State of the CITRUS Project

Max Welz
welz@ese.eur.nl

Econometric Institute
Erasmus School of Economics

October 13, 2021

1 Predictive Models

1.1 Risk Modeling

I have implemented the risk modeling as in Kent et al. (2020) for both an ordinary and survival setup, where the latter is based on a completing risk Cox proportional subdistribution hazard model (Fine and Gray, 1999), which generalizes the famous approach of Cox (1972) to multiple causes of failure. Below a summary.

- Implementation of the first stage, both for the ordinary and survival setup, has been straightforward, as existing software for regularized regression with an elastic net penalty (Zou and Hastie, 2005) could be used (`glmnet` and `fastcmprsk` packages). Note that no hierarchical penalty as in Bien et al. (2013) is required here, since the first stage neither includes interaction effects nor a treatment assignment variable.

- We cannot guarantee accurate inference on the coefficients of a risk model. While reliable estimates of the coefficients can be obtained, their confidence intervals will be too narrow due to the additional estimation uncertainty from the first stage. This problem is related to post-selection inference and inference in multi-stage models. Thus, obtaining reliable confidence intervals in risk models (in particular for survival models) is an interesting area of further (technical) research, and beyond the scope of CITRUS. Luckily, for the purposes of CITRUS, point estimates of the coefficients suffice.

- Risk models (especially in a survival setup!) are arguably not robust. We could use a robust regularized Median-of-Means-type estimator (Lecué and Lerasle, 2020) for the first stage and robust logistic regression in the second stage in the ordinary setup, but I'm not sure how well these will work in the survival setup due to a lack of literature. Again, further research is required here, which is beyond the scope of CITRUS. We could easily show that risk models are not robust to situations where the data are not identically distributed and outline this as a problem in CITRUS, but I think that offering a methodological solution is beyond the scope of this project (but still an interesting area of further research!).

- The issue with the discrepancy between the model description in Rekkas et al. (2019) and Kent et al. (2020) has been solved.
- Overall, we are ready to run the risk models.

1.2 Effect Modeling

I am currently working on effect modeling as in Kent et al. (2020) with a hierarchical penalty as in Bien et al. (2013), both in an ordinary and survival context. Luckily, I found the useful reference Du et al. (2021), which adapts the method of Bien et al. (2013) to the analysis of treatment effect heterogeneity, and I am currently implementing their proposed method. Below a summary.

- We have the same situation for inference in effect modeling as for risk modeling, namely that there does not yet exist a method for valid inference, hence this is an area of further research that is beyond the scope of CITRUS. Luckily, for the purposes of CITRUS, point estimates of the coefficients suffice.
- Effect models have the same robustness limitations as risk models.
- We are almost ready to run effect models, I should have an implementation by early next week.

2 Considered models

- one-variable-at-a-time analysis (can only estimate the ARTE);
- risk models and effect models, both with and without survival, as in Kent et al. (2020);
- causal random forest, both with and without survival (Athey et al., 2019 and Cui et al., 2021);
- Double Machine Learning (Chernozhukov et al., 2018) (can only estimate the ATE);
- Generic Machine Learning (Chernozhukov et al., 2020) (give qualitative evaluation of presence of treatment effect heterogeneity).

3 Issues that we want to model

- Ability to capture nonlinear heterogeneity;
- Small samples;

- Improper randomization and unequal randomization (e.g.. 2-1 randomization);
- Robustness against false positives;
- Effects of missing data (demonstrate which imputation techniques are appropriate and which ones aren't; e.g. Valdiviezo and Van Aelst, 2015) and extreme observations (heavy smokers; e.g. under or overreporting of smoking behavior);
- Low representation of risk factors. This means that we only have noisy information on important predictors. (which strictly speaking violates the unconfoundedness assumption).
- Variable transformation (some variables always get transformed in medicine)
- Different type of variables: rating-scale, continuous, categorical, and their encoding. And mix thereof.
- Categorization of continuous variables.
- Added value of survival models, vilation of proportional hazards assumption (e.g. Stensrud and Hernán, 2020).
- For the simulation design, we might want to get inspiration from Knaus et al. (2021).

4 Personal Notes

It would be useful to have a self-written code of the optimizatition routines in Friedman et al. (2010) and Simon et al. (2011), also for later projects in this PhD. I could implement them in R, but R will probably be too slow. Hence, we should consider C/C++ for this purpose. I however lack the experience in C/C++ in such applications, so it would probably take me a few days to get this to run.

If we want to take a deeper look into inference, Guo and He (2021) seems useful.

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2):1148–1178.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111 – 1141.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2020). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint: arXiv1712.04802*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2021). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv preprint: arXiv2001.09887*.

Du, Y., Chen, H., and Varadhan, R. (2021). Lasso estimation of hierarchical interactions for analyzing heterogeneity of treatment effect. *Statistics in Medicine*. In press.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Guo, X. and He, X. (2021). Inference on selected subgroups in clinical trials. *Journal of the American Statistical Association*, 116(535):1498–1506.

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*, 172(1):35–45.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161.

Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906 – 931.

Rekkas, A., Paulus, J. K., Raman, G., Wong, J. B., Steyerberg, E. W., Rijnbeek, P. R., Kent, D. M., and van Klaveren, D. (2019). Predictive approaches to heterogeneous treatment effects: a systematic review. *medRxiv preprint: medRxiv:19010827*.

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5):1–13.

Stensrud, M. J. and Hernán, M. A. (2020). Why test for proportional hazards? *Journal of the American Medical Association*, 323(14):1401–1402.

Valdiviezo, H. C. and Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, 311:163–181.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.