and stability selection takes much more computational resources than the classical boosted model or the deselection approach.

4 Joint modelling of LDL and HDL cholesterol

We illustrate our deselection approach on high-dimensional genomic cohort data from the UK Biobank, modelling the joint genetic predisposition for two continuous phenotypes, LDL and HDL cholesterol, in dependence of different genetic variants. For both phenotypes, the 1000 variants (typically single nucleotide polymorphisms) with the largest marginal associations with each of the two phenotypes were selected in a pre-screening process. Overall, the data set includes 20,000 sampled observations and 1,179 variants (803 variants selected for both phenotypes). The log-logistic distribution was considered as marginal distribution for both phenotypes and the Gumbel copula was used for modelling the dependency structure based on the comparison of the predictive risk. All variants were included with simple linear models as base-learners.
Figure 1 illustrates the resulting estimated absolute coefficients for every distribution parameter (similar to Manhattan plots). The classical boosted model selected several variants for each distribution parameter. After the deselection approach with $\tau = 0.01$, only some variants for μ_1 and μ_2 are left. This means that with the deselection approach we can not only reduce

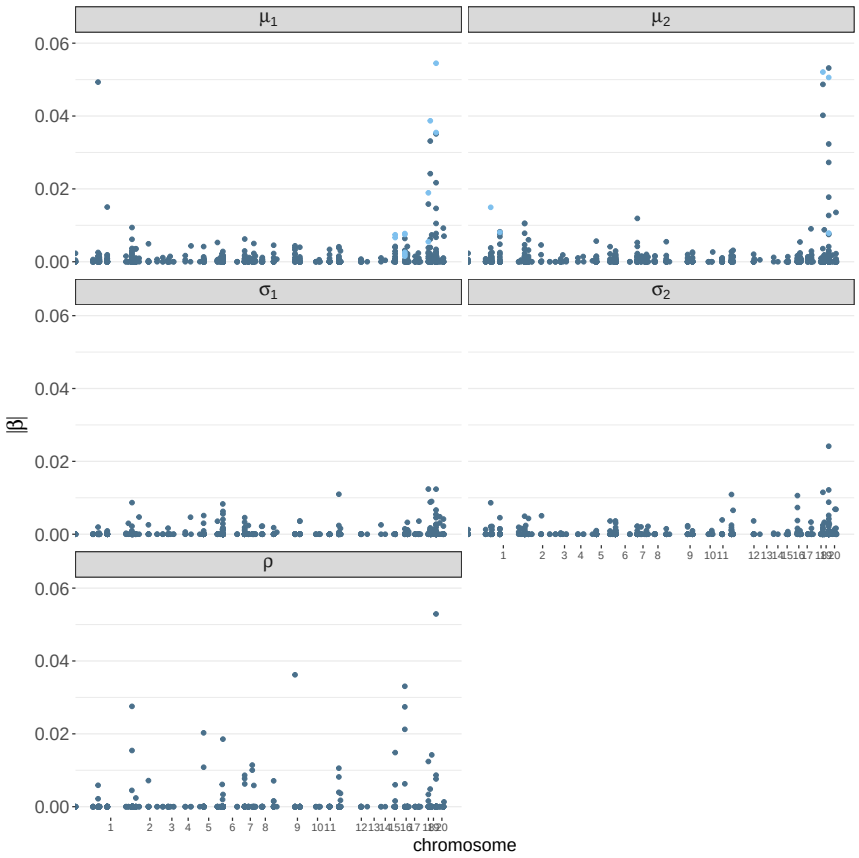

FIGURE 1. Manhattan-type plots (chromosomes on x-axis) for the absolute coefficients of boosted copula regression for the joint analysis of LDL and HDL cholesterol. The dark blue dots are the selected variants by classical boosting, the lighter blue points are the remaining variables after the deselection approach.

the included variables and obtain a much sparser model with a potentially simpler interpretation: In this case the approach also further reduces the overall complexity by completely deselecting all variants of distribution parameters resulting in two simple univariate models for both phenotypes. Furthermore, the deselection leads to a comparable predictive performance on test data as the classical boosted model.

5 Conclusion

We presented a pragmatic deselection approach for boosting multivariate distributional copula regression models. The new deselection approach re-

sults in much sparser models and can even lead to more simple univariate regression models, reducing the complexity of the overall analysis. The prediction accuracy usually does not improve but can lead to comparable accuracy as the classical boosted model with less predictors. Consequently, the interpretability of resulting prediction models is improved.
The presented deselection procedure is controlled via a threshold value τ, which represents the minimum amount of total risk reduction which should be attributed to a corresponding base-learner in order to avoid deselection. This can be interpreted as a threshold-value for the importance of the particular predictor variable. In the simulation study, a threshold of $\tau = 0.01$ (i.e. 1% of total risk reduction) was considered to be appropriate. However, depending on the research question and the context of the problem, the choice of τ is a trade-off between more complex models with the highest prediction accuracy and a sparser, more interpretable model with potentially reduced prediction accuracy.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776).

References

Hans, N., Klein, N., Faschingbauer, F., Schneider, M. and Mayr, A. (2022). Boosting distributional copula regression. *Biometrics*, **00**: 1 – 13.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, **26** (4), 841 – 860.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C*, **61** (3): 403 – 427.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4): 417 – 473.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507 – 554.

Strömer A., Staerk C., Klein N., Weinhold L., Titze S. and Mayr A. (2022). Deselection of base-learners for statistical boosting — with an application to distributional regression. *Statistical Methods in Medical Research*, **31**(2): 207 – 224.

Thomas, J., Hepp, T., Mayr, A. and Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Computational and Mathematical Methods in Medicine*, **2017** 1 – 8.