Online Appendix

Description of the Deletion/Substitution/Addition algorithm

The DSA routine implements a general data-adaptive estimation procedure based on cross-validation and the L2 loss function. The final estimator is selected from a set of candidate estimators defined with polynomial generalized linear models generated by the Deletion/Substitution/Addition (D/S/A) algorithm. The space of candidate estimators is parameterized with four variables: 'maxsize', 'maxorderint', 'maxsumofpow' and 'rank.cutoffs'. The final model returned minimizes the empirical risk on the learning set among all estimators considered and characterized by the "optimum" size, order of interactions and set of candidate variables selected by cross-validation.

The D/S/A algorithm is an aggressive model search algorithm which iteratively generates polynomial generalized linear models based on the existing terms in the current 'best' model and the following three steps: 1) a deletion step which removes a term from the model, 2) a substitution step which replaces one term with another, and 3) an addition step which adds a term to the model. The search for the 'best' estimator starts with the base model specified with 'formula': typically the intercept model except when the user requires a number of terms to be forced in the final model.

The search for the 'best' estimator is limited by four user-specified arguments: 'maxsize', 'maxorderint', 'maxsumofpow' and 'rank.cutoffs'. The first argument limits the maximum number of terms in the models considered (excluding the intercept). The second argument limits the maximum order of interactions for the models considered. All terms in the models considered are composed of interactions of variables raised to a given

power. The third argument limits the maximum sum of powers in each term. The fourth argument limits the set of candidate variables to be considered in each model. Only the variables whose ranks specified by 'candidate.rank' are smaller than the threshold(s) specified by 'rank.cutoffs' are allowed in the models considered (and, if applicable, the variable(s) forced in the model). Note that the default ranking of the candidate variables in 'data' is based on their univariate association with the independent variable(s) and obtained with the routine 'candidateReduction'.

This data-adaptive estimation procedure allows comparison of models based on different number of observations and can account for informative censoring through the use of weights in each regression. These weights are provided to the DSA routine with the argument 'weights'. The DSA routine currently supports data-adaptive estimation for continuous or binomial outcomes. When the outcome is binomial, the estimators considered are based on polynomial generalized linear models where the link function is the logit function. Factors can be candidate variables with the caveat that there are currently limitations to the use of factors in terms forced in the final model (see 'formula' above).

The default cross-validation splitting scheme is v-fold where the value for v is specified with the argument 'vfold'. The DSA routine performs the data splits based on the value for 'vfold', 'nsplits', 'id' and the 'userseed' arguments. The argument'nsplits' specifies the number of v-fold splits, e.g. if 'nsplits=2' and 'vfold=5' then the data is split twice based on the 5-fold splitting scheme and thus the DSA call relies on 10 data splits. The argument 'id' identifies the independent experimental units in the data and ensures that the training and validation sets are independent. The argument 'userseed' is

used to set the seed of the R random number generators with the routine 'set.seed()'. This allows for reproducible results. The user can specify an alternative cross-validation splitting scheme with the argument 'usersplits'.

Author(s):

Romain Neugebauer and James Bullard based on the original C code from Sandra Sinisi.

References:

1. Mark J. van der Laan and Sandrine Dudoit, "Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples" (November 2003). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130. http://www.bepress.com/ucbbiostat/paper130

2. Mark J. van der Laan, Sandrine Dudoit, and Aad W. van der Vaart, "The Cross-Validated Adaptive Epsilon-Net Estimator" (February 2004). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 142. http://www.bepress.com/ucbbiostat/paper142

3. Sandra E. Sinisi and Mark J. van der Laan, "Loss-Based Cross-Validated Deletion/Substitution/Addition Algorithms in Estimation" (March 2004). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 143. http://www.bepress.com/ucbbiostat/paper143