MR-Hevo: statistical model and methods

Statistical model

- X exposure
- Y outcome
- $\boldsymbol{Z}$ vector of genotypic instruments, of length equal to the number J of unlinked loci
- $\boldsymbol{\alpha}$ vector of coefficients of effects of instruments Z on exposure X
- $\boldsymbol{\beta}$ vector of coefficients of direct (pleiotropic) effects of instruments on outcome Y
- $\boldsymbol{X_u}$ unpenalized covariates
- θ parameter for causal effect of X on Y
- $\beta 0$, $\boldsymbol{\beta_u}$ parameters for intercept and unpenalized covariates $\boldsymbol{X_u}$

We have a dataset of N individuals with measurements of the outcome Y and the genetic instruments $\boldsymbol{Z}$. From summary statistics we have estimates $\hat{\boldsymbol{\alpha}}$ of the effects $\boldsymbol{\alpha}$ of the instrument on the exposure, with corresponding standard errors $a_1, \dots, a_j$.

We specify a Bayesian full probability model as below

$$\alpha_j \sim N\left(\hat{\alpha}_j, a_j^2\right)$$

$$\mathbb{E}\langle X \rangle = \alpha_0 + \boldsymbol{Z\alpha}$$

$$g\left(\mathbb{E}\langle Y \rangle\right) = \beta_0 + \boldsymbol{X_u \beta_u} + \boldsymbol{Z\beta} + \theta \mathbb{E}\langle X \rangle$$

where $g\left(\right)$ is a link function.

To calculate the likelihood as a function of the causal effect parameter θ, we have to marginalize over the distribution of the direct effects $\boldsymbol{\beta}$ given the data

The regression coefficients are given a regularized horseshoe prior

$$\beta_j \sim N\left(0, \tau^2 \tilde{\lambda}_j^2\right), \tilde{\lambda}_j^2 = \frac{\eta \lambda_j^2}{\eta + \tau^2 \lambda_j^2}$$

Half-Cauchy priors are specified on the unregularized local scale parameters λ_j.

$$\lambda_j \sim C^+\left(0, 1\right)$$

A weakly informative gamma distribution is specified for η:

$$\eta \sim \text{Gamma}\left(0.5\nu_{\text{slab}}, 0.5\nu_{\text{slab}} s_{\text{slab}}^2\right)$$

The heavy tail of the half-Cauchy distribution allows some of the regression coefficients to escape the shrinkage imposed by the global parameter τ. The regularization parameter η regularizes the scale of the nonzero coefficients (those that are in the slab of the spike and slab distribution). Even the largest coefficients will be regularized as a Gaussian with variance η.

The value of ν_{slab} controls the shape of the distribution of η. Piironen and Vehtari recommend setting $\nu_{\text{slab}} = 1$, but setting $\nu_{\text{slab}} = 2$ may be required to regularize the sampler so that it does not draw very large values of η.

The scaling factor s_{slab} is specified based on prior information about the size of the largest direct effects. This information will usually be available from genome-wide association studies of the outcome.

A half-t distribution is chosen for the global scale parameter τ

$$\tau \sim t^{+}\left(0, s_{\text{global}}, \nu_{\text{global}}\right)$$

Specifying $\nu_{\text{global}} = 1$ gives a half-Cauchy prior. Setting $\nu_{\text{aglobal}} = 2$ may be required to regularize the sampler so that it does not draw very large values of τ. This regularization shrinks the right tail of the distribution of τ, and thus limits narrowness of the spike component.

The scaling factor s_{global} is specified to encode a prior guess about the number r_0 of nonzero coefficients for the direct effects. Piironen and Vehtari show that this implies that most of the prior mass for τ is located near the value

$$\tau_0 = \frac{r_0}{J - r_0} \frac{\sigma_y}{\sqrt{N}}$$

We specify s_{global} so that the median of the prior on τ is τ_0 calculated as above.

For the jth instrument, the *shrinkage coefficient* κ_j is

$$\kappa_j = \frac{1}{1 + \tilde{\lambda}_j^2}$$

This takes values from 0 (no shrinkage) to 1 (complete shrinkage). The prior on this parameter has a horseshoe shape.

The effective number m of nonzero coefficients is then

$$m = \sigma(1 - \kappa_j)$$

Extension to instruments that are calculated from multiple SNPs

For each clump of exposure-associated SNPs and each individual in the target dataset, a locus-specific score is calculated from the vector $\hat{\gamma_{\boldsymbol{u}}}$ of univariate summary statistics for the effect of SNPs on the exposure X. Estimates of the multivariable coefficients $\hat{\gamma}$ are calculated by premultiplying the univariate coefficients by the correlation matrix between the SNP genotypes (obtained from a reference population). Where this correlation matrix is singular or ill-conditioned, a pseudo-inverse can be used to calculate the multivariable coefficients.

The locus-specific score S is then calculated as $\boldsymbol{G} \cdot \boldsymbol{\gamma}$.

Because S is calculated from the genotypes and and the coefficients for the effect of genotypes on exposure, we cannot simply substitute it for the genetic instrument Z in the model above. We can however factor the dot product $\boldsymbol{G} \cdot \boldsymbol{\gamma}$ as the product of two scalars: the magnitude of the coefficient vector $|\boldsymbol{\gamma}|$ and a pseudo-genotype $|\boldsymbol{G}| \cos\phi$, where ϕ is the angle between the vectors $\boldsymbol{G}$ and $\boldsymbol{\gamma}$. We can then substitute $|\boldsymbol{\gamma}|$ for the scalar coefficient α and $S/|\boldsymbol{\gamma}|$ for the scalar instrument Z as a pseudo-genotype in the statistical model defined above.

This procedure ensures that all exposure-associated SNPs can be used in constructing the genotypic instruments, and that these instruments are unlinked so that their pleiotropic effects can be modelled as independent.

Computational methods

To generate the posterior distribution given the model and the data, we use the program `Stan`. Scripts for a linear regression (continuous outcome) and a logistic regression (binary outcome) are here.

The likelihood from the posterior distribution of θ by dividing by the prior. This is done by fitting a kernel density to the posterior samples of θ, weighting each observation by the inverse of the prior. We take the logarithm of this likelihood, and fit a quadratic function to the log-likelihood function. The maximum likelihood estimate and test of the null hypothesis $\theta = 0$ are obtained from this quadratic approximation to the log-likelihoood.