Introduction to the R code accompanying "De-biased Lasso for Generalized Linear Models with A Diverging Number of Covariates" by Xia, Nan and Li

1 Description

This file provides introduction to the two R code files in the Supporting Information. The R code file `DBL_GLMs_functions.R` provides two functions for inference in generalized linear models (GLMs), one for implementing the proposed de-biased lasso approach by directly inverting the Hessian matrix, and the other for implementing the original de-biased lasso approach (van de Geer et al., 2014). The R code file `DBL_GLMs_example.R` provides some example code to compare the aforementioned two de-biased lasso methods and the maximum likelihood estimation (MLE) with simulated data.

2 Functions in `DBL_GLMs_functions.R`

2.1 Function `REF_DS_inf()`

This function implements the proposed refined de-biased lasso approach by directly inverting the Hessian matrix.

The input parameters are:

- `x`: The covariate matrix, with observations in rows and covariates in columns. A column of 1's will be automatically added, so users do not need to supply the intercept in `x`;
- `y`: The vector of responses;
- `family`: Can be either "binomial" for logistic regression, or "poisson" for Poisson regression;
- `lasso_est`: The lasso estimates for the regression coefficients, which can be obtained using R package `glmnet` prior to the de-biasing step and must be of the same length as `ncol(x)+1`.

The following values are returned:

- `est`: The refined de-biased lasso estimates for the regression coefficients, where the first element corresponds to the intercept;
- `se`: The model-based standard error estimates for `est`;
- `pvalue`: The p-values for two-sided tests of whether each coefficient is zero based on the proposed de-biased lasso approach;
- `theta`: The inverse of Hessian matrix $\widehat{\Theta} = \widehat{\Sigma}_{\widehat{\xi}}^{-1}$.

2.2 Function `ORIG_DS_inf()`

This function implements the original de-biased lasso approach (van de Geer et al., 2014) that exploits the node-wise lasso estimation for the inverse information matrix approximation. The node-wise lasso is implemented using R package `glmnet` with cross-validation.

The input parameters are:

- `x`: The covariate matrix, with observations in rows and covariates in columns. A column of 1's will be automatically added, so users do not need to supply the intercept in `x`;
- `y`: The vector of responses;
- `family`: Can be either "binomial" for logistic regression, or "poisson" for Poisson regression;
- `lasso_est`: The lasso estimates for the regression coefficients, which can be obtained using R package `glmnet` prior to the de-biasing step and must be of the same length as `ncol(x)+1`;
- `nfold`: The number of folds for cross-validation when using node-wise lasso to estimate each row of $\widetilde{\Theta}$;
- `n_lambda`: The number of tuning parameter lambda values for cross-validation when using node-wise lasso to estimate each row of $\widetilde{\Theta}$;
- `lambda_ratio`: The ratio between the smallest and the largest lambda values for cross-validation when using node-wise lasso to estimate each row of $\widetilde{\Theta}$, which is used to generate the sequence of `n_lambda` lambda values.

The following values are returned:

- `est`: The original de-biased lasso estimates for the regression coefficients, where the first element corresponds to the intercept;
- `se`: The model-based standard error estimates for `est`;

- `pvalue`: The p-values for two-sided tests of whether each coefficient is zero based on the original de-biased lasso approach;
- `theta`: The node-wise lasso estimate $\widetilde{\Theta}$ for the inverse information matrix.

3 A simulated example in `DBL_GLMs_example.R`

A dataset with $n = 500$ observations and $p = 100$ covariates is generated in a logistic regression setting. Covariates follow a multivariate normal distribution with mean zero and AR(1) covariance matrix ($\rho = 0.7$), and then truncated at ± 6. The lasso estimator is first obtained using `glmnet`, and then supplied to `REF_DS_inf()` and `ORIG_DS_inf()` for de-biased lasso inferences. 95% confidence intervals for regression coefficients are calculated. For comparison, we also include the MLE fitted by `glm.fit`. Please refer to the code for more detailed notes.

References

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.