CIMBAL Software Documentation

March 17, 2022

Version 0.7
Date March 17, 2022
Title Meta-analysis method for confounder imbalance
Correspondence Debashree Ray, Ph.D. `<dray@jhu.edu>`
Description CIMBAL leverages asymptotic relations between adjusted and unadjusted effect estimates to impute adjusted summary statistics (effect estimate and its corresponding standard error) for a cohort that reported only unadjusted estimates. To do this, CIMBAL borrows information on unadjusted and adjusted estimates from cohorts that report both. CIMBAL finally provides the appropriately meta-analyzed adjusted estimates across all cohorts either reporting unadjusted estimates only or both.
Depends R (>= 3.0.1)

`cimbal` *Meta-analysis of cohorts with confounder imbalance after imputing adjusted summary statistics for those reporting unadjusted summary statistics only.*

Description

CIMBAL uses adjusted and unadjusted summary statistics– effect estimate and its standard error (SE)– from cohorts with measurement on the same set of confounders to impute adjusted summary statistics for all cohorts without any confounder measurement. Finally, it meta-analyzes the available adjusted estimates and the imputed adjusted estimates in a fixed-effect meta-analysis framework after appropriately accounting for the dependence between estimates arising due to borrowing of information across cohorts. While we described CIMBAL in the context of cohort studies with disparate confounder measurements, it is also relevant for a meta-analysis of randomized controlled trials where imbalance in measuring the effect modifiers across trials is prevalent.

Usage

```
cimbal(dat, mute.msgs=FALSE, ncohort.thresh=25)
```

Arguments

`dat` A data frame consisting of K rows (corresponding to K studies to be meta-analyzed) and at least 6 columns. The required 6 columns to be labeled: ‘cohort’ (cohort identifier), ‘samplesize’ (sample size of a given cohort), ‘b.unadj’ (unadjusted effect estimate), ‘se.unadj’ (unadjusted SE estimate), ‘b.adj’ (adjused effect estimate), ‘se.adj’ (adjusted SE estimate). All cohorts in this data frame must have non-missing unadjusted effect and SE estimates. Cohorts with missing adjusted estimates should report NA under columns ‘b.adj’ and ‘se.adj’.

`mute.msgs` Logical; if `TRUE`, different messages during CIMBAL implementation are muted. Default value is `FALSE`.

`ncohort.thresh` Minimum number of complete cohorts to use for estimating covariance between unadjusted and adjusted effect estimates. This covariance is needed to impute the adjusted SE estimate of incomplete cohort and also for computing the appropriate weight to be used in the meta-analysis step. Default value is 25.

Details

Consider the following measurements from an epidemiologic study: response or outcome Y, exposure X and a set of q possible confounders $\boldsymbol{\mathcal{C}}$. To assess exposure-outcome association, a cohort with no information on any confounder will consider an unadjusted model

$$Y = \alpha_{\text{unadj}} + \beta_{\text{unadj}} X + \epsilon_u,\ \epsilon_u \sim N(0, \sigma^2_{\text{unadj}}) \qquad \text{(if continuous response)}$$

$$\text{logit}\left(P(Y=1)\right) = \alpha_{\text{unadj}} + \beta_{\text{unadj}} X \qquad \text{(if binary outcome)}$$

and report unadjusted estimate of association ($\hat{\beta}_{\text{unadj}}$) and its SE ($\hat{\text{se}}_{\text{unadj}}$). On the other hand, a cohort that measured all confounders in $\boldsymbol{\mathcal{C}}$ will consider an adjusted model

$$\begin{aligned} Y &= \alpha_{\text{adj}} + \beta_{\text{adj}} X + \gamma' \boldsymbol{\mathcal{C}} + \epsilon_a,\ \epsilon_a \sim N(0, \sigma^2_{\text{adj}}) && \text{(if continuous response)} \\ \text{logit}\left(P(Y=1)\right) &= \alpha_{\text{adj}} + \beta_{\text{adj}} X + \gamma' \boldsymbol{\mathcal{C}} && \text{(if binary outcome)} \end{aligned}$$

and report adjusted estimate of association ($\hat{\beta}_{\text{adj}}$) and its SE ($\hat{\text{se}}_{\text{adj}}$). Let us assume there are K cohorts, of which K_c cohorts have complete information on confounders while K_m cohorts have no confounder information. Thus, one can gather the unadjusted estimates of exposure-outcome association from all cohorts $\left\{\left(\hat{\beta}_{j,\text{unadj}}, \hat{\text{se}}_{j,\text{unadj}}\right),\ j = 1, 2, ..., K\right\}$ and the adjusted estimates from the complete cohorts $\left\{\left(\hat{\beta}_{j,\text{adj}}, \hat{\text{se}}_{j,\text{adj}}\right),\ j = 1, 2, ..., K_c\right\}$.

For the ease of exposition, let us first consider only two cohorts and later generalize CIMBAL's pipeline for multiple cohorts. If there are two independent cohorts– cohort 1 with no information on the confounder and is able to report only $\hat{\beta}_{\text{unadj}}$, and cohort 2 with complete information to be able to report both $\hat{\beta}_{\text{adj}}$ and $\hat{\beta}_{\text{unadj}}$– the investigator can impute the adjusted association estimate for the cohort with missing confounder information as $\tilde{\beta}_{1,\text{adj}} = \hat{\beta}_{2,\text{adj}} - \hat{\beta}_{2,\text{unadj}} + \hat{\beta}_{1,\text{unadj}}$. This imputation approach is based on the observation that the difference in effect estimates is asymptotically independent of the sample size (Ray et al., 2022). The adjusted SE estimate for cohort 1 may, then, be obtained from first principles as $\tilde{\text{se}}^2_{1,\text{adj}} = \hat{\text{se}}^2_{2,\text{adj}} + \hat{\text{se}}^2_{2,\text{unadj}} + \hat{\text{se}}^2_{1,\text{unadj}} - 2\text{Cov}(\hat{\beta}_{2,\text{unadj}}, \hat{\beta}_{2,\text{adj}})$. Although cohorts 1 and 2 are independent, the imputed estimate $\tilde{\beta}_{1,\text{adj}}$ and the available estimate $\hat{\beta}_{2,\text{adj}}$ are not uncorrelated due to borrowing of information between cohorts. As a result, the inverse-variance weights are no longer optimal for a fixed-effect meta-analysis. In fact, the appropriate meta-analysis of our imputed adjusted estimate ($\tilde{\beta}_{1,\text{adj}}$) from one cohort and the available adjusted estimate ($\hat{\beta}_{2,\text{adj}}$) from another cohort as implemented by CIMBAL is given by $\hat{\beta}_{\text{adj}} = \hat{w}_1 \tilde{\beta}_{1,\text{adj}} + \hat{w}_2 \hat{\beta}_{2,\text{adj}}$ where the optimal weights minimizing the variance of $\hat{\beta}_{\text{adj}}$ are $\hat{w}_1 = \frac{\text{Cov}(\hat{\beta}_{2,\text{unadj}}, \hat{\beta}_{2,\text{adj}})}{\hat{\text{se}}^2_{1,\text{unadj}} + \hat{\text{se}}^2_{2,\text{unadj}}}$ and $\hat{w}_2 = 1 - \hat{w}_1$, and the variance achieves its minimum at $\hat{\text{se}}^2_{\text{adj}} = \hat{\text{se}}^2_{2,\text{adj}} - \frac{\text{Cov}(\hat{\beta}_{2,\text{unadj}}, \hat{\beta}_{2,\text{adj}})^2}{\hat{\text{se}}^2_{1,\text{unadj}} + \hat{\text{se}}^2_{2,\text{unadj}}}$.

If there are $K_c (> 1)$ complete cohorts and $K_m (> 1)$ incomplete cohorts, we first meta-analyze these cohorts in each group using fixed-effect inverse-variance weighted meta-analysis: $\left(\hat{\beta}^{(\text{meta})}_{2,\text{adj}}, \hat{\text{se}}^{(\text{meta})}_{2,\text{adj}}\right)$ and $\left(\hat{\beta}^{(\text{meta})}_{2,\text{unadj}}, \hat{\text{se}}^{(\text{meta})}_{2,\text{unadj}}\right)$ from complete cohorts, and $\left(\hat{\beta}^{(\text{meta})}_{1,\text{unadj}}, \hat{\text{se}}^{(\text{meta})}_{1,\text{unadj}}\right)$ from incomplete cohorts. Now we can apply the afore-mentioned imputation approach to obtain CIMBAL-imputed adjusted estimates for the pooled no-confounder cohort $\left(\tilde{\beta}^{(\text{meta})}_{1,\text{unadj}}, \tilde{\text{se}}^{(\text{meta})}_{1,\text{unadj}}\right)$. For the final meta-analysis of all cohorts, we first estimate $\text{Cov}(\hat{\beta}_{2,\text{unadj}}, \hat{\beta}_{2,\text{adj}})$ from the K_c complete cohorts and then use the above meta-analysis formulae.

For more details on how and when CIMBAL may be used, please refer to Ray et al. (2022). We request that this reference be cited if any part of this software is used in any publication.

Value

`dat.imputed` A dataframe consisting of $K_c + 1$ rows and at least 6 columns (‘cohort’, ‘samplesize’, ‘b.unadj’, ‘se.unadj’, ‘b.adj’, ‘se.adj’). The rows correspond to the K_c complete cohorts and the meta-analyzed incomplete cohort. The ‘b.unadj’ and ‘se.unadj’ columns for the meta-analyzed incomplete cohort provide the inverse-variance weighted fixed-effect meta-analysis of unadjusted estimates, while the ‘b.adj’ and ‘se.adj’ columns provide the CIMBAL-imputed adjusted estimates.

`beta.meta.cimbal` Overall effect estimate from meta-analyzing complete and incomplete cohorts using CIMBAL.

`se.meta.cimbal` Overall SE estimate corresponding to the effect estimate `beta.meta.cimbal`.

Reference

Ray, D., Muñoz, A., Zhang, M., Li, X., Chatterjee, N., Jacobson, L.P. and Lau, B. “Meta-analysis under imbalance in measurement of confounders in cohort studies using only summary-level data”. *In revision.*
(Please check `https://github.com/RayDebashree/CIMBAL` for updated citation.)

Example

```
#-------- Download or directly source CIMBAL
source("CIMBAL_v0.7.R")
# require(devtools)
# source_url("https://github.com/RayDebashree/CIMBAL/blob/master/
CIMBAL_v0.7.R?raw=TRUE")
#--------
```

```
### For an example, let’s first simulate a toy set of summary
### statistics from 60 cohorts on a binary outcome (Y), binary exposure (X)
### and 2 binary confounders (C1, C2)

    # function to simulate data with Y, X, C1 and C2 for sample size n
    getdata <- function(n, a, b, bx, model="glm"){
        # simulate confounders
        c1 <- rbinom(n,1,0.1)
        c2 <- rbinom(n,1,0.6)
        # get the binary exposure with prob px
        px <- 1/(1+exp(-(a[1]+a[2]*c1+a[3]*c2)))
        x <- rbinom(n,1,px)
        # simulate binary response Y if "glm" else normal response
        if(model=="glm"){
            py <- 1/(1+exp(-(b[1]+b[2]*c1+b[3]*c2+bx*x)))
            y <- rbinom(n,1,py)
        }else{
            e <- rnorm(n,0,1)
            y <- b[1]+b[2]*c1+b[3]*c2+bx*x+e
        }
        return(data.frame(y,x,c1,c2))
    }

    # function to get summary statistics from cohorts
    getcohorts <- function(samplesizes, a, b, bx, model, type){
        # no. of cohorts
        K <- length(samplesizes)
        # initialize the data frame of summary stats
        mydat <- as.data.frame(cbind(1:K, samplesizes))
        colnames(mydat) <- c("cohort","samplesize")
        #cohortnames <- sapply(1:K, function(i) paste0("Cohort",i))
        mydat$cohort <- 1:K
        mydat$se.adj <- mydat$b.adj <- mydat$se.unadj <- mydat$b.unadj <- NA
        # initialize the full data
        set.seed(2022)
        for (cohort in 1:K){
            ### simulate the data on cohort of sample size n
            dat1 <- getdata(samplesizes[cohort], a, b, bx)
            ### analysis output of dat1
            if(type=="unadj") outu <- glm(y˜x, data=dat1, family="binomial",
control=list(maxit=1e4))
            if(type=="padj") outu <- glm(y˜x+c1, data=dat1, family="binomial",
control=list(maxit=1e4))
            outa <- glm(y˜x+c1+c2, data=dat1, family="binomial",
control=list(maxit=1e4))
            rm(dat1)
            mydat[cohort,3:6] <- c(coef(summary(outu))[’x’,c(’Estimate’,
’Std. Error’)], coef(summary(outa))[’x’,c(’Estimate’,’Std. Error’)])
        }
        return(mydat)
    }

### Simulating 60 cohorts and obtaining their unadjusted and adjusted
### estimates; data generation models used are:
###     model for X: logit(P(X=1)) = a0 + a1*C1 + a2*C2
###     model for Y: logit(P(Y=1)) = b0 + b1*C1 + b2*C2 + bx*X
samplesizes <- rep(150, 60)
a <- c(log(0.5/0.5), 0.5, 0.5)
b <- c(log(0.3/0.7), 0.5, 0.5)
bx <- log(1)
mydat <- getcohorts(samplesizes, a, b, bx, model="glm", type="unadj")
### randomly assign 30 cohorts to have only unadjusted estimates
set.seed(1)
noconfcohorts <- sort(sample(1:60, size=30, replace=F))
mydat[which(mydat$cohort %in% noconfcohorts), c(’b.adj’,’se.adj’)] <- NA
mydat

### Implementing CIMBAL to impute adjusted estimates for the
### combined no-confounder cohort and meta-analyze adjusted
### estimates from all 60 cohorts
out <- cimbal(dat=mydat)
# imputed adjusted estimate for combined 30 no-confounder cohorts
out$dat.imputed[nrow(out$dat.imputed),]
# final exposure-outcome adjusted effect estimate meta-analyzing all
# 60 cohorts
out$beta.meta.cimbal
```

```
# corresponding adjusted SE estimate
out$se.meta.cimbal

### One can also implement CIMBAL if some cohorts report partially
### adjusted estimates and others report fully adjusted estimates
mydat <- getcohorts(samplesizes, a, b, bx, model="glm", type="padj")
# randomly assign 30 cohorts to have only C1-adjusted estimates
set.seed(1)
noconfcohorts <- sort(sample(1:60, size=30, replace=F))
mydat[which(mydat$cohort %in% noconfcohorts), c('b.adj','se.adj')] <- NA
# implement CIMBAL
out <- cimbal(dat=mydat)
# imputed fully adjusted estimate for combined 30 cohorts with only
# C1 confounder
out$dat.imputed[nrow(out$dat.imputed),]
# final exposure-outcome fully adjusted effect estimate meta-analyzing
# all 60 cohorts
out$beta.meta.cimbal
# corresponding adjusted SE estimate
out$se.meta.cimbal
```

`impute.summstat.single` *To impute adjusted summary statistics of a cohort*

Description

The R function `impute.summstat.single` allows the user to impute adjusted effect estimate and its SE of a single cohort that reported only unadjusted summary statistics.

Usage

```
impute.summstat.single(est.unadj.2.vec, var.unadj.2.vec,
        est.adj.2.vec, var.adj.2.vec, est.unadj.1, var.unadj.1,
        mute.msgs=FALSE, ncohort.thresh=25)
```

Arguments

`est.unadj.2.vec` It is the vector (length K_c) of unadjusted effect estimates (β or $\log(\mathrm{OR})$) for the K_c 'complete' cohorts that reported both unadjusted and adjusted summary statistics.

`var.unadj.2.vec` It is the vector (length K_c) of squared SE estimates corresponding to the unadjusted effect estimates `est.unadj.2.vec` of the complete cohorts.

`est.adj.2.vec` It is the vector (length K_c) of adjusted effect estimates (β or $\log(\text{OR})$) for the K_c complete cohorts that reported both unadjusted and adjusted summary statistics.

`var.adj.2.vec` It is the vector (length K_c) of squared SE estimates corresponding to the adjusted effect estimates `est.adj.2.vec` of the complete cohorts.

`est.unadj.1` It is the unadjusted effect estimate (β or $\log(\text{OR})$) for one ‘incomplete’ cohort that reported only unadjusted summary statistics.

`var.unadj.1` It is the squared SE estimate corresponding to the unadjusted effect estimate `est.unadj.1`.

`mute.msgs` Logical; if `TRUE`, different messages during CIMBAL implementation are muted. Default value is `FALSE`.

`ncohort.thresh` Minimum number of complete cohorts to use for estimating covariance between unadjusted and adjusted effect estimates. This covariance is needed to impute the adjusted SE estimate of the incomplete cohort. Default value is 25.

Value

`beta.adj.imputed` The imputed adjusted effect estimate for the incomplete cohort that reported only unadjusted estimates.

`var.adj.imputed` The imputed adjusted squared SE estimate for the incomplete cohort that reported only unadjusted estimates.

`meta.fixed.invvar` *Inverse-variance weighted fixed effect meta-analysis*

Description

The R function `meta.fixed.invvar` implements the popular fixed effect meta-analysis of effect estimates using inverse-variance weights.

Usage

```
meta.fixed.invvar(beta.vec, var.vec, returnSE=FALSE)
```

Arguments

`beta.vec` It is the vector of effect estimates (β or $\log(\mathrm{OR})$) for the cohorts to be meta-analyzed.

`var.vec` It is the vector of squared SE estimates corresponding to the effect estimates in `beta.vec`.

`returnSE` Logical; if `TRUE`, meta-analyzed SE estimate is returned instead of squared SE estimate. Default value is `FALSE`.

Value

`beta.meta` Inverse-variance weighted fixed effect meta-analysis effect estimate.

`var.meta` Variance estimate corresponding to `beta.meta`. If `returnSE` is `TRUE`, SE estimate is returned as `se.meta`.