Review for the MA8701 Data Analysis Project "Water Flow Modeling"

Authors: Group 7 (Eirik Berge, Camilla Idina Jensen Elvebakken, Martin Ludvigsen)

Reviewers: Group 5 (Helene Minge Olsen, Florian Beiser, Yaolin Ge)

General Comments

The group provides an summary to their data analysis repository in the README.md. Therein and the Introduction, the authors highlight the relevance of their prediction models for safety as well as energy production and mention the current operational HBV modelling. This gives a clear motivation for their data-driven prediction modelling in the subsequent report. For reference, a linear regression model is presented which already proves the potential of shrinkage by subset selection. In the shrinkage section, many different methods have been applied. It is nice to see the implementation and result of different methods. In general, it worked as expected to reduce the variance of the predicted model. The conclusion provides a summary of the qualitative results of the different models and points out the implicit assumption in the project that have to be considered for generalisation of the results.

Introduction

The report provides a fundation for the understanding of the practical problem setting and the provided data. The preprocessing of the data base is cleary comprehensible. However, a bunch of presentations of relation between single variables makes the Introduction lengthy and we recommend to discard some of them.

Technical details:

- In the figure for missing data the y-axis could be transformed to a scala with years.
- Are the orange lines linear regression results?
- We cannot see very little information in the `dato`/`day` versus `vannføring` plots and hence recommend to delete them - the boxplot is much more illustrative and significant. This helps to keep the report constructive.
- Mention quantile value in boxplot description.

Linear Models

Already in this section whose principal purpose is mainly to produce reference results, the authors already show impressively the potential of shrinkage methods. The section helps to understand the characteristics of the data set. As the rest of the work, this section has a high technical quality. For the seek of the project the MSE is an acceptable loss function, but similarly to the honesty that there are may other relevant risk attitudes, the reviewers would appreciate that the adjusted R-value is set into relation to other potential measures for the goodness-of-fit, since the chosen one improves by definition with the number of variables, what may not suited for shrinkage methods.

Technical details:

- It is a bit confusing confusing that you show that your `data$train` has 54 variables including `vannstand`, `dato` and `modellertvannføring` despite those are not considered, but this is personal preference. Please indicate this when printing `dim(train)`.
- Do you have an idea why already linear regression is better than HBV? Is it the chosen loss function? If you have an educated guess, it would be nice to include.
- The visualisation of the predicion is very illustrative. Please always use the same time range for those plots.

Shrinkage

The reviewers like it to see multiple implementation on lasso, elastic net and fused lasso. It is interesting to see the effect from fused lasso on the predicted water level, it achieves the goal of reducing variance as a shrinkage method even though it might hamper the capability for the fused lasso model to have higher accuracy. It is good to see the comparison in terms of the test MSE to dig out which one performs the best. Although linear regression works pretty perfect, 50 numbers of covariates are certainly not easy to interpret. Therefore, you did a great job by applying different lassos for proving its ability of shrinking. In the lasso section, the reviewers are wondering why the regularisation lambda is so small to reduce the number of the covariates dramatically and would appreciate a statement on this fact. Also, in the plot showing the seasonability variation comparing different effects from multiple methods. It seems that the variances in those significant months are shrinked, but not in other months, so it would be nice if you could have a discussion on why this happened. But in general, it is a very comprehensive part to read and understand.

Comparison and Conclusion

The authors encapsulate shortly the previous results. Again the data explanation is very long, whereas the discussion of the results and the different properties of the previous outcomes is more of interest in this project. The stated questions for further application are relevant, maybe a comparison of different losses or other criteria for the suitability of a model should be added to the outlook.

Summary

Additionally to the detailed report the group provides a summary in the README.md. However, in the summary the reviewers would prefer to see less exploration of the data but more comments on the models: what is the statistical modelling approach? Why are particularly those different shrinkage methods applied?

Conclusion

The authors perform a statistically sound analysis of the given data and develop interpretable prediction models, where they deliberately respect the time-dependence of the variables. The methods are more than appropriate for the scope of the project and the results are clearly presented. The report is very well written. It stands out that the authors implement, interpret and compare five different approaches!

We suggest the following minor amendments:

- Change the scope of the summary from explaining the practical application towards methodical statistical focus and reasons for the selection of shrinkage methods, supplemented by a bit longer result presentation.
- Please keep the descriptions compact and reduce the lengthiness in some part, but emphasis more why you are choosing particular methods and what are advantages/disadvantages. (You do this already but

it disappears in the long paragraphs.)

- The discussion of the questions posed in the review above would be appreciated, if you have substantive ideas.