MASSIVE: Tractable and Robust Bayesian Learning of Many-Dimensional Instrumental Variable Models – Supplementary Material

Ioan Gabriel Bucur **Tom Claassen** **Tom Heskes**
Department of Data Science
Institute for Computing and Information Sciences
Radboud University
Nijmegen, The Netherlands

A DERIVING THE MAXIMUM LIKELIHOOD ESTIMATES

In our model, we assume that the data is generated from the following linear structural equation model:

$$\begin{aligned} U &:= \epsilon_U \\ G_j &:= \epsilon_{G_j} \\ X &:= \sum_j \gamma_j G_j + \kappa_X U + \epsilon_X \\ Y &:= \sum_j \alpha_j G_j + \kappa_Y U + \beta X + \epsilon_Y \end{aligned} \quad . \tag{1}$$

In order to avoid any scaling issues, we first divide each structural equation in (1) by the scale of the noise term. We then define priors on the scale-free interactions. The scaled structural parameters are

$$\begin{aligned} &\tilde{\gamma}_j = \sigma_{G_j}\sigma_X^{-1}\gamma_j; \quad \tilde{\alpha}_j = \sigma_{G_j}\sigma_Y^{-1}\alpha_j; \\ &\tilde{\beta} = \sigma_X\sigma_Y^{-1}\beta; \quad \tilde{\kappa}_X = \sigma_X^{-1}\kappa_X; \quad \tilde{\kappa}_Y = \sigma_Y^{-1}\kappa_Y. \end{aligned}$$

We assume that the data consists of N i.i.d. observations $\mathbf{D} = (\boldsymbol{G}_i, X_i, Y_i)_{1\leq i\leq N}$. The conditional Gaussian observed data likelihood reads

$$\mathcal{L}\left(\begin{bmatrix} X \\ Y \end{bmatrix} \middle| \boldsymbol{G}\right) = (4\pi^2|\mathbf{\Sigma}|)^{-\frac{N}{2}} \exp\left\{-\frac{N}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right\}, \tag{2}$$

with

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\left\{\begin{bmatrix} X_i \\ Y_i \end{bmatrix} - \boldsymbol{\mu}(\boldsymbol{G}_i)\right\}\left\{\begin{bmatrix} X_i \\ Y_i \end{bmatrix} - \boldsymbol{\mu}(\boldsymbol{G}_i)\right\}^\intercal$$

and

$$\boldsymbol{\mu}(\boldsymbol{G}) = \begin{bmatrix}\boldsymbol{\gamma} & \beta\boldsymbol{\gamma}+\boldsymbol{\alpha}\end{bmatrix}^\intercal \boldsymbol{G} = \begin{bmatrix}\boldsymbol{\gamma} & \mathbf{\Gamma}\end{bmatrix}^\intercal \boldsymbol{G}.$$

The maximum of the conditional likelihood function occurs at $\mathbf{S} = \mathbf{\Sigma}$. Our model has $2J + 5$ (scaled) parameters, $\widetilde{\mathbf{\Theta}} = (\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\alpha}}, \tilde{\beta}, \log\sigma_X, \log\sigma_Y, \tilde{\kappa}_X, \tilde{\kappa}_Y)$, which is more than the number of independent constraints $(2J{+}3)$ imposed by maximizing the likelihood. This makes the problem of finding the maximum likelihood estimate undetermined, but if we fix the values of $\widetilde{\mathbf{C}} = (\tilde{\kappa}_X, \tilde{\kappa}_Y)$, we can analytically derive the other parameters ($\widetilde{\mathbf{B}}$) such that the likelihood is maximized.

We have as input sufficient statistics the first and second-order empirical (raw) moments of the data:

$$\begin{aligned} \overline{\boldsymbol{G}} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i \to \mathbb{E}\left[\boldsymbol{G}\right]; \\ \overline{X} &= \frac{1}{N}\sum_{i=1}^{N}X_i \to \mathbb{E}\left[X\right]; \\ \overline{Y} &= \frac{1}{N}\sum_{i=1}^{N}Y_i \to \mathbb{E}\left[Y\right]; \\ \overline{\boldsymbol{G}\boldsymbol{G}^\intercal} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_i\boldsymbol{G}_i^\intercal \to \mathbb{E}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right]; \\ \overline{\boldsymbol{G}X} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_iX_i \to \mathbb{E}\left[\boldsymbol{G}X\right]; \\ \overline{\boldsymbol{G}Y} &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{G}_iY_i \to \mathbb{E}\left[\boldsymbol{G}Y\right]; \\ \overline{X^2} &= \frac{1}{N}\sum_{i=1}^{N}X_i^2 \to \mathbb{E}\left[X^2\right]; \\ \overline{Y^2} &= \frac{1}{N}\sum_{i=1}^{N}Y_i^2 \to \mathbb{E}\left[Y^2\right]; \\ \overline{XY} &= \frac{1}{N}\sum_{i=1}^{N}X_iY_i \to \mathbb{E}\left[XY\right]. \end{aligned}$$

The maximum likelihood estimator here coincides with the method of moments estimator, so we will derive the

ML estimates using moment matching, which is straightforward. The conditional moments relate to the parameters as follows:

$$\begin{aligned}
\mathbb{E}\left[X|\boldsymbol{G}\right] &= \boldsymbol{\gamma}^\intercal \boldsymbol{G} \\
\mathbb{E}\left[Y|\boldsymbol{G}\right] &= \boldsymbol{\Gamma}^\intercal \boldsymbol{G} \\
\mathrm{Var}\left[X|\boldsymbol{G}\right] &= \sigma_X^2 + \kappa_X^2 \\
\mathrm{Cov}\left[X,Y|\boldsymbol{G}\right] &= \beta(\sigma_X^2 + \kappa_X^2) + \kappa_X\kappa_Y \\
\mathrm{Var}\left[Y|\boldsymbol{G}\right] &= \sigma_Y^2 + \beta^2\sigma_X^2 + (\kappa_Y + \beta\kappa_X)^2.
\end{aligned}$$

We now relate the previous statements to the unconditional moments:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{G}X\right] &= \mathbb{E}\left[\boldsymbol{G}\mathbb{E}\left[X|\boldsymbol{G}\right]\right] = \mathbb{E}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right]\boldsymbol{\gamma} \\
\mathbb{E}\left[\boldsymbol{G}Y\right] &= \mathbb{E}\left[\boldsymbol{G}\mathbb{E}\left[Y|\boldsymbol{G}\right]\right] = \mathbb{E}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right]\boldsymbol{\Gamma} \\
\mathbb{E}\left[X^2\right] &= \boldsymbol{\gamma}^\intercal\mathbb{E}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right]\boldsymbol{\gamma} + \sigma_X^2 + \kappa_X^2 \\
\mathbb{E}\left[XY\right] &= \boldsymbol{\gamma}^\intercal\mathbb{E}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right]\boldsymbol{\Gamma} + \beta(\sigma_X^2 + \kappa_X^2) + \kappa_X\kappa_Y \\
\mathbb{E}\left[Y^2\right] &= \boldsymbol{\Gamma}^\intercal\mathbb{E}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right]\boldsymbol{\Gamma} + \sigma_Y^2 + \beta^2\sigma_X^2 + (\kappa_Y + \beta\kappa_X)^2.
\end{aligned}$$

We therefore obtain the constraints

$$\begin{aligned}
\boldsymbol{\gamma} &= (\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}X} \\
\beta\boldsymbol{\gamma} + \boldsymbol{\alpha} &= (\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}Y} \\
\sigma_X^2 + \kappa_X^2 &= \widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right] \\
&= \overline{X^2} - \overline{X\boldsymbol{G}^\intercal}(\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}X} \\
\beta(\sigma_X^2 + \kappa_X^2) + \kappa_X\kappa_Y &= \widehat{\mathrm{Cov}}\left[X,Y|\boldsymbol{G}\right] \\
&= \overline{XY} - \overline{X\boldsymbol{G}^\intercal}(\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}Y} \\
\sigma_Y^2 + \beta^2\sigma_X^2 + (\kappa_Y + \beta\kappa_X)^2 &= \widehat{\mathrm{Var}}\left[Y|\boldsymbol{G}\right] \\
&= \overline{Y^2} - \overline{Y\boldsymbol{G}^\intercal}(\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}Y}.
\end{aligned}$$

The next step is to express the above constraints in terms of the scaled parameters:

$$\begin{aligned}
\mathrm{Var}\left[\boldsymbol{G}\right]\tilde{\boldsymbol{\gamma}}\sigma_X^{-1} &= (\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}X} \\
\mathrm{Var}\left[\boldsymbol{G}\right](\tilde{\beta}\tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\alpha}})\sigma_Y^{-1} &= (\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}Y} \\
\sigma_X^2(1 + \tilde{\kappa}_X^2) &= \widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right] \\
&= \overline{X^2} - \overline{X\boldsymbol{G}^\intercal}(\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}X} \\
\sigma_X\sigma_Y\left[\tilde{\beta}(1 + \tilde{\kappa}_X^2) + \tilde{\kappa}_X\tilde{\kappa}_Y\right] &= \widehat{\mathrm{Cov}}\left[X,Y|\boldsymbol{G}\right] \\
&= \overline{XY} - \overline{X\boldsymbol{G}^\intercal}(\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}Y} \\
\sigma_Y^2\left[1 + \tilde{\beta}^2 + (\tilde{\kappa}_Y + \tilde{\beta}\tilde{\kappa}_X)^2\right] &= \widehat{\mathrm{Var}}\left[Y|\boldsymbol{G}\right] \\
&= \overline{Y^2} - \overline{Y\boldsymbol{G}^\intercal}(\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1}\overline{\boldsymbol{G}Y}.
\end{aligned}$$

From the above constraints, given fixed values for $\tilde{\kappa}_X$ and $\tilde{\kappa}_Y$, we obtain the following (scaled) parameter values that maximize the likelihood in (2):

$$\begin{aligned}
(\sigma_X^{\mathrm{ML}})^2 &= \frac{\widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right]}{1 + \tilde{\kappa}_X^2} \\
(\sigma_Y^{\mathrm{ML}})^2 &= \left(\widehat{\mathrm{Var}}\left[Y|\boldsymbol{G}\right] - \frac{(\widehat{\mathrm{Cov}}\left[X,Y|\boldsymbol{G}\right])^2}{\widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right]}\right)\frac{1 + \tilde{\kappa}_X^2}{1 + \tilde{\kappa}_X^2 + \tilde{\kappa}_Y^2} \\
\tilde{\beta}^{\mathrm{ML}} &= \frac{\widehat{\mathrm{Cov}}\left[X,Y|\boldsymbol{G}\right](\sigma_X^{\mathrm{ML}}\sigma_Y^{\mathrm{ML}})^{-1} - \tilde{\kappa}_X\tilde{\kappa}_Y}{1 + \tilde{\kappa}_X^2} \\
\tilde{\boldsymbol{\gamma}}^{\mathrm{ML}} &= \sqrt{\widehat{\mathrm{Var}}\left[\boldsymbol{G}\right]}(\widehat{\mathbb{E}}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right])^{-1}\widehat{\mathbb{E}}\left[\boldsymbol{G}X\right](\sigma_X^{\mathrm{ML}})^{-1} \\
\tilde{\boldsymbol{\alpha}}^{\mathrm{ML}} &= \sqrt{\widehat{\mathrm{Var}}\left[\boldsymbol{G}\right]}(\widehat{\mathbb{E}}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right])^{-1}\widehat{\mathbb{E}}\left[\boldsymbol{G}Y\right](\sigma_Y^{\mathrm{ML}})^{-1} - \tilde{\beta}^{\mathrm{ML}}\tilde{\boldsymbol{\gamma}}^{\mathrm{ML}}
\end{aligned} \quad . \tag{3}$$

B USING SUMMARY STATISTICS AS INPUT

To run MASSIVE, we must provide the first and second-order moments of the observed data $\mathbf{Z} = (\boldsymbol{G}, X, Y)$ as input to plug into the data likelihood from Equation (2). If we have access to the whole data set, then the moments can immediately be derived. Much more often, however, individual-level data is unavailable and instead we have to rely on published GWAS results, which typically come in the form of regression coefficients together with their standard errors. In this section we show how the first and second-order moments can be derived from this summary data, thereby making MASSIVE applicable on a much broader set of data sources.

To obtain all the necessary input sufficient statistics, we require the following summary data:

- $\hat{p}_j$: the effect allele frequency (EAF) of G_j
- m: the number of allele copies (almost always equal to two, since humans are diploid organisms)
- $\hat{\gamma}_j, \hat{\sigma}_{\hat{\gamma}_j}, N_{\hat{\gamma}_j}$: for the gene-exposure associations, we require the coefficient obtained by regressing X on G_j , its standard error and the sample size
- $\hat{\Gamma}_j, \hat{\sigma}_{\hat{\Gamma}_j}, N_{\hat{\Gamma}_j}$: for the gene-outcome associations, we require the coefficient obtained by regressing Y on G_j, its standard error and the sample size
- $\hat{\beta}$: the coefficient obtained by regressing X on Y (observational exposure-outcome association)

Summary data on gene-exposure and gene-outcome associations from GWAS is widely available, so we can typically get estimates for $\hat{\gamma}_j$, $\hat{\Gamma}_j$ together with the associated standard errors and sample sizes. The effect allele frequency $\hat{p}_j$ is usually also reported. In addition, we require a measure of the association between the exposure and the outcome ($\hat{\beta}$) to derive an estimate of $\mathrm{Cov}\left[X, Y\right]$. This estimate can be obtained from observational studies for determining potential risk factors for the outcome.

To estimate the second-order moments, we employ the following well-known approximations from simple linear regression:

$$
\begin{aligned}
\hat{\gamma}_j &\approx \frac{\mathrm{Cov}\left[G_j, X\right]}{\mathrm{Var}\left[G_j\right]} \\
\hat{\Gamma}_j &\approx \frac{\mathrm{Cov}\left[G_j, Y\right]}{\mathrm{Var}\left[G_j\right]} \\
\hat{\beta} &\approx \frac{\mathrm{Cov}\left[X, Y\right]}{\mathrm{Var}\left[X\right]} \\
\hat{\sigma}^2_{\hat{\gamma}_j} &\approx \frac{1}{N_\gamma}\left(\frac{\mathrm{Var}\left[X\right]}{\mathrm{Var}\left[G_j\right]} - \hat{\gamma}^2\right) \\
\hat{\sigma}^2_{\hat{\Gamma}_j} &\approx \frac{1}{N_\Gamma}\left(\frac{\mathrm{Var}\left[Y\right]}{\mathrm{Var}\left[G_j\right]} - \hat{\Gamma}^2\right)
\end{aligned}
.
$$

Note that these approximations also apply in a multivariate setting when the regressors are independent. Moreover, to compute the expected values and variances for the genetic variants, we assume a binomial distribution, so we plug in the EAF as the estimated success probability and then use the appropriate formulas. We use all these approximations to finally derive the following estimates for the moments from summary statistics:

$$
\begin{aligned}
\mathbb{E}\left[G_j\right] &\approx m \cdot \hat{p}_j \; (= \widehat{\mathbb{E}\left[G_j\right]}) \\
\mathbb{E}\left[X\right] &\approx \sum_j \widehat{\mathbb{E}\left[G_j\right]} \cdot \hat{\gamma}_j \\
\mathbb{E}\left[Y\right] &\approx \sum_j \widehat{\mathbb{E}\left[G_j\right]} \cdot \hat{\Gamma}_j \\
\mathrm{Var}\left[G_j\right] &\approx m \cdot \hat{p}_j \cdot (1 - \hat{p}_j) \; (= \widehat{\mathrm{Var}\left[G_j\right]}) \\
\mathrm{Cov}\left[G_j, X\right] &\approx \widehat{\mathrm{Var}\left[G_j\right]} \cdot \hat{\gamma}_j \\
\mathrm{Cov}\left[G_j, Y\right] &\approx \widehat{\mathrm{Var}\left[G_j\right]} \cdot \hat{\Gamma}_j \\
\mathrm{Var}\left[X\right] &\approx \widehat{\mathrm{Var}\left[G_j\right]} \cdot (\hat{\gamma}_j^2 + N_{\hat{\gamma}_j} \cdot \hat{\sigma}^2_{\hat{\gamma}_j}) \; (= \widehat{\mathrm{Var}\left[X\right]}) \\
\mathrm{Var}\left[Y\right] &\approx \widehat{\mathrm{Var}\left[G_j\right]} \cdot (\hat{\Gamma}_j^2 + N_{\hat{\Gamma}_j} \cdot \hat{\sigma}^2_{\hat{\Gamma}_j}) \\
\mathrm{Cov}\left[X, Y\right] &\approx \widehat{\mathrm{Var}\left[X\right]} \cdot \hat{\beta}
\end{aligned}
. \quad (4)
$$

When we have information on multiple genetic variants, we obtain multiple estimates of $\mathrm{Var}\left[X\right]$ and $\mathrm{Var}\left[Y\right]$ in (4), in which case we take the median over the estimates. Our approach also requires specifying a sample size. Since the summary statistics are likely to be computed from different samples, we conservatively choose the minimum of their sizes as input to `MASSIVE` in order not to overestimate the precision of the data. If the sample size for the exposure-outcome association measure is also available, we take it into consideration when calculating the minimum of the sample sizes.

C SMART INITIALIZATION PROCEDURE FOR THE POSTERIOR OPTIMIZATION

We propose to start the search for posterior local optima from the bivariate maximum likelihood manifold. Since we are looking for sparse parameter solutions, we also start from points on the manifold that exhibit some degree of sparsity.

The first starting point corresponds to the *no confounding* sparse solution, where we fix $\tilde{\kappa}_X = \tilde{\kappa}_Y = 0$. The other parameters can be derived using (3):

$$
\begin{aligned}
(\sigma_X^{\mathrm{ML}})^2 &= \widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right] \\
(\sigma_Y^{\mathrm{ML}})^2 &= \widehat{\mathrm{Var}}\left[Y|\boldsymbol{G}\right] - \frac{(\widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right])^2}{\widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right]} \\
\tilde{\beta}^{\mathrm{ML}} &= \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right] (\sigma_X^{\mathrm{ML}} \sigma_Y^{\mathrm{ML}})^{-1} \\
\tilde{\boldsymbol{\gamma}}^{\mathrm{ML}} &= \widehat{\mathrm{Var}}\left[\boldsymbol{G}\right] (\widehat{\mathbb{E}}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right])^{-1} \widehat{\mathbb{E}}\left[\boldsymbol{G}X\right] (\sigma_X^{\mathrm{ML}})^{-1} \\
\tilde{\boldsymbol{\alpha}}^{\mathrm{ML}} &= \widehat{\mathrm{Var}}\left[\boldsymbol{G}\right] (\widehat{\mathbb{E}}\left[\boldsymbol{G}\boldsymbol{G}^\intercal\right])^{-1} \widehat{\mathbb{E}}\left[\boldsymbol{G}Y\right] (\sigma_Y^{\mathrm{ML}})^{-1} - \tilde{\beta}^{\mathrm{ML}} \tilde{\boldsymbol{\gamma}}^{\mathrm{ML}}.
\end{aligned}
$$

The second starting point corresponds to the *no causal effect* solution, where we fix $\tilde{\beta} = 0$. By solving the equation system in (3) with $\tilde{\beta} = 0$, we obtain the following constraint:

$$
\tilde{\kappa}_X \tilde{\kappa}_Y = \frac{\widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}{1 - \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}.
$$

We have one degree of freedom left for choosing $\tilde{\kappa}_X$ and $\tilde{\kappa}_Y$. We propose to additionally set $|\tilde{\kappa}_X| = |\tilde{\kappa}_Y|$ and assume $\tilde{\kappa}_X > 0$. Finally, we obtain:

$$
\begin{aligned}
\tilde{\kappa}_X^{\mathrm{ML}} &= \sqrt{\left|\frac{\widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}{1 - \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}\right|} \\
\tilde{\kappa}_Y^{\mathrm{ML}} &= \sqrt{\left|\frac{\widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}{1 - \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}\right|} \cdot \mathrm{sign}\left\{\frac{\widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}{1 - \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right]}\right\}.
\end{aligned}
$$

The rest of the parameters can be derived given $(\tilde{\kappa}_X, \tilde{\kappa}_Y)$ from (3).

The third starting point corresponds to minimizing the pleiotropic effects sum of squares. If we consider the constraint (at the maximum likelihood estimate):

$$
\boldsymbol{\alpha} = (\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1} \overline{\boldsymbol{G}Y} - \beta (\overline{\boldsymbol{G}\boldsymbol{G}^\intercal})^{-1} \overline{\boldsymbol{G}X} = r^{Y|\boldsymbol{G}} - \beta r^{X|\boldsymbol{G}},
$$

where $r^{X|\boldsymbol{G}}$ and $r^{Y|\boldsymbol{G}}$ are the coefficients obtained by regressing $\boldsymbol{G}$ on X and Y, respectively,

$$
\beta^* = \arg\min \sum_{j=1}^{J} \alpha_j^2 = \arg\min \left(r_j^{Y|\boldsymbol{G}} - \beta r_j^{X|\boldsymbol{G}}\right)^2.
$$

The solution to this minimization problem is:

$$\beta^* = \frac{1}{J}\frac{\sum_{j=1}^{J} r_j^{X|\boldsymbol{G}} r_j^{Y|\boldsymbol{G}}}{\sum_{j=1}^{J} r_j^{X|\boldsymbol{G}} r_j^{X|\boldsymbol{G}}}.$$

For independent instruments, the right-hand side ratios above corresponds to the instrumental variable estimates. By solving the equation system in (3) with $\beta = \beta^*$, we obtain the following constraint:

$$\tilde{\kappa}_X \tilde{\kappa}_Y = \frac{C}{1-C},$$

where

$$C = \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right] - \\ - \frac{\beta^* \widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right]}{\sqrt{\widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right] (\widehat{\mathrm{Var}}\left[Y|\boldsymbol{G}\right] + (\beta^*)^2 \widehat{\mathrm{Var}}\left[X|\boldsymbol{G}\right] - 2\beta^* \widehat{\mathrm{Cov}}\left[X, Y|\boldsymbol{G}\right])}}.$$

We have one degree of freedom left for choosing $\tilde{\kappa}_X$ and $\tilde{\kappa}_Y$. We propose to additionally set $|\tilde{\kappa}_X| = |\tilde{\kappa}_Y|$ and assume $\tilde{\kappa}_X > 0$. Finally, we obtain:

$$\tilde{\kappa}_X^{\mathrm{ML}} = \sqrt{\left|\frac{C}{1-C}\right|}$$
$$\tilde{\kappa}_Y^{\mathrm{ML}} = \sqrt{\left|\frac{C}{1-C}\right|} \cdot \mathrm{sign}\left\{\frac{C}{1-C}\right\}.$$

D DETERMINING THE PRIOR HYPERPARAMETERS EMPIRICALLY

We base our choice of prior hyperparameters on how likely it is for the observed genetic associations to have come from the prior. We start by choosing a reasonable hyperparameter for the 'slab' component (σ_{slab}). Since instrument candidates are chosen based on the robustness of their association with the exposure X, we can use the size of these associations as a measure of the effect size of relevant effects, i.e., those corresponding to the 'slab' component. In our framework, this translates to the assumption that all the instrument strengths $\tilde{\gamma}_j$ arise from the 'slab' distribution, and will therefore give a good indication of the expected effect size for relevant parameters. Consequently, we want to find the hyperparameter value that maximizes the (log-)likelihood of the genetic associations coming from $\mathcal{N}(0, \sigma^2_{\text{slab}})$:

$$\sigma^*_{\text{slab}} = \arg\max_{\sigma_{\text{slab}}} \sum_{j=1}^{J} \left[-\frac{1}{2}\log(2\pi\sigma^2_{\text{slab}}) - \frac{\tilde{\gamma}_j^2}{2\sigma^2_{\text{slab}}}\right]. \quad (5)$$

Maximizing the above log-likelihood is straightforward if we know the instrument strengths $\tilde{\gamma}_j$ on the right-hand side from data. Instead, we will plug in an empirical estimate of the scaled instrument strengths. We use the fact that the unscaled maximum likelihood estimate for the instrument strengths γ_j is identifiable as

$$\boldsymbol{\gamma}^{\mathrm{ML}} = (\overline{\boldsymbol{G}\boldsymbol{G}^{\intercal}})^{-1}\overline{\boldsymbol{G}X}.$$

For the scaled parameters we then have:

$$(\tilde{\gamma}_j^{\mathrm{ML}})^2 = \sigma^2_{G_j}(\gamma_j^{\mathrm{ML}})^2(\sigma_X^{\mathrm{ML}})^{-2} = \frac{\sigma^2_{G_j}(\gamma_j^{\mathrm{ML}})^2(1+\tilde{\kappa}_X^2)}{\mathrm{Var}\left[X|\boldsymbol{G}\right]}.$$

These values are undetermined because we de not know the confounding coefficient $\tilde{\kappa}_X$. We propose to compute an average estimate by integrating out $\tilde{\kappa}_X$, which we have assumed follows a $\mathcal{N}(0, 10)$ distribution a-priori. We average over all possible values of $\tilde{\kappa}_X$ to get

$$\begin{aligned}\mathbb{E}\left[(\tilde{\gamma}_j^{\mathrm{ML}})^2\right] &= \frac{\sigma^2_{G_j}(\gamma_j^{\mathrm{ML}})^2}{\mathrm{Var}\left[X|\boldsymbol{G}\right]}\int_{-\infty}^{\infty}(1+\tilde{\kappa}_X^2)\,\mathcal{N}(\tilde{\kappa}_X; 0, 10)\,\mathrm{d}\tilde{\kappa}_X \\ &= \frac{\sigma^2_{G_j}(\gamma_j^{\mathrm{ML}})^2}{\mathrm{Var}\left[X|\boldsymbol{G}\right]}\cdot 101 \\ &\stackrel{!}{=} 101 D_j^2.\end{aligned} \quad (6)$$

We plug in the derived estimate into (5) to get

$$\sigma^*_{\text{slab}} = \arg\max_{\sigma_{\text{slab}}} \sum_{j=1}^{J}\left[-\log\sigma_{\text{slab}} - \frac{101 D_j^2}{2\sigma^2_{\text{slab}}}\right].$$

From this we finally obtain our first empirically determined hyperparameter

$$(\sigma^*_{\text{slab}})^2 = \frac{101}{J}\sum_{j=1}^{J} D_j^2 = \frac{101}{J}\sum_{j=1}^{J}\frac{\sigma^2_{G_j}(\gamma_j^{\mathrm{ML}})^2}{\mathrm{Var}\left[X|\boldsymbol{G}\right]}. \quad (7)$$

We now derive a reasonable hyperparameter for the 'spike' component (σ_{spike}), relative to the previously determined σ^*_{slab}. The potential gain (or penalty) in moving $\tilde{\gamma}_{\text{min}}$ from the slab to the spike component in the prior is

$$G(\sigma_{\text{spike}}, \sigma_{\text{slab}}) = \log\mathcal{N}(\tilde{\gamma}_{\text{min}}; 0, \sigma_{\text{spike}}) - \log\mathcal{N}(\tilde{\gamma}_{\text{min}}; 0, \sigma_{\text{slab}}).$$

The penalty in the likelihood (approximated by a normal distribution) due to the parameter shrinkage from its current value to zero is

$$P(\sigma_{\text{slab}}) = N\cdot[\log\mathcal{N}(0; \tilde{\gamma}_{\text{min}}, \sigma_{\text{slab}}) - \log\mathcal{N}(\tilde{\gamma}_{\text{min}}; \tilde{\gamma}_{\text{min}}, \sigma_{\text{slab}})].$$

The empirical argument we employ is to choose σ_{spike} so small such that changing the component of the minimal instrument strength ($\tilde{\gamma}_{\text{min}} = \min_j \tilde{\gamma}_j$) from slab to spike would incur a greater penalty than the one induced on the log-likelihood by shrinking that parameter to zero. This way, fitting any $\tilde{\gamma}_j$ into the 'spike' component is strongly discouraged, in line with our assumption that these are relevant values coming from the 'slab' component. Consequently, as our second empirically determined hyperparameter, we choose the value σ^*_{spike} solving the equation $G(\sigma_{\text{spike}}, \sigma^*_{\text{slab}}) = P(\sigma^*_{\text{slab}})$, where σ^*_{slab} is given in (7) and our estimate of $\tilde{\gamma}_{\text{min}}$ is the smallest of the J expected value estimates derived in (6). It is straightforward to show that the constraint boils down to

$$(N + 1 - C)\left(\frac{101 \min_j D_j^2}{\sigma^*_{\text{slab}}}\right)^2 + \log C = 0,$$

where $C = \left(\frac{\sigma^*_{\text{slab}}}{\sigma_{\text{spike}}}\right)^2$. It can be easily shown that the above equation in C has a unique solution greater than one ($C > 1$ by definition because $\sigma_{\text{spike}} < \sigma_{\text{slab}}$). Via our empirical argument, we have thus arrived at an easily computable, unique pair of hyperparameters $(\sigma^*_{\text{slab}}, \sigma^*_{\text{spike}})$. We emphasize that this choice of parameters is independent of the true causal effect and relies solely on the estimated values of the instrument strengths to calibrate the appropriate size of relevant ('slab') and irrelevant ('spike') effects.