TEV Vignette

Hua Yun Chen
Division of Epidemiology and Biostatistics,
School of Public health, University of Illinois Chicago.

1 Introduction

TEV (Total Explained Variation) is an R package developed by Hua Yun Chen for estimating and inferring the proportion of the variation of an outcome explained by a set of high-dimensional covariates in a linear model. The approach estimates the proportion of the explained variation of outcome Y explained by covariates $X_1, \cdots, X_p$, i.e.,

$$r^2 = \frac{\text{var}\left\{E(Y \mid X_1, \cdots, X_p)\right\}}{\text{var}(Y)}. \tag{1}$$

in a linear model by

$$\hat{r}^2_\lambda(\Omega) = \frac{tr\left[W_\lambda(\Omega)\left\{\hat{\sigma}_Y^{-2}(Y - \bar{Y}\mathbf{1})(Y - \bar{Y}\mathbf{1})^\top - (I_n - \mathbf{1}\mathbf{1}^\top/n)\right\}\right]}{tr\left[W(\Omega)\left\{M(\Omega) - (I_n - \mathbf{1}\mathbf{1}^\top/n)\right\}\right]}, \tag{2}$$

where Ω is the inverse of the covariance matrix of the covariates, M and W are matrix function of covariates, and $\hat{\sigma}^2_Y$ is the empirical estimate of the variance of Y. The package also estimates the explained variation

$$\sigma_s^2 = \text{var}\left\{E(Y \mid X_1, \cdots, X_p)\right\}$$

by $\hat{\sigma}_s^2 = \hat{\sigma}_Y^2 \hat{r}^2_\lambda(\Omega)$.

Three functions are developed for implementing the estimation and inference approach. Depending on the data you have, different functions need to be called. There are three different situations.

1. If covariates are assumed independent, call `"RVee"`.
2. If covariates are assumed dependent, and the sample size $n > p$, call `"RVls"`.
3. If covariates are assumed dependent, and supplementary covariate data are available such that the total number of covariate data sample size is greater than p, call `"RVsd"`.

If your data do not fall into one of the three cases, i.e., covariates are dependent with unknown correlation matrix and $n \leq p$ without supplementary covariate data, the package TEV cannot be used.

2 Installation

If you do not have the package "devtools" installed in R, you can run the following command in R to install it.

```
install.packages("devtools")
```

Once the package "devtools" is installed or if you had installed the package `devtools` in your computer previously, run

```
library(devtools)
```

to include the package in your current R session. You can now install the development version of TEV from github with:

```
devtools::install_github("hychen-uic/TEV")
```

3 Example data

The package has an internal data set called `"example.rda"`. The data set has 3262 records, 107 variables. Variables 1 to 7 are outcomes, variables 8 to 19 are confounders, variables 20 to 107 are exposures. Outcome data are available for the first 1,000 records, and are not available for the rest records.

```
n=1000

N=2000

y=as.vector(example[1:n,1])

cfd=as.matrix(example[1:n, 8:19])

x=as.matrix(example[1:n,20:107])

CFD=as.matrix(example[(n+1):(n+N), 8:19])

xsup=as.matrix(example[(n+1):(n+N), 20:107])
```

Outcomes are in `y`, confounders to be adjusted are in `cfd`, exposures are in `x`, and supplementary confounders data are in `CFD`, and supplementary exposures are in `X`.

4 Data analysis without confounder adjustment

4.1 Independent exposures and no supplementary data

Suppose that we do not have supplementary data. We use the first y component as the outcome and x as the exposure, we can estimate the proportion of the explained variation and the explained variation by exposures x by

```
RVee(y, x, alpha=0.05)

[[1]]

[1] 0.176380930 0.001846638 0.001817497
```
Estimates of the proportion (r^2) of variance explained by the exposures, estimated variance of $\hat{r}^2$ in general, and under the normal random error assumption.

```
[[2]]

[1] 0.09215633 0.26060553
```
The 95% confidence interval for r^2 in general.

```
[[3]]

[1] 0.09282352 0.25993834
```
The 95% confidence interval for r^2 under the normal random error assumption.

```
[[4]]

[1] 94.92776 820.83561 819.54988
```
Estimates of the explained variance (σ_s^2) by the exposures, estimated variance of $\hat{\sigma}_s^2$ in general, and under the normal random error assumption.

```
[[5]]
[1] 38.77435 151.08118
```
The 95% confidence interval for σ_s^2 in general.

```
[[6]] [1] 38.81834 151.03719
```
The 95% confidence interval for σ_s^2 under the normal random error assumption.

These estimates assume that the exposures are independent and link to the outcome through a linear model.

4.2 Dependent exposures and no supplementary data

If we do not know if the exposures are independent and need to account for the possible dependence of exposures, we can use the least-square approach because $n > p$.

```
RVls(y, x, alpha=0.05)

[[1]]

[1] 0.176380930 0.001846638 0.001817497
```
Estimates of the proportion (r^2) of variance explained by the exposures, estimated

variance of $\hat{r}^2$ in general, and under the normal random error assumption.

```
   [[2]]

   [1] 0.09215633 0.26060553
```
The 95% confidence interval for r^2 in general.

```
   [[3]]

   [1] 0.09282352 0.25993834
```
The 95% confidence interval for r^2 under the normal random error assumption.

```
   [[4]]

   [1] 94.92776 820.83561 819.54988
```
Estimates of the explained variance (σ_s^2) by the exposures, estimated variance of $\hat{\sigma}_s^2$ in general, and under the normal random error assumption.

```
   [[5]]

   [1] 38.77435 151.08118
```
The 95% confidence interval for σ_s^2 in general.

```
   [[6]]

   [1] 38.81834 151.03719
```
The 95% confidence interval for σ_s^2 under the normal random error assumption.

The analysis can also be performed using `RVsd` without the supplementary data as

```
RVsd(y, x, alpha=0.05)
```

This differs from `RVls(y, x, alpha=0.05)` in the variance estimate for the explained variation estimator. This method uses the simulation approach. It takes longer to compute and can be more accurate than that obtained from `RVls(y, x, alpha=0.05)` based on simulation studies.

These estimators are based on the assumption that the exposures and link to the outcome through a linear model.

4.3 Dependent exposures with supplementary data

If we do not know if the exposures are independent and need to account for the possible dependence of exposures, we can include supplementary covariate data in the analysis. *This may take more time to produce results as estimating the variances include Monte Carlo simulation.*

```
   RVsd(y, x, xsup, alpha=0.05)
   [[1]]

   [1] 0.2514390188 0.0007355247 0.0006794820
Estimates of the proportion (r^2) of variance explained by the exposures, estimated
variance of r̂^2 in general, and under the normal random error assumption.

   [[2]]

   [1] 0.1982837 0.3045943
The 95% confidence interval for r^2 in general.

   [[3]]

   [1] 0.2003489 0.3025292
The 95% confidence interval for r^2 under the normal random error assumption.

   [[4]]

   [1] 135.3238 288.7580 287.6783
Estimates of the explained variance (σ_s^2) by the exposures, estimated variance
of σ̂_s^2 in general, and under the normal random error assumption.

   [[5]]

   [1] 102.0184 168.6293
The 95% confidence interval for σ_s^2 in general.

   [[6]]

   [1] 102.0807 168.5669
The 95% confidence interval for σ_s^2 under the normal random error assumption.
```

These estimates assume that the exposures link to the outcome through a linear model. In this particular data set, we have $n > p$. *In general, if $n < p$, we need to have a sample size for the supplementary data, say N, such that $N + n > p$.*

5 Data analysis with confounder adjustment

In inferring the proportion of the explained variation and the explained variation itself by the exposures, we often need to adjust for possible confounders. The way to adjust for the confounders in the TEV package is through projection. The exposures are projected to the orthogonal complementary of the linear space generated by the confounders. The projected data will be used as the exposures in `RVee, RVeels, RVeesd`. TEV has a function to

perform the projection.

Specifically, if no supplementary data are available, the project can be carried out by

```
z=project(cfd, x)[[1]]
```

We then use z in place of x in calling `RVee` and `RVeels` as follows

```
RVee(y, z, alpha=0.05)

[[1]]

[1] 0.174557401 0.001816604 0.001787979

[[2]]

[1] 0.09102052 0.25809428

[[3]]

[1] 0.0916813 0.2574335

[[4]]

[1] 93.94635 801.89718 800.61499

[[5]]

[1] 38.4445 149.4482

[[6]]

[1] 38.48889 149.40380
```

For least-square approach,

```
RVls(y, z, alpha=0.05)

[[1]]

[1] 0.2513327642 0.0006554077 0.0006008460

[[2]]

[1] 0.2011559 0.3015097
```

```
[[3]]

[1] 0.2032898 0.2993757

[[4]]

[1] 135.2666 250.4958 249.3437

[[5]]

[1] 104.2462 166.2871

[[6]]

[1] 104.3176 166.2157
```

If supplementary data are available, the projection is carried out by

```
W=project(rbind(cfd, CFD), rbind(x,X))[[1]]

z=W[1:n,]

zsup=W[n+c(1:N), ]
```

We then use z and Z respectively in place of x and X in calling `RVeesd` as follows

```
RVsd(y, z, zsup, alpha=0.05)

[[1]]

[1] 0.2503042964 0.0007307610 0.0006752166

[[2]]

[1] 0.1973214 0.3032872

[[3]]

[1] 0.1993748 0.3012338

[[4]]

[1] 134.7131 285.9223 284.8436

[[5]]
```

```
[1] 101.5716 167.8546

[[6]]

[1] 101.6342 167.7921
```

References

1. Chen, H. Y. (2022). Statistical inference on explained variation in high-dimensional linear model with dense effects. arXiv:2201.08723.

2. Chen, H.Y., Li, H., Argos, M., Persky, V.W., Turyk, M.E. (2022). Statistical methods for assessing explained variation of a health outcome by mixture of exposures. *Int. J. Environ. Res. Public Health*, 18.

3. Chen, H. Y., Zhang, B., and Pan, G (2023). Estimation and inference on explained variation with possible supplementary covariate data. Manuscript.