# MASSIVE: Tractable and Robust Bayesian Learning of Many-Dimensional Instrumental Variable Models

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

## Abstract

The recent availability of huge, many-dimensional data sets, like those arising from genome-wide association studies (GWAS), provides many opportunities for strengthening causal inference. One popular approach is to utilize these many-dimensional measurements as instrumental variables (instruments) for improving the causal effect estimate between other pairs of variables. Unfortunately, searching for proper instruments in a many-dimensional set of candidates is a daunting task due to the intractable model space and the fact that we cannot directly test which of these candidates are valid, so most existing search methods either rely on overly stringent modeling assumptions or fail to capture the inherent model uncertainty in the selection process. We show that, as long as at least some of the candidates are (close to) valid, without knowing a priori which ones, they collectively still pose enough restrictions on the target interaction to obtain a reliable causal effect estimate. We propose a general and efficient causal inference algorithm that accounts for model uncertainty by performing Bayesian model averaging over the most promising many-dimensional instrumental variable models, while at the same time employing weaker assumptions regarding the data generating process. We showcase the efficiency, robustness and predictive performance of our algorithm through experimental results on both simulated and real-world data.

## 1 INTRODUCTION

Causal inference is a fundamental topic of research in the biomedical sciences, where the relationship between an exposure to a putative risk factor and a disease outcome or marker is often studied. The gold standard for answering causal questions – e.g., does an intake of vitamin D supplements reduce the risk of developing schizophrenia? – is to perform a *randomized controlled trial* (RCT), in which the exposure (treatment) is assigned randomly to the participants. The purpose of randomization is to eliminate potential confounding due to variables influencing both the exposure and the outcome. Unfortunately, performing an RCT is often unfeasible due to monetary, ethical, or practical constraints (Benson and Hartz, 2000). On the other side of the fence, there are vast amounts of medical data available from observational studies, but estimating a causal effect from such data is prone to confounding, reverse causation, and other biases (Sheehan et al., 2008).

With the advent of high-throughput genomics, an enormous amount of observational genetic data has been collected in large-scale *genome-wide association studies* (GWAS). There is great potential in using this genetic information for strengthening causal inference in observational designs, where the causal effect is obfuscated by potentially unmeasured confounding (Visscher et al., 2017). One popular and powerful systematic approach that can be exploited is to make use of so-called *instrumental variables* or *instruments* (Angrist et al., 1996). In recent years, instrumental variable analysis has become prevalent in the field of genetic epidemiology under the moniker *Mendelian randomization*. Mendelian randomization (MR) refers to the random segregation and assortment of genes from parent to offspring, as stated by Mendel's laws, which can be seen as analogous to the randomization induced in an RCT (Hingorani and Humphries, 2005). In MR studies, *genetic variants*, such as the allele at a particular location in the genome, ful-

*Proceedings of the 36$^{th}$ Conference on Uncertainty in Artificial Intelligence (UAI)*, PMLR volume 124, 2020.

fill the role of instruments (Lawlor et al., 2008). For example, a gene encoding a major enzyme for alcohol metabolism (ALDH2) has been used as a proxy measure for alcohol consumption with the goal of investigating the latter's effect on the risk of coronary heart disease (Davey Smith and Hemani, 2014).

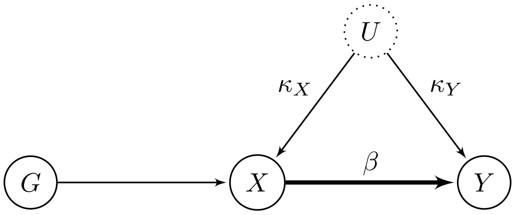

Figure 1: Graphical description of the causal model assumed in instrumental variable analyses. In the figure above, $X$ is the exposure, $Y$ is the outcome variable, $G$ is the instrument, and $U$ represents potentially unmeasured confounding. Note that the association between $G$ and $X$ need not be causal, but we can assume it here for simplicity without losing any generality.

Formally, an instrumental variable (IV) is a third variable in regression analysis that is correlated with both exposure and outcome, but affects the outcome only through its association with the exposure. A *valid* instrument follows the causal model depicted in Figure 1. An IV thus acts as a proxy for the exposure that is not susceptible to the same degree of confounding. A key challenge in instrumental variable methods is finding the right instrument(s) for performing the analysis (John et al., 2019). Due to the unmeasured confounding of the $X - Y$ association, this model cannot be elucidated from observed data unless we are willing to make strong assumptions about the generating process (Cornia and Mooij, 2014; Silva and Shimizu, 2017).

Genetic variants are particularly suitable as candidate instrumental variables, since they are fixed at conception and more robust against confounding due to environmental factors (Davey Smith et al., 2007). Nevertheless, the validity of genetic instruments is also not easily testable from data. To make matters worse, many genes affect multiple traits, meaning that the outcome variable $Y$ could be influenced by $G$ via different causal pathways. This violation of the instrumental variable assumptions is known in the Mendelian randomization literature as *horizontal pleiotropy* (Chesmore et al., 2018). Horizontal pleiotropy, to which we will refer from now simply as *pleiotropy*, is usually shown as a directed arrow from genetic variant ($G$) to outcome ($Y$) and the implied (direct) causal effect from $G$ to $Y$ is called a *pleiotropic effect*.

Searching for instruments in a haystack of potentially relevant genetic variants with unknown biological function is akin to a many-dimensional variable selection problem. To solve this problem, we adopt a *spike-and-slab* prior on the pleiotropic effects ($G \to Y$) to encourage sparse solutions through *selective shrinkage* (Ishwaran and Rao, 2005). The 'spike' captures the prior distribution of coefficients that are close to zero, corresponding to valid instruments, while the 'slab' models the prior distribution of coefficients that are significantly different from zero. Even though we do not know *a priori* which of the genetic variants are (close to being) valid instruments, by using the *wisdom of the crowd* (Surowiecki, 2005), where the crowd is the many-dimensional set of potential candidates, we are able to separate the wheat from the chaff, as we will later see in Section 5. We show that, as long as there are at least some valid instruments to be found in the haystack, the causal effect of interest can be reasonably estimated by using the proposed prior.

In this work, we consider a general Bayesian causal model subsuming the IV model in which a large number of (genetic) covariates have the potential to act as instrumental variables. We assume a hierarchical discrete scale mixture (spike-and-slab) prior on the pleiotropic effects to consider every possible combination of valid and invalid instruments. We then introduce an algorithm (`MASSIVE`) which we use to perform Bayesian model averaging (BMA) over this mixture space so as to properly handle the uncertainty in choosing the covariates to be used as instruments. The algorithm features two components: (1) a Markov Chain Monte Carlo Model Composition (`MC3`) stochastic search procedure (Madigan et al., 1995) and (2) an approximation procedure based on *Laplace's method* (Bishop, 2006) for determining the model evidence (marginal likelihood). We show the robustness and tractability of our approach in both simulated studies and real-world examples.

## 2 RELATED WORK

A number of methods have been suggested for selecting instrumental variables out of a rich set of candidates. Swerdlow et al. (2016) have outlined a set of principles for selecting instruments in MR analyses using a combination of statistical criteria and relevant biological knowledge. Belloni et al. (2012), on the other hand, have proposed a data-driven approach for model selection based on Lasso methods. Agakov et al. (2010) have built an approach for extracting the most reliable instruments by using approximate Bayesian inference with sparseness-inducing priors on linear latent variable models. Finally, Berzuini et al. (2020) have developed a Bayesian solution in which the horseshoe shrinkage prior is imposed on potential pleiotropic effects. These methods, however, are designed to select the most likely IV model and do not account for potential model uncertainty. Moreover, some of these methods require individual patient data, which is often unavailable, as input.

A number of model averaging solutions have also been proposed. Eicher et al. (2009) have used BMA to average over the set of potential models in the first stage of two-stage least squares (2SLS), which means that the selection of instruments is based on the strength of their association with the exposure. The model evidences are approximated using the *Bayesian information criterion* (Schwarz, 1978). Eicher et al. (2009) later extended their approach in (Lenkoski et al., 2014) by also accounting for model uncertainty in the second stage of 2SLS. In a similar vein, Karl and Lenkoski (2012) developed the `IVBMA` algorithm to incorporate model uncertainty into IV estimation by exploring the model space using stochastic search guided by analytically derived conditional Bayes factors. More recently, Shapland et al. (2019) have proposed using the `IVBMA` approach for Mendelian randomization with dependent instruments. The above-mentioned methods, however, work under the assumption that the chosen candidates are all valid instruments. This means that the algorithms are no longer consistent if any of the IV assumptions are violated.

Gkatzionis et al. (2019) have introduced a comparable Bayesian model averaging method (`JAM-MR`) in which genetic variants likely to exhibit horizontal pleiotropy, thereby violating the IV assumptions, are penalized via a pleiotropic-loss function. `JAM-MR` implements a standard reversible-jump MCMC stochastic search scheme for exploring the model space. However, the estimated causal effect for each model is obtained using the classical *inverse-variance weighted* (IVW) estimator (Burgess and Thompson, 2015), meaning that there is no complete description of the parameter uncertainty.

## 3 MODEL

Currently no published method offers a complete Bayesian solution for handling both the uncertainty in selecting the most promising candidates out of a many-dimensional set of potential instruments and the uncertainty in estimating the causal effect using those instruments. We propose to address this shortcoming with our `MASSIVE` (Model Assessment and Stochastic Search for Instrumental Variable Estimation) Bayesian approach, which is designed to reliably estimate the studied causal effect as long as at least one of the candidate instruments is close to valid. This condition is weaker than causal assumptions typically made in related work, e.g., a plurality of the candidate instruments are valid (the most common pleiotropic effect is zero) or the pleiotropic effects are balanced (on average they cancel each other out).

Our method incorporates Bayesian model averaging to further relax the IV causal assumptions by searching for the most plausible many-dimensional IV models, thereby properly accounting for uncertainty in the model selection. Our algorithm provides as output a posterior distribution over the causal effect that appropriately reflects the uncertainty in the estimate, as well as posterior inclusion probabilities indicating which candidates are likely to be valid instruments. Finally, our approach does not rely on having access to individual-level data, and instead can use publicly available summary data from large-scale GWAS as input. This constitutes a significant practical advantage, as access to information about individuals is often restricted, for instance due to privacy concerns (Pasaniuc and Price, 2017).

In our model, we assume that the data is generated from the following *structural equation model* (Bollen, 1989):

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

The associated generating model is depicted graphically in Figure 2. We are interested in estimating the (linear) causal effect from exposure ($X$) to outcome ($Y$), denoted by $\beta$. To aid estimation, we have measurements from $J$ covariates, denoted by $G_j$, at our disposal. Each covariate is associated in the model with both the exposure $X$, via the $\gamma_j$ parameters, and the outcome $Y$, via the $\alpha_j$ parameters. Finally, the unmeasured confounding is characterized by the coefficients $\kappa_X$ and $\kappa_Y$.

We assume that the noise terms of $X$, $Y$, and the unmeasured confounder $U$ are normally distributed. We can assume without loss of generality that $\epsilon_U \sim \mathcal{N}(0, 1)$ by appropriately rescaling the confounding coefficients. The exposure and outcome terms are normally distributed with unknown scale parameters, i.e., $\epsilon_X \sim \mathcal{N}(0, \sigma_X^2)$ and $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$. The random vector $(X, Y)|\boldsymbol{G}$ then follows the Conditional Gaussian distribution (*CG-distribution* in (Lauritzen and Wermuth, 1989)):

$$\begin{bmatrix} X \\ Y \end{bmatrix} \Bigg| \, \boldsymbol{G} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{G}), \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}(\boldsymbol{G}) = \begin{bmatrix} \boldsymbol{\gamma} & \beta\boldsymbol{\gamma} + \boldsymbol{\alpha} \end{bmatrix}^\intercal \boldsymbol{G}$ and $\boldsymbol{\Sigma} =$

$$= \begin{bmatrix} \sigma_X^2 + \kappa_X^2 & \beta(\sigma_X^2 + \kappa_X^2) + \kappa_X \kappa_Y \\ \beta(\sigma_X^2 + \kappa_X^2) + \kappa_X \kappa_Y & \sigma_Y^2 + \beta^2 \sigma_X^2 + (\kappa_Y + \beta \kappa_X)^2 \end{bmatrix}.$$

We now assume that $N$ independent and identically distributed observations $\mathbf{D} = (\boldsymbol{G}_i, X_i, Y_i)_{1 \le i \le N}$ are drawn

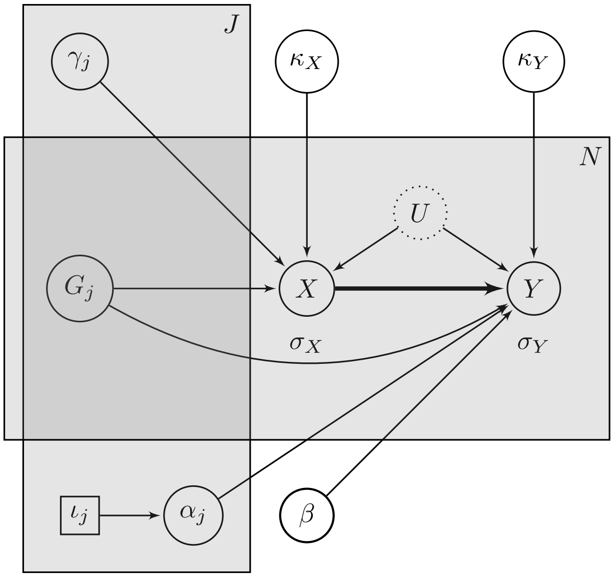

Figure 2: Graphical description of our assumed generative model. We denote the exposure variable by $X$ and the outcome variable by $Y$. We are interested in the causal effect from $X$ to $Y$, which is denoted by $\beta$. The association between $X$ and $Y$ is obfuscated by the unobserved variable $U$, which we use to model unmeasured confounding explicitly. The shaded plate indicates replication across the $J$ independent genetic variants $G_j, j \in \{1, 2, ..., J\}$. Note that the replication also applies to the parameters $\gamma_j$ and $\alpha_j$.

from the structural equation model described in (1). The conditional Gaussian observed data likelihood reads

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

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

### 3.1 PRIORS

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

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

For each scaled pleiotropic effect ($\tilde{\alpha}_j$), we propose a scale mixture of two normal distributions (Ishwaran and Rao, 2005), where the scale is determined by the value of a latent indicator variable $\iota_j$. The component with lower (higher) variance encompasses our prior belief that the pleiotropic effect is a priori 'weak' / irrelevant ('strong' / relevant). This hierarchical prior is identical to the one proposed by George and McCulloch (1993) for their Stochastic Search Variable Selection (SSVS) algorithm.

The standard deviation of the 'spike' (lower variance) component and of the 'slab' (higher variance) component can be set based on our prior knowledge or assumptions regarding the size of relevant and irrelevant parameters. For example, George and McCulloch (1993) have proposed a semiautomatic approach for selecting the spike-and-slab hyperparameters based on the intersection point of the two mixture components and the relative heights of the component densities at zero. For the more general situation when prior knowledge is not available, we propose a simple empirical approach for choosing these hyperparameters starting from the belief (assumption) that the measured interactions between $\boldsymbol{G}$ and $X$ are all relevant, which we can expect in most analyses since the first criterion by which potential instruments are chosen is the relevance of their association with the exposure. We describe the procedure for empirically determining prior hyperparameters in the supplement.

For the scaled instrument strengths $\tilde{\gamma}_j$, we propose a normal prior with the same variance as the slab component, under the mild assumptions that genetic interactions with different traits are of the same size and that the instrument strengths correspond are strong (relevant) interactions. For the causal effect ($\tilde{\beta}$) and the confounding coefficients ($\tilde{\kappa}_X$ and $\tilde{\kappa}_Y$), we choose a very weakly informative normal prior proposed by Gelman et al. (2020). For the scale parameters ($\sigma_X$ and $\sigma_Y$), we propose an improper uniform prior on the log-scale, corresponding to Jeffreys's scale-invariant prior (Gelman et al., 2013). The final Bayesian generating model is

$$\begin{gathered}\iota_j \sim \text{Bernoulli}(0.5);\\ \tilde{\alpha}_j \sim \iota_j \cdot \mathcal{N}(0, \sigma^2_{\text{slab}}) + (1 - \iota_j) \cdot \mathcal{N}(0, \sigma^2_{\text{spike}});\\ \tilde{\gamma}_j \sim \mathcal{N}(0, \sigma^2_{\text{slab}}); \quad \tilde{\beta} \sim \mathcal{N}(0, 10);\\ \tilde{\kappa}_X \sim \mathcal{N}(0, 10); \quad \tilde{\kappa}_Y \sim \mathcal{N}(0, 10);\\ p(\log\sigma_X) \propto 1; p(\log\sigma_Y) \propto 1;\\ \begin{bmatrix}X\\Y\end{bmatrix}\middle|\,\boldsymbol{G} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\gamma}^{\intercal}\boldsymbol{G}\\(\beta\boldsymbol{\gamma} + \boldsymbol{\alpha})^{\intercal}\boldsymbol{G}\end{bmatrix}, \boldsymbol{\Sigma}\right).\end{gathered} \tag{3}$$

### 3.2 BAYESIAN MODEL AVERAGING

In our approach we use the general framework of Bayesian Model Averaging to incorporate the uncertainty in instrument candidate validity by combining the causal effect estimates from reasonable instrument combinations. Instead of relying on a single model for estimating our causal effect $\beta$, we average the estimates over a number ($K$) of promising models, weighing each result by the model posterior

$$p(\beta|\mathbf{D}) = \sum_{k=1}^{K} p(\beta|M_k, \mathbf{D}) p(M_k|\mathbf{D}).$$

Our assumed generating model in (3) has $2J + 5$ parameters, $\boldsymbol{\Theta} = (\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\beta}, \tilde{\kappa}_X, \tilde{\kappa}_Y, \log\sigma_X, \log\sigma_Y)$, where $J$ is the number of candidates. There are $J$ latent indicator variables $\iota_j$ corresponding to the parameters $\tilde{\alpha}_j$ which indicate whether each parameter is 'weak' (generated by the 'spike' component) or 'strong' (generated by the 'slab' component). The full multivariate prior thus is a mixture of $K = 2^J$ multivariate Gaussian priors (the uniform prior on the log-scale parameters can be seen as a limiting case of a Gaussian prior). We refer to each mixture component as a different *model*. The difference between these models lies solely in the prior beliefs we assume on the pleiotropic effect strengths.

It is intractable to consider the entire space of $2^J$ models (multivariate indicator instances), so we instead search for a subset that best fits the data using MCMC Model Composition (MC3) (Madigan et al., 1995). If an unspecified subset of the $J$ candidates are close to being valid instruments, then only a small number of models will be a good fit to the data. We can thus obtain a good approximation of the model posterior probabilities without averaging over the entire model space. The idea of MC3 is to construct a Markov chain that moves through the class of models $\mathcal{M} = \{0, 1\}^J$. For each model $M$ we define a neighborhood consisting of the $J$ models that have only one indicator variable different than $M$, and we allow transitions only into the set of neighbors, with equal probability. A new model $M'$ in the neighborhood is then accepted with probability

$$\min\left\{1, \frac{p(M'|\mathbf{D})}{p(M|\mathbf{D})}\right\},$$

where $p(M|\mathbf{D})$ is the posterior probability of model $M$. The posterior probability is given by Bayes's theorem

$$p(M|\mathbf{D}) = \frac{p(\mathbf{D}|M)p(M)}{\sum p(\mathbf{D}|M')p(M')},$$

where

$$p(\mathbf{D}|M) = \int_{\boldsymbol{\Theta}} p(\mathbf{D}|\boldsymbol{\Theta}, M) p(\boldsymbol{\Theta}|M) \, \mathrm{d}\boldsymbol{\Theta}$$

is the model evidence. Here, the latent indicators $\iota_j$ are part of the model definition and their choice determines the parameter prior given the model, i.e., $p(\boldsymbol{\Theta}|M)$. As prior over the model space, we consider the simple uniform prior $p(M) = 2^{-J}$. This prior corresponds to the assumption that each parameter is as likely to be 'relevant' as 'irrelevant' a priori, i.e., $\iota_j \sim \mathrm{Bernoulli}(0.5)$ in (3). Other priors on the model space could be easily accommodated to indicate a prior belief in the presence or absence of pleiotropic effects.

A key challenge when considering a general approach such as the one proposed here is estimating the *evidence* (*marginal likelihood*) for each model. Since the integral is not analytically tractable for the proposed likelihood and priors, we have to resort to approximation methods. One idea would be to approximate the evidence with a *nested sampling* algorithm (Skilling, 2006), but this procedure is relatively slow, so we instead propose to approximate the evidence more efficiently using Laplace's method, similar to Rue et al. (2009).

# 4 ALGORITHM

## 4.1 FINDING THE POSTERIOR OPTIMA

When sampling a certain combination of indicator variables, we need to compute the corresponding approximate model evidence using Laplace's method. We need to find local posterior optima over the $2J + 5$ parameters $\widetilde{\boldsymbol{\Theta}} = (\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\beta}, \log\sigma_X, \log\sigma_Y, \tilde{\kappa}_X, \tilde{\kappa}_Y)$. Despite the simplicity of our chosen priors, we are dealing with a many-dimensional multimodal optimization problem. We tackle the issue by first separating our model parameters into those pertaining to observed variables, denoted by $\widetilde{\mathbf{B}} = (\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\beta}, \log\sigma_X, \log\sigma_Y)$, and those pertaining to the unobserved variable, denoted by $\widetilde{\mathbf{C}} = (\tilde{\kappa}_X, \tilde{\kappa}_Y)$.

To guide the optimization, we use the fact that for each value of the confounding coefficients in $\widetilde{\mathbf{C}}$, we can analytically derive the maximum likelihood estimate for $\widetilde{\mathbf{B}}$. For the details of deriving the ML estimate, please see the supplement. Thus, if we attempt to perform inference via maximum likelihood estimation, we arrive at a two-dimensional manifold of equally good solutions for the equation system. We propose to start the posterior optimization procedure from the bivariate ML manifold, for each considered model. We develop a smart procedure for choosing starting points on the manifold, described in the supplement, in which we look for (sparse) parameter combinations where some of the parameters are close to zero. The optimization initialization list $\mathcal{L}$ is given as input to the posterior approximation in Algorithm 1.

By analyzing the optimization results in the $\widetilde{\mathbf{C}}$ space, we have identified at most five local optima for each model. Note that these optima constitute pairs that are symmetric with respect to the origin. This is because the value of the posterior does not change if we replace $(\tilde{\kappa}_X, \tilde{\kappa}_Y)$ with $(-\tilde{\kappa}_X, -\tilde{\kappa}_Y)$. One possible optimum occurs at the critical point corresponding to the *no confounding* scenario, when the confounding coefficients are close to zero. We can find this optimum efficiently, if it exists, by starting the posterior optimization from the maximum likelihood parameters obtained when setting $\tilde{\kappa}_X = \tilde{\kappa}_Y = 0$.

### 4.2 COMPUTING THE APPROXIMATION

**Algorithm 1** `Approximate Posterior`

```
**Input:** data $\mathbf{Z} = [\boldsymbol{G}_i, X_i, Y_i]_{1 \le i \le N}$, model $M$, optimization initialization list $\mathcal{L}$
**for** $(\tilde{\kappa}_X, \tilde{\kappa}_Y)$ in $\mathcal{L}$ **do**
    $\widetilde{\boldsymbol{\Theta}}^{\text{ML}}$ = get_ML_estimate($\mathbf{Z}, \tilde{\kappa}_X, \tilde{\kappa}_Y$)
    $\widetilde{\boldsymbol{\Theta}}^{\text{MAP}}$ = optimize($posterior(\mathbf{Z}, M), \widetilde{\boldsymbol{\Theta}}^{\text{ML}}$)
    LA = Laplace_approximation($\widetilde{\boldsymbol{\Theta}}^{\text{MAP}}$)
    **Save:** $\widetilde{\boldsymbol{\Theta}}^{\text{MAP}}(\tilde{\kappa}_X, \tilde{\kappa}_Y)$, LA($\widetilde{\boldsymbol{\Theta}}^{\text{MAP}}$)
**end for**
Eliminate potential duplicates from optima list;
Compute total model evidence from LA list;
**Output:** Mixture of LA($\widetilde{\boldsymbol{\Theta}}^{\text{MAP}}$), model evidence
```

We conjecture that there are at most five posterior local optima for any choice of latent indicator variables, which means that the mixture we intend to use as a posterior approximation will consist of at most five Laplace approximations. We can simplify the optimization by using only three preset initialization points (please see details in supplement) and symmetry. This is typically sufficient to find all the local optima in the full parameter space, or at least the global posterior mode. In Figure 3, we show an example of posterior surface for which all five local optima are present. The posterior is projected in the confounder space by computing the optimal posterior value for each pair of values $(\tilde{\kappa}_X, \tilde{\kappa}_Y)$. We use the results from the posterior optimization described above to construct an approximation to the posterior density using Laplace's method. We apply the method to each of the (at most five) local optima and then approximate the model evidence by computing the normalization constant for the approximate (unnormalized) posterior, which is a mixture of Laplace approximations (output of Algorithm 1).

### 4.3 SAMPLING OVER IV MODELS

We use the approximated model evidence in a MC3 scheme to search over the different models. To improve the sampling over causal models, we first run a greedy search procedure to arrive at a good (high-evidence) starting model. The approximations computed during this phase are cached and passed on to the MC3 stochastic search, after which we prune the explored model list in line with *Occam's window* (Madigan and Raftery, 1994) and average over the remaining IV models. By pruning out very-low probabilities estimated models, we arrive at a slimmer, less noisy, and more robust BMA posterior. Finally, we sample the causal effect estimates from the derived BMA posterior distribution. The full set of steps are shown in Algorithm 2.

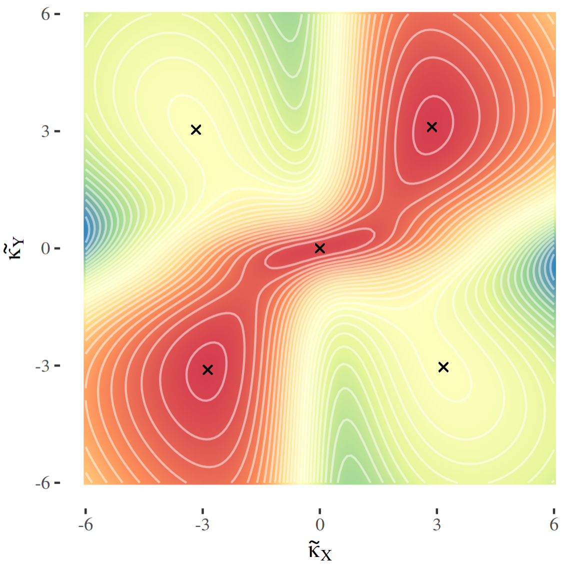

Figure 3: Surface of parameter posterior projected in confounder space, in which five local optima (the black X marks) can be observed.

**Algorithm 2** `MASSIVE` (Model Assessment and Stochastic Search for Instrumental Variable Estimation)

```
**Input:** data $\mathbf{Z} = [\boldsymbol{G}_i, X_i, Y_i]_{1 \le i \le N}$
$greedy\_start$ = greedy_search($\mathbf{Z}$)
$model\_list$ = MC3_search($\mathbf{Z}$, $greedy\_start$)
$pruned\_list$ = prune($model\_list$)
$BMA\_posterior$ = average($pruned\_list$)
$posterior\_samples$ = sample($BMA\_posterior$)
**Output:** $BMA\_posterior, posterior\_samples$
```

## 5 EMPIRICAL RESULTS

In this experiment we show that our algorithm is accurate in predicting the (lack of) causal effect from $X$ to $Y$ when there are least some measured variables that can act as potential instruments. The first and second order statistics for the observed variables $(\boldsymbol{G}, X, Y)$ are sufficient statistics for computing the likelihood specified in Equation (2). If individual-level data is not available, the sufficient statistics can also be derived from summary (regression) data, as shown in the supplement. This means that our approach can leverage the public results obtained from large-sample GWAS.

The *selective shrinkage* property of the Gaussian scale mixture leads to an improved causal effect estimate in the scenario under investigation. Without any priors on the pleiotropic effects, the problem is undetermined and for all values of $(\tilde{\kappa}_X, \tilde{\kappa}_Y)$ we can find a set of parameters that maximizes the data likelihood (please see supplement). By introducing sparsifying priors on the parameters, however, the symmetry among these different sets is broken, leading to a preference for smaller values. The key advantage of the 'spike-and-slab' prior is the ability to distinguish between relevant and irrelevant effects. We illustrate this difference in Figure 4. With the spike-and-slab prior, we obtain a much more confident estimate compared to when using a Gaussian prior. In practice, we do not know which of the pleiotropic effects are relevant and which are irrelevant, but with our `MASSIVE` BMA approach, we can infer this distinction from data, thereby significantly improving the causal effect estimate.

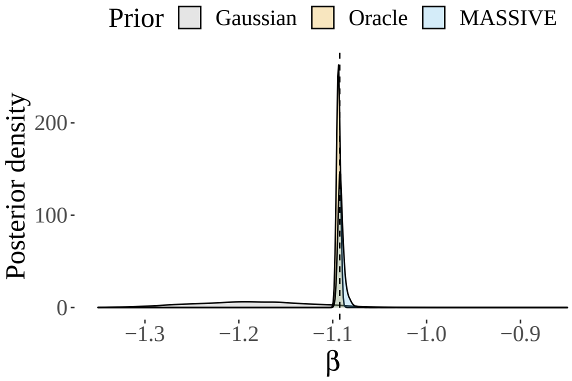

Figure 4: Comparison of estimated causal effect with different sparsifying priors when five out of 50 candidates are valid instruments. The true causal effect value ($\beta = -1.093$) is indicated with a dashed vertical line. **Gaussian**: We estimate a single model with fixed Gaussian priors on the genetic associations. **Oracle:** We estimate a single model with a spike-and-slab prior, where the latent indicators on the pleiotropic effects are chosen to correspond to the ground truth, i.e., $\iota_j = 0$ if the effect is irrelevant and $\iota_j = 1$ if it is relevant. **MASSIVE:** We use a spike-and-slab prior over the pleiotropic effects and learn the latent indicators with BMA.

We simulated two different scenarios starting from the setup described in (Gkatzionis et al., 2019): one in which there is no causal effect ($\beta = 0$), and one in which there is a strong positive causal effect ($\beta = 0.3$). The other simulation parameters we varied are the number of generated observations $N$ and the noise $\sigma$, which characterizes the degree of both intrinsic noise and confounding. We considered three simulation configurations: (1) $N = 10^3, \sigma = 1$ (less data, less noise); (2) $N = 10^3, \sigma = 4$ (less data, more noise); and (3) $N = 10^5, \sigma = 4$ (more data, more noise). The full parameters specifications for the linear SEM from Equation (1) used in the simulated experiments are outlined in (4).

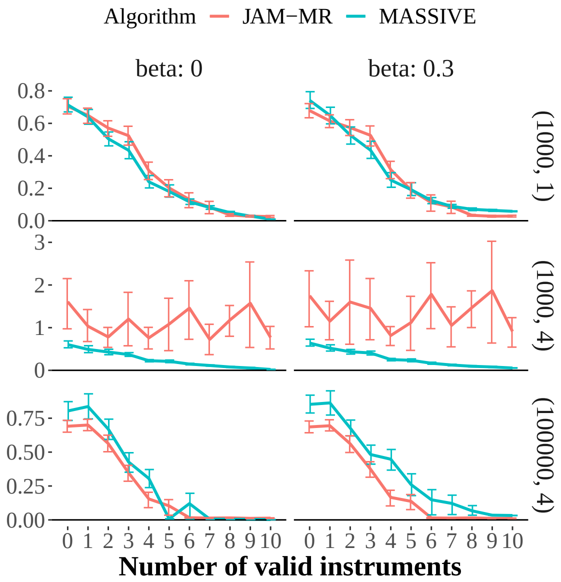

Figure 5: Comparison of `MASSIVE` and `JAM−MR` results averaged over one hundred simulated data sets. `MASSIVE` returns a posterior distribution, unlike `JAM−MR` which outputs point estimates. For `MASSIVE`, we took the median value as the causal point estimate for each data set. We then computed the root mean squared error (RMSE) of the different point causal estimates for both algorithms, as well as the bootstrapped RMSE confidence interval. We ran `JAM−MR` using the default settings, according to which a grid search is used to set the tuning parameter $w$ (Gkatzionis et al., 2019).

$$\begin{aligned} N \in \{10^3, 10^5\}; J = 10; \quad K \in \{1, 2, ..., J\};\\ \forall j \ p_j \sim \mathcal{U}(0.1, 0.9);\\ \forall j \ \gamma_j \sim 0.5 + |\mathcal{N}(0.0, 0.5^2)|;\\ \forall j \ \alpha_j \sim \pm \mathbf{1}_{j \leq K} \mathcal{N}(0, 0.2^2);\\ \beta \in \{0, 0.3\};\\ \kappa_X = \kappa_Y = \sigma_X = \sigma_Y = \sigma \in \{1, 4\}. \end{aligned} \tag{4}$$

We illustrate the simulation results in Figure 5, where we compare our approach against the competing `JAM−MR` algorithm (Gkatzionis et al., 2019). We report the *root mean square error* (RMSE) as a measure of estimation precision. As expected, the estimate generally improves with the number of valid instruments and with noise reduction for both algorithms. In the first configuration, the (potential) instruments are strong, accounting for about

60% of the variability in $X$, while in the other two configurations, they are weak, accounting for around 10% of the variability. The last configuration is typical for MR studies, which are characterized by large sample sizes but small genetic associations (Davey Smith and Hemani, 2014). Our approach is competitive in the first (less data, less noise) and third (more data, more noise) configuration, and much more robust than `JAM-MR` for the second configuration (less data, more noise).

# 6 REAL-WORLD APPLICATIONS

## 6.1 DETERMINANTS OF MACROECONOMIC GROWTH

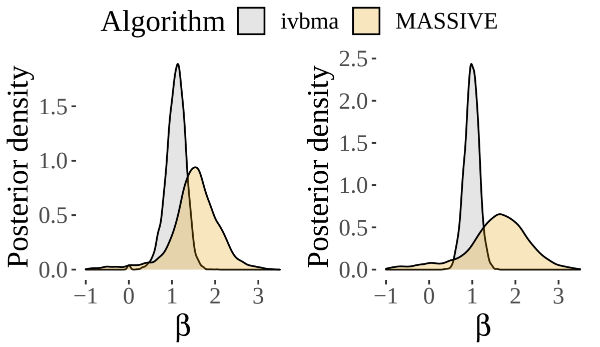

Figure 6: **Left:** Estimated effect of institutions (rule of law) on macroeconomic growth **Right:** Estimated effect of economic integration on macroeconomic growth. We used empirically determined values for the `MASSIVE` hyperparameters $\sigma_{\text{slab}}$ and $\sigma_{\text{spike}}$.

In this experiment, we use `MASSIVE` to model uncertainty in macroeconomic growth determinants on a data set compiled by Rodrik et al. (2004). This data set has been previously analyzed by Karl and Lenkoski (2012) using the `IVBMA` approach. The goal of the analysis was to find the best determinants (markers) of macroeconomic growth. Karl and Lenkoski (2012) found strong evidence indicating *institutions*, as measured by the strength of rule of law, and *economic integration* as the leading determinants of macroeconomic growth. In their analysis, they split the data into the two endogenous variables (exposures), rule of law and integration, four potential instrumental variables and 18 additional covariates. The authors treat these two types of variables distinctly in their model: the instrumental variables are only associated with the exposure, while the covariates are associated with both exposure and outcome. In our model, these two types of variables are considered the same as we do not make any assumptions regarding the candidates' validity a priori, but instead attempt to learn it from the data. Since the `IVBMA` model does not include location parameters, an intercept term is included in the data set, which we also use when running `MASSIVE`. In Figure 6 we compare the results obtained with `MASSIVE` and `IVBMA` on the macroeconomic growth data set. The output of `MASSIVE` is in line with previously computed estimates and provides further evidence for a significant causal effect of institutions (rule of law) and economic integration on macroeconomic growth.

## 6.2 INVESTIGATING THE RELATIONSHIP BETWEEN BMI AND PSORIASIS

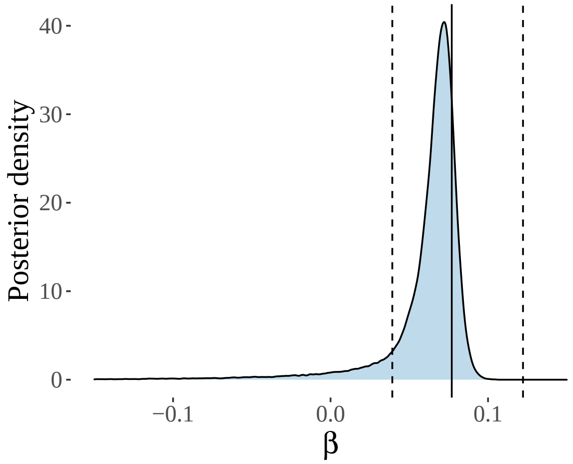

Figure 7: The posterior estimate of the causal effect of BMI, measured in $\mathrm{kg\,m^{-2}}$, on the log-odds of psoriasis risk obtained when running the `MASSIVE` algorithm with empirically determined hyperparameters. The vertical lines correspond to the results (estimate and error bars) from the IV analysis performed by Budu-Aggrey et al. (2019) on the UK Biobank data set.

Psoriasis is a common, chronic inflammatory skin disease, which affects approximately 2-4% of the population. Psoriasis is presumed to be influenced by both genetic and environmental risk factors, as there are a number of recognized determinants such as family history, smoking, stress, obesity, and alcohol consumption (Parisi et al., 2013). Establishing a causal link between obesity and psoriasis would be of great clinical interest both for understanding the precise mechanism underlying the association and for guiding treatment recommendations. Recently, Budu-Aggrey et al. (2019) have attempted to quantify this putative causal relationship by performing an instrumental variable analysis using 97 *single-nucleotide polymorphisms* (SNPs) associated with the *Body Mass Index* (BMI), a common measure of obesity, as genetic instruments. Their study provides evidence that higher BMI leads to a higher risk of psoriasis. The authors report that “higher BMI causally increased the odds of psoriasis by 9% per 1 unit increase in BMI”.

In this experiment, we have reproduced their analysis using the MASSIVE algorithm. We have applied our approach on the UK Biobank data set analyzed in (Budu-Aggrey et al., 2019), containing 5,676 psoriasis cases and 372,598 controls. Our algorithm returned 58 models, which were used to compute the model mixture posterior approximation. We then sampled $10^5$ parameter posterior samples from the mixture. In Figure 7 we show the posterior density estimate for the causal effect $\beta$. The result obtained is very similar to that reported in (Budu-Aggrey et al., 2019). It provides further evidence for increased BMI leading to a higher occurrence of psoriasis.

# 7 DISCUSSION

It is crucial to take model uncertainty into account when making inferences so as to mitigate the pitfalls of model misspecification (Hoeting et al., 1999). Bayesian Model Averaging (BMA) is a principled approach of incorporating this uncertainty into the analysis, but it is limited in scope due to the intractability of evaluating the model evidence for a considerable number of interesting models. In light of the computational limitations, the researcher often turns to approximating the evidence, but common solutions such as the BIC approximation might not be suitable for complex models (Fragoso et al., 2018). Through a combination of clever model choices and a hybrid inference scheme, combining MC3 stochastic search with fast Laplace approximations, MASSIVE is the first algorithm that can provide a reliable posterior estimate of the causal effect in IV settings with hundreds of candidate instruments.

Our proposed model provides a flexible and general solution for instrumental variable analyses. Thanks to the ‘spike-and-slab’ type prior on the interaction strengths, potential background knowledge regarding the sparsity and effective size of interactions can easily be incorporated into the model in an intuitive fashion. In this work, we have chosen to model the confounding coefficients explicitly in order to provide a unified view of causal interactions. Another possibility would have been to model the confounding effect as variance terms in a correlated errors model (Jones et al., 2012), a possibility we leave for future work.

In our Bayesian approach, we have proposed simple but flexible priors both over the model and parameter space to permit a more accurate approximation of the posterior using Laplace’s method. This approach allows for a tractable search through the model space, and parameter samples can be immediately derived from the approximation. The approach also lends itself to straightforward parallelization. In future work we plan to refine and speed up the process by, for example, including more starting points in the optimization procedure and distributing them across multiple cores. Furthermore, there is great potential in combining our approach with other means of (pre-)selecting instruments such as Lasso-based methods (Belloni et al., 2012) or the sparse IV (SPIV) approach (Agakov et al., 2010).

### Acknowledgements

This research has been partially financed by the Netherlands Organisation for Scientific Research (NWO) under project 617.001.451 (CHiLL).

### References

Agakov, F. V. et al. (2010). “Sparse Instrumental Variables (SPIV) for Genome-Wide Studies”. In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., pp. 28–36.

Angrist, J. D. et al. (1996). “Identification of Causal Effects Using Instrumental Variables”. In: *J. Am. Stat. Assoc.* 91.434, pp. 444–455.

Belloni, A. et al. (2012). “Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain”. In: *Econometrica* 80.6, pp. 2369–2429.

Benson, K. and A. J. Hartz (2000). “A Comparison of Observational Studies and Randomized, Controlled Trials”. In: *N. Engl. J. Med.* 342.25, pp. 1878–1886. pmid: 10861324.

Berzuini, C. et al. (2020). “A Bayesian Approach to Mendelian Randomization with Multiple Pleiotropic Variants”. In: *Biostatistics* 21.1, pp. 86–101.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer-Verlag.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Budu-Aggrey, A. et al. (2019). “Evidence of a Causal Relationship between Body Mass Index and Psoriasis: A Mendelian Randomization Study”. In: *PLoS Med* 16.1. pmid: 30703100.

Burgess, S. and S. G. Thompson (2015). *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. CRC Press. 222 pp. Google Books: WYSbBgAAQBAJ.

Chesmore, K. et al. (2018). “The Ubiquity of Pleiotropy in Human Disease”. In: *Hum Genet* 137.1, pp. 39–44.

Cornia, N. and J. M. Mooij (2014). “Type-II Errors of Independence Tests Can Lead to Arbitrarily Large Errors in Estimated Causal Effects: An Illustrative Example”. In: *Proceedings of the UAI 2014 Conference on*

*Causal Inference: Learning and Prediction - Volume 1274*. CI'14. Quebec City, Canada: CEUR-WS.org, pp. 35–42.

Davey Smith, G. and G. Hemani (2014). "Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies". In: *Hum Mol Genet* 23.R1, R89–R98.

Davey Smith, G. et al. (2007). "Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology". In: *PLOS Medicine* 4.12, e352.

Eicher, T. S. et al. (2009). *Bayesian Model Averaging and Endogeneity Under Model Uncertainty: An Application to Development Determinants*. Working Paper UWEC-2009-19-FC. University of Washington, Department of Economics, p. 29.

Fragoso, T. M. et al. (2018). "Bayesian Model Averaging: A Systematic Review and Conceptual Classification". In: *Int. Stat. Rev.* 86.1, pp. 1–28.

Gelman, A. et al. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. et al. (2020). *Prior Choice Recommendations · Stan-Dev/Stan Wiki*.

George, E. I. and R. E. McCulloch (1993). "Variable Selection via Gibbs Sampling". In: *J. Am. Stat. Assoc.* 88.423, pp. 881–889.

Gkatzionis, A. et al. (2019). "Bayesian Variable Selection with a Pleiotropic Loss Function in Mendelian Randomization". In: *bioRxiv*, p. 593863.

Hingorani, A. and S. Humphries (2005). "Nature's Randomised Trials". In: *The Lancet* 366.9501, pp. 1906–1908. pmid: 16325682.

Hoeting, J. A. et al. (1999). "Bayesian Model Averaging: A Tutorial". In: *Stat. Sci.* 14.4, pp. 382–401. JSTOR: 2676803.

Ishwaran, H. and J. S. Rao (2005). "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies". In: *Ann. Stat.* 33.2, pp. 730–773.

John, E. R. et al. (2019). "Assessing Causal Treatment Effect Estimation When Using Large Observational Datasets". In: *BMC Med Res Methodol* 19.1, p. 207.

Jones, E. M. et al. (2012). "On the Choice of Parameterisation and Priors for the Bayesian Analyses of Mendelian Randomisation Studies". In: *Stat. Med.* 31.14, pp. 1483–1501.

Karl, A. and A. Lenkoski (2012). "Instrumental Variable Bayesian Model Averaging via Conditional Bayes Factors". In: arXiv: 1202.5846 [stat].

Lauritzen, S. L. and N. Wermuth (1989). "Graphical Models for Associations between Variables, Some of Which Are Qualitative and Some Quantitative". In: *Ann. Stat.* 17.1, pp. 31–57. JSTOR: 2241503.

Lawlor, D. A. et al. (2008). "Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology". In: *Stat. Med.* 27.8, pp. 1133–1163.

Lenkoski, A. et al. (2014). "Two-Stage Bayesian Model Averaging in Endogenous Variable Models". In: *Econom. Rev.* 33.1-4, pp. 122–151.

Madigan, D. and A. E. Raftery (1994). "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window". In: *J. Am. Stat. Assoc.* 89.428, pp. 1535–1546.

Madigan, D. et al. (1995). "Bayesian Graphical Models for Discrete Data". In: *Int. Stat. Rev.* 63.2, p. 215. JSTOR: 1403615?origin=crossref.

Parisi, R. et al. (2013). "Global Epidemiology of Psoriasis: A Systematic Review of Incidence and Prevalence". In: *Journal of Investigative Dermatology* 133.2, pp. 377–385.

Pasaniuc, B. and A. L. Price (2017). "Dissecting the Genetics of Complex Traits Using Summary Association Statistics". In: *Nat Rev Genet* 18.2, pp. 117–127.

Rodrik, D. et al. (2004). "Institutions Rule: The Primacy of Institutions Over Geography and Integration in Economic Development". In: *Journal of Economic Growth* 9.2, pp. 131–165.

Rue, H. et al. (2009). "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71.2, pp. 319–392.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *Ann. Statist.* 6.2, pp. 461–464.

Shapland, C. Y. et al. (2019). "A Bayesian Approach to Mendelian Randomisation with Dependent Instruments". In: *Stat. Med.* 38.6, pp. 985–1001.

Sheehan, N. A. et al. (2008). "Mendelian Randomisation and Causal Inference in Observational Epidemiology". In: *PLoS Med* 5.8. pmid: 18752343.

Silva, R. and S. Shimizu (2017). "Learning Instrumental Variables with Structural and Non-Gaussianity Assumptions". In: *J. Mach. Learn. Res.* 18.1, pp. 4321–4369.

Skilling, J. (2006). "Nested Sampling for General Bayesian Computation". In: *Bayesian Anal.* 1.4, pp. 833–859.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Knopf Doubleday Publishing Group. 335 pp. Google Books: hHUsHOHqVzEC.

Swerdlow, D. I. et al. (2016). "Selecting Instruments for Mendelian Randomization in the Wake of Genome-Wide Association Studies". In: *Int J Epidemiol* 45.5, pp. 1600–1616.

Visscher, P. M. et al. (2017). "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1, pp. 5–22.