The Journal of Open Source Software

robustHD: An R package for robust regression with high-dimensional data

Andreas Alfons*[1]

1 Erasmus School of Economics, Erasmus University Rotterdam, Netherlands

DOI: 10.21105/joss.03786

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @valentint
- @msalibian

Submitted: 29 September 2021
Published: 25 October 2021

Summary

In regression analysis with high-dimensional data, variable selection is an important step to (i) overcome computational problems, (ii) improve prediction performance by variance reduction, and (iii) increase interpretability of the resulting models due to the smaller number of variables. However, robust methods are necessary to prevent outlying data points from distorting the results. The add-on package `robustHD` (Alfons, 2021) for the statistical computing environment R (R Core Team, 2021) provides functionality for robust linear regression and model selection with high-dimensional data. More specifically, the implemented functionality includes robust least angle regression (Khan et al., 2007), robust groupwise least angle regression (Alfons et al., 2016), as well as sparse least trimmed squares regression (Alfons et al., 2013). The latter can be seen as a trimmed version of the popular lasso regression estimator (Tibshirani, 1996). Selecting the optimal model can be done via cross-validation or an information criterion, and various plots are available to illustrate model selection and to evaluate the final model estimates. Furthermore, the package includes functionality for pre-processing such as robust standardization and winsorization. Finally, `robustHD` follows a clear object-oriented design and takes advantage of C++ code and parallel computing to reduce computing time.

Statement of need

While the authors of Khan et al. (2007) did provide an R script for robust least angle regression (although the link in their paper appears to be broken), the implementation in `robustHD` utilizes C++ code for computational efficiency and provides the convenience of an R package. Moreover, code for robust groupwise least angle regression and sparse least trimmed squares regression is not available elsewhere. Package `robustHD` therefore provides researchers with access to several popular methods for robust regression and variable selection with high-dimensional data. It has been used in many benchmarking studies in the statistical literature (e.g., Chang et al., 2018; Cohen Freue et al., 2019; Kurnaz et al., 2018b; Smucler & Yohai, 2017), as well as in empirical research (e.g., Antczak-Orlewska et al., 2021; Stadlbauer et al., 2020).

Example: Robust groupwise least angle regression

Robust least angle regression (Khan et al., 2007) and robust groupwise least angle regression (Alfons et al., 2016) follow a hybrid model selection strategy: first obtain a sequence of

*Corresponding author

important candidate predictors, then fit submodels along that sequence via robust regressions. Here, data on cars featured in the popular television show *Top Gear* are used to illustrate this functionality.

The response variable is fuel consumption in miles per gallon (MPG), with all remaining variables used as candidate predictors. Information on the car model is first removed from the data set, and the car price is log-transformed. In addition, only observations with complete information are used in this illustrative example.

```
# load package and data
library("robustHD")
data("TopGear")

# keep complete observations and remove information on car model
keep <- complete.cases(TopGear)
TopGear <- TopGear[keep, -(1:3)]
# log-transform price
TopGear$Price <- log(TopGear$Price)
```

As the *Top Gear* data set contains several categorical variables, robust groupwise least angle regression is used. Through the formula interface, function `rgrplars()` by default takes each categorical variable (`factor`) as a group of dummy variables while all remaining variables are taken individually. However, the group assignment can be defined by the user through argument `assign`. The maximum number of candidate predictor groups to be sequenced is determined by argument `sMax`. Furthermore, with `crit = "BIC"`, the optimal submodel along the sequence is selected via the Bayesian information criterion (BIC). Note that each submodel along the sequence is fitted using a robust regression estimator with a non-deterministic algorithm, hence the seed of the random number generator is supplied for reproducibility.

```
# fit robust groupwise least angle regression and print results
fit <- rgrplars(MPG ~ ., data = TopGear, sMax = 15,
                crit = "BIC", seed = 20210507)
fit
```

```
## Call:
## rgrplars(formula = MPG ~ ., data = TopGear, sMax = 15, crit = "BIC",
##     seed = 20210507)
##
## Sequence of moves:
##       1 2 3 4  5 6  7  8 9 10 11 12 13 14 15
## Group 6 4 8 1 10 5 12 13 9 18 27 16 28 17 26
##
## Coefficients of optimal submodel:
##     (Intercept)      FuelPetrol    Displacement DriveWheelFront
##   149.512783142   -12.905795146    -0.003968404     5.374692058
##  DriveWheelRear             BHP    Acceleration        TopSpeed
##     0.612416948     0.015265327     0.530554736    -0.191667161
##          Weight           Width          Height
##    -0.001817037    -0.013641306    -0.029560052
##
## Optimal step: 9
```

The output prints information on the sequence of predictor groups, as well as the results of the final model fit. Here, 9 predictor groups consisting of 10 individual covariates are selected into the final model.

Several plots are available for the results: `coefPlot()` visualizes the coefficient path along the sequence of submodels, `critPlot()` plots the values of the optimality criterion against the step along the sequence, and `diagnosticPlot()` allows to produce various diagnostic plots for the final model fit. Examples of these plots are shown in Figure 1.

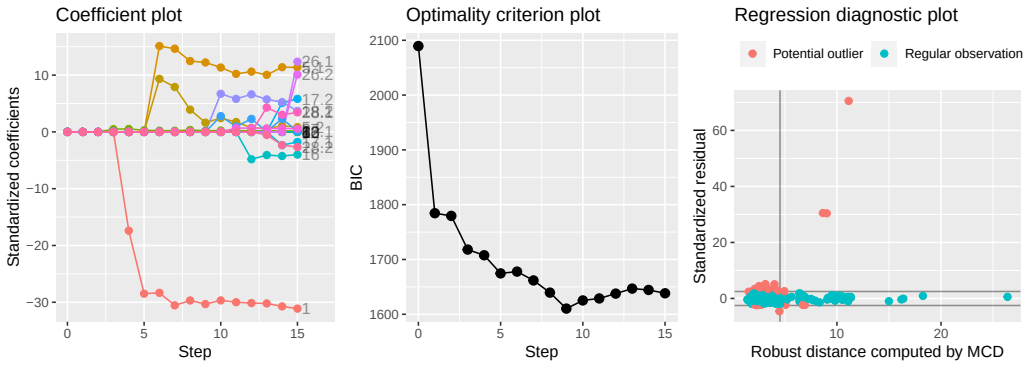

Figure 1: Examples of the coefficient plot (*left*), the optimality criterion plot (*center*), and the regression diagnostic plot (*right*) for output of function `rgrplars()` in package `robustHD`.

Example: Sparse least trimmed squares regression

The well-known NCI-60 cancer cell panel (Reinhold et al., 2012) is used to illustrate the functionality for sparse least trimmed squares regression. The protein expressions for a specific protein are selected as the response variable, and the gene expressions of the 100 genes that have the highest (robustly estimated) correlations with the response variable are screened as candidate predictors.

```
# load package and data
library("robustHD")
data("nci60")  # contains matrices 'protein' and 'gene'

# define response variable
y <- protein[, 92]
# screen most correlated predictor variables
correlations <- apply(gene, 2, corHuber, y)
keep <- partialOrder(abs(correlations), 100, decreasing = TRUE)
X <- gene[, keep]
```

Sparse least trimmed squares is a regularized estimator of the linear regression model, whose results depend on a non-negative regularization parameter (see Alfons et al., 2013). In general, a larger value of this regularization parameter yields more regression coefficients being set to zero, which can be seen as a form of variable selection.

For convenience, `sparseLTS()` can internally estimate the smallest value of the regularization parameter that sets all coefficients to zero. With `mode = "fraction"`, the values supplied via the argument `lambda` are then taken as fractions of this estimated value (i.e., they are multiplied with the internally estimated value). In this example, the optimal value of the the regularization parameter is selected by estimating the prediction error (`crit = "PE"`) via 5-fold cross-validation with one replication (`splits = foldControl(K = 5, R = 1)`). The default prediction loss function is the root trimmed mean squared prediction error. Finally, the seed of the random number generator is supplied for reproducibility.

```
# fit sparse least trimmed squares regression and print results
lambda <- seq(0.01, 0.5, length.out = 10)
fit <- sparseLTS(X, y, lambda = lambda, mode = "fraction", crit = "PE",
                 splits = foldControl(K = 5, R = 1), seed = 20210507)
fit

## Final model:
##
## Call:
## sparseLTS(x = X, y = y, lambda = 0.0448151412422958)
##
## Coefficients:
##  (Intercept)         8502        21786          134         4454
## -3.709498642  0.593549132  0.033366829  0.115955965  0.015899659
##         1106        20125         8510        14785        17400
##  0.020447909 -0.091451556  0.111369625 -0.014556471  0.002262256
##         8460         8120        18447        15622         7696
## -0.003669024  0.112165149 -0.229292900 -0.008785651  0.020915212
##         5550        16784        13547
##  0.005880150  0.015398316  0.037262275
##
## Penalty parameter:       0.04481514
## Residual scale estimate: 0.62751742
```

Among other information (which is omitted above), the output prints the results of the final model fit, which here consists of 17 genes with non-zero coefficients. Similar plots as in Figure 1 are available to visualize the results.

Related software

Package `enetLTS` (Kurnaz et al., 2018a) provides robust elastic-net-regularized estimators based on trimming for the linear and logistic regression models. Package `pense` (Kepplinger et al., 2021) contains implementations of robust S- and MM-type estimators with elastic net regularization for linear regression.

Acknowledgements

Andreas Alfons is partially supported by a grant of the Dutch Research Council (NWO), research program Vidi (project number VI.Vidi.195.141).

References

Alfons, A. (2021). *robustHD: Robust methods for high-dimensional data*. https://github.com/aalfons/robustHD/

Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, *7*(1), 226–248. https://doi.org/10.1214/12-AOAS575

Alfons, A., Croux, C., & Gelper, S. (2016). Robust groupwise least angle regression. *Computational Statistics & Data Analysis*, *93*, 421–435. https://doi.org/10.1016/j.csda.2015.02.007

Antczak-Orlewska, O., Płóciennik, M., Sobczyk, R., Okupny, D., Stachowicz-Rybka, R., Rzodkiewicz, M., Siciński, J., Mroczkowska, A., Krąpiec, M., Słowiński, M., & Kittel, P. (2021). Chironomidae morphological types and functional feeding groups as a habitat complexity vestige. *Frontiers in Ecology and Evolution*, *8*(583831), 1–16. https://doi.org/10.3389/fevo.2020.583831

Chang, L., Roberts, S., & Welsh, A. (2018). Robust lasso regression using Tukey's biweight criterion. *Technometrics*, *60*(1), 36–47. https://doi.org/10.1080/00401706.2017.1305299

Cohen Freue, G. V., Kepplinger, D., Salibian-Barrera, M., & Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *The Annals of Applied Statistics*, *13*(4), 2065–2090. https://doi.org/10.1214/19-AOAS1269

Kepplinger, D., Salibián-Barrera, M., & Cohen Freue, G. (2021). *pense: Penalized elastic net S/MM-estimator of regression*. https://CRAN.R-project.org/package=pense

Khan, J. A., Van Aelst, S., & Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, *102*(480), 1289–1299. https://doi.org/10.1198/016214507000000950

Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018a). *enetLTS: Robust and sparse methods for high dimensional linear and logistic regression*. https://CRAN.R-project.org/package=enetLTS

Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018b). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, *172*, 211–222. https://doi.org/10.1016/j.chemolab.2017.11.017

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J., & Pommier, Y. (2012). CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Research*, *72*(14), 3499–3511. https://doi.org/10.1158/0008-5472.can-12-1370

Smucler, E., & Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, *111*, 116–130. https://doi.org/10.1016/j.csda.2017.02.002

Stadlbauer, V., Engertsberger, L., Komarova, I., Feldbacher, N., Leber, B., Pichler, G., Fink, N., Scarpatetti, N., Schippinger, W., Schmidt, R., & Horvath, A. (2020). Dysbiosis, gut barrier dysfunction and inflammation in dementia: A pilot study. *BMC Geriatrics*, *20*(248), 1–13. https://doi.org/10.1186/s12877-020-01644-2

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x