## Report on paper COST-D-20-00534

## Using reference models in variable selection

This paper is a continuation of Pironen et al. (2020) where a methodology to make predictive inference and feature selection in high-dimensional problems is proposed. This methodology uses a two-state approach: first, it is built a model (possibly non-sparse) that predicts well and tries to approximate the true data generation mechanism of future data. This model is named as a *reference model*. Then, a minimal subset of features that gives similar predictions as this reference model is selected. This selection is carried out using a variable selection algorithm (named as *projpred*) that is described in the appendix A of the present paper. This algorithm uses a projection of the predictive distribution of the *reference model* to the explored reduced models to calculate their predictive distributions, and it uses some cross-validation predictive performance measures as, for instance, the log predictive density (MLPD) over the validation points, to carry out the selection process. Three different types of predictive prediction techniques are proposed: draw-by-draw (that is used in the present paper), single point and clustered.

The authors propose to use an approximate leave-one-out cross validation via Pareto-smoothed importance sampling procedure to calculate a predictive utility score of each explored model, and to choose the smallest sub-model (size) that is *"sufficiently close"* to the reference model's predictive utility. Additionally, a complete variable selection algorithm, which tries to find all relevant variables that have some predictive information about the target is proposed. This algorithm is based on the use of iterative projections that repeats the *projpred* selection for different iterations, at each time excluding the variables selected in the previous iterations from the search. Using simulated and real examples the authors try to show that the use of reference models generally translates into better and more stable variable selection, and that the projection predictive approach shows superior performance as compared to alternative variable selection methods independently of whether or not they use reference models.

The paper is well written and organized an the topic analyzed is very relevant and practice. However, in my view, the goals of the paper, described in the introduction section, have already been reached in the excellent paper of Pironen et al. (2020) and, in this sense, the contributions of the current paper is very marginal. The framework of

Pironen et al. (2020) is more general (they consider exponential family models and generalized linear models) and provide some theoretical results that, as they said "… *help us to understand when the reference model could be helpful for parameter learning in linear submodels"* (section 8 of Piironen et al., 2020). The only innovation that I can appreciate in the current paper is the proposal of an algorithm to solve a complete variable selection problem that, unfortunately, does not work very well as the authors explicitly recognize: "*The great performance of projpred in minimal subset selection is not carried out over for the complete variable selection with iterative projpred …, and the methods specifically designed for the complete variable selection perform better"* (page 12, lines 24-27), even though they observe a better selection to iterative lasso. In my view, this is due to that the algorithm is too simple (as the authors explicitly recognize in the conclusions) and that the projective prediction approach was not designed for the complete variable selection. You should modify this algorithm in such a way that a more exhaustive exploration of the space of models was carried out, perhaps using Bayesian variable selection procedures based on the use of Occam's window or something similar.

For all these reasons, I think that the paper should be rejected and resubmitted with a better projective prediction approach for complete variable selection problems.

Additionally, and as an interested reader in the application of the predictive prediction approach to real problems, I would like that, if possible, the authors answer to the following questions:

1) My main concern with respect to the projective predictive approach is the robustness of the results to the reference model. Apparently, you can build very different references models to the same data set: have you study the sensitivity of the results with different reference models?. You use Bayesian models as reference models and, in these cases, the determination of prior distribution is not very clear. You can build hundreds and hundreds of prior distribution: for instance in (7) you can have use standard inverted gamma priors on $\tau^2$ and $\sigma^2$ by determining the parameters of the prior in order to the prior mixture gives reasonable results about the data using the prior mixture. How can you study the robustness of your results to this important aspect?

2) Perhaps a solution to the previous problem would be to determine the reference model using frequentist MLE procedures that avoid the construction of prior distributions and use a single point prediction or, alternatively, to use the sampling distribution of MLE as a posterior distribution and take this “Bayesian” model as reference model. Do you have any experience in this line?

3) What if the true generator process is a mixture model, in such a way that there is not a unique solution to the minimal subset variable selection problem? In fact, how can you assure that the minimal subset variable selection problems have only one selection? I think that the algorithm described in appendix A should take into account this possibility, and the step 5 not to choose the smallest (model) that is sufficient close to the reference model’s predictive utility score but to use, for instance, an Occam window, that reflects more adequately the uncertainty associated with the model selection process something that, apparently, your methodology does not consider. In fact, and even though in your examples your results in the paper were not sensitive to the specific choice of how “sufficiently close” is defined, I guess that there can be situations where this statement is not true, and more of one solution to the minimal subset variable selection problem and to the complete variable selection problem could exist.

4) In most of real situations, we have an M-open problem (particularly in big data problems) where the true generator mechanism is too complex and it is not included in the family of considered models. What would be the performance of your methodology if the reference model is far from the true generator process? How can you detect these situations?

5) In big data problems with n << p and the explanatory variables are weakly related (in such a way that the number of significant principal components is large), how do you build a reference model? Is it possible?

6) The same question that 5) but if n is small and your data have not power to discriminate between two alternative reference models.

**Minor comments**

1) Section 3.1: you say that in this simulation you use a Bayesian stepwise forward selection procedure. Is this really the case? In your description of the procedure in page

6 you exclude (and not include variables) in this step of the algorithm. In my view, this procedure is a Bayesian stepwise backward selection procedure.

2) In the Bayesian stepwise selection procedures of section 3.1 you use posterior pvalues to exclude variables, and you exclude the variable with the highest Bayesian p-value. Shouldn't you take into account whether the value of the pvalue is large or small? Shouldn't you calibrate these pvalues?

3) Section 3.1 steplm procedure: Why have you chosen the AIC criterion which, as it is well-known tend to select non-parsimonious models? Have you tried to use the BIC criterion instead? In my experience, this criterion tend to select more parsimonious models and in a M Complete approach, it is usually consistent.

4) Section 4.1, page 8, line 38: $eld_{base}$ should be $eld_{best}$

5) How is calculated the probability used in the stopping rule (2)? Using the reference model?

6) It is not very exigent in (2) to impose that $elpd_i - elpd_{best} > 0$? Perhaps you should consider other limits similar to those used in Bayes factors (Kass and Raftery, 1995) and not to choose only on solution but to explore other possible solutions in the spirit of Occam's window

7) Section 4.1: It is no clear how the value of $\alpha$ should be chosen. $\alpha = 0.16$ is a magic number and a more thoroughly discussion of the influence of this value should be provided.

8) Section 4.1, line 50 "*expect*" should be "*except*"