# Analysis of multiple specifications

Supporting Information

Christoph Semken David Rossell

01 April 2021

## Contents

This is the appendix includes extended methods for our paper "Analysis of multiple specifications: statistical validity and a Bayesian proposal". In Section 1 we briefly review the fundamentals behind Bayesian regression, Bayesian model selection (BMS) and Bayesian model averaging (BMA). In Section 2, we compare the statistical properties of the BMA estimator to those of the Specification Curve Analysis (SCA) median permutation test. The simulation study also shows the needed R code to obtain inference on individual treatment/outcome combinations and on averaged treatment effects.

The code to reproduce our empirical analysis and build a Bayesian Specification Curve Analysis (BSCA), robustness checks and a discussion of analytical choices can be found in our Open Science Framework repository at https://osf.io/m8d4n/.

# 1 Introduction to the Bayesian framework

This section provides a short introduction to the Bayesian regression, BMA and BMA frameworks. It is designed to allow readers who are unfamiliar with Bayesian statistics to follow the main text. We recommend anyone who wants to use BSCA to first get a deeper understanding of the Bayesian framework. Two excellent introductions are 'Statistical Rethinking' by McElreath (2020) and 'Bayesian Data Analysis' by Gelman et al. (2013). Two more applied texts are Hoeting et al.'s (1999) BMA tutorial and the open textbook (and course) 'An Introduction to Bayesian Thinking' by Clyde et al. (2020).

## 1.1 Bayesian regression

Bayesian statistics is based on a law of probability, known as Bayes theorem (or Bayes rule). It states that for two events $A$ and $B$ with non-zero probability, the probability of $A$ occurring given that $B$ occurred is

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

where $p(A)$ and $p(B)$ are the marginal (or unconditional) probabilities of $A$ and $B$ occurring, respectively.

To apply Bayes' rule to regression models, consider the standard linear regression model $y_i = \beta^T x_i + \alpha^T z_i + \epsilon_i$, where $y_i$ is the value of the dependent variable, $x_i$ contains one or more treatment variables of interest, $z_i$ are adjustment covariates, and $\epsilon_i \sim N(0, \sigma^2)$ are independent errors for observations $i = 1, \ldots, n$. The parameters $(\beta, \alpha, \sigma^2)$ describe the probabilistic dependence of the outcome $y_i$ on the treatment(s) $x_i$, after accounting for the effect of control covariates $z_i$. Specifically, $\beta$ is a vector with one or more parameters that captures the association between the outcome and the treatment(s) of interest. $\alpha$ is a parameter vector capturing the association with control covariates which, despite not being of immediate interest, is needed to reduce biases and variance when estimating $\beta$. Although we describe the linear regression setting for simplicity, an analogous construction applies to any other regression model, including the logistic regression model used in the main paper.

Applying Bayes rule, the information about $(\beta, \alpha, \sigma^2)$ after observing the data is contained in a posterior probability distribution that is described by the density function

$$\overbrace{p(\beta, \alpha, \sigma^2 \mid \text{data})}^{\text{posterior prob.}} = \frac{\overbrace{p(\text{data} \mid \beta, \alpha, \sigma^2)}^{\text{likelihood}} \overbrace{p(\beta, \alpha, \sigma^2)}^{\text{prior prob.}}}{p(\text{data})}$$

where 'data' denotes the observed data $y_1, x_1, z_1, \ldots, y_n, x_n, z_n$. Values of $(\beta, \alpha, \sigma^2)$ receiving higher $p(\beta, \alpha, \sigma^2 \mid \text{data})$ are more supported, *a posteriori* after observing the data, than values receiving lower $p(\beta, \alpha, \sigma^2 \mid \text{data})$.

Two important quantities in the above equation are the likelihood function $p(\text{data} \mid \beta, \alpha, \sigma^2)$ and the prior distribution on the parameters $p(\beta, \alpha, \sigma^2)$. The likelihood is the probability (or probability density, for continuous data) that we would observe our actual data, given the parameters. The denominator $p(\text{data})$ does not depend on $\beta, \alpha, \sigma^2$, it is a normalizing constant that we do not need to calculate directly and follows from the fact that (like all probability density functions) $p(\beta, \alpha, \sigma^2 \mid \text{data})$ integrates to 1.

The posterior information about our treatment effect coefficient(s) of interest $\beta$ given the data is contained in $p(\beta \mid \text{data})$, which can be obtained by integrating the posterior distribution with respect to $\alpha$ and $\sigma^2$, that is

$$p(\beta \mid \text{data}) = \int p(\beta, \alpha, \sigma^2 \mid \text{data}) d\alpha d\sigma^2.$$

The posterior $p(\beta \mid \text{data})$ contains all the probabilistic information needed in a Bayesian analysis to make inference on the treatment effect(s) $\beta$. In particular, one may obtain a point estimate by taking the posterior mean of $p(\beta \mid \text{data})$, and quantify uncertainty via a 95% posterior interval (an interval that is assigned 95% probability by $p(\beta \mid \text{data})$). Of particular importance for BSCA, $p(\beta \mid \text{data})$ can be expressed via BMA as a weighted average across models, as we outline next.

## 1.2 Bayesian model selection

A common issue in regression analysis – and one that SCA and BSCA try to address – is that there are many potential treatment and control covariates a researcher can choose from. Specifically, for a total of $p = J+Q$ variables, where $J$ is the number of treatments and $Q$ the number of controls there are $2^p$ models $M_1, \ldots, M_{2^p}$, corresponding to the $2^p$ configurations of the variables in $(x_i, z_i)$ that can potentially be included. Bayesian model selection (BMS) addresses this issue. It uses the posterior probability of the model to quantify the evidence in favor of a configuration of covariates being the optimal one (Kass and Raftery 1995). Here optimal refers to dropping any variable such that its regression coefficient in $\alpha$ is truly equal to 0, while preserving variables that are truly associated with the outcome (i.e. having a non-zero entry in $\alpha$). Two common strategies are then to either report estimates for the model with the highest posterior probability or use it to create weighted estimates. In Bayesian SCA we advocate for the latter, as described in the next section, so that any reported parameter estimates and intervals formally acknowledge the uncertainty in what is the right set of control variables.

The posterior probability of each model is given by Bayes' rule as

$$p(M_m \mid \text{data}) = \frac{p(\text{data} \mid M_m) p(M_m)}{p(\text{data})},$$

where $p(M_m)$ is a user-specified prior model probability and $p(\text{data} \mid M_m)$ is the so-called integrated likelihood (or marginal likelihood, or evidence) for model $M_m$. It can be computed using standard methods, either via closed-form expressions (when available), via deterministic approximations given by the Bayesian information criterion (BIC; Schwarz 1978), the extended BIC (EBIC; Chen and Chen 2008) or Laplace's method (Kass, Tierney, and Kadane 1990), or via stochastic approximations based on Markov Chain Monte Carlo methods (Friel and Wyse 2012).

For BSCA, again to avoid contentious prior choices, we use a uniform prior on the model size (the Beta-Binomial(1,1) prior; Scott and Berger 2010). This prior is uninformative in terms of how many covariates should be included in the regression. This also makes the estimation of the marginal likelihood computationally easy, since with these priors $p(\text{data} \mid M_m)p(M_m) \approx e^{-\frac{1}{2}\text{EBIC}_m}$, and hence we may use approximate posterior probabilities

$$p(M_m|\text{data}) \approx \frac{e^{-\frac{1}{2}\text{EBIC}_m}}{\sum_{m'=1}^{2^p} e^{-\frac{1}{2}\text{EBIC}_{m'}}}$$

where EBIC is the expected Bayesian information criterion (Chen and Chen 2008). Under fairly general conditions, the approximation becomes accurate as the sample size $n$ grows (Schwarz 1978; Rossell 2018 Section 3). As a result, the BSCA model score is just a scaled log-transformation of the EBIC, which is a correction of the widely-used BIC to prevent the inclusion of false positives when the number of treatments $J$ or controls $Q$ are large.

### 1.3 Bayesian model averaging

In the Bayesian SCA, we use Bayesian model averaging (BMA) to get a combined estimate for each coefficient of interest from a large number of models, as well as 95% intervals that consider all possible specifications. Specifically, BMA expresses the posterior distribution as the weighted average

$$p(\beta, \alpha, \sigma^2 \mid \text{data}) = \sum_{m=1}^{2^p} p(\beta, \alpha, \sigma^2 \mid M_m, \text{data})p(M_m \mid \text{data}),$$

where $p(\beta, \alpha, \sigma^2 \mid M_m, \text{data})$ is the posterior distribution on $(\beta, \alpha, \sigma^2)$ after setting a subset of elements in $\alpha$ to zero, and $p(M_m \mid \text{data})$ is the posterior probability of the model $M_m$. When the number of models $2^p$ is very large it is unfeasible to conduct the summation above exactly. In such situations one can use Markov Chain Monte Carlo methods, such as Gibbs sampling algorithms implemented in the R packages used in our examples.[1]

The point estimates and 95% intervals for treatment effects $\beta$ – obtained from the BMA-weighted posterior $p(\beta \mid \text{data})$ – have several desirable properties. Under general conditions $p(M_m \mid \text{data})$ converges to 1 (as $n$ grows) for the optimal model that only selects the covariates that are truly associated with the outcome (Dawid 1999). Under suitable mathematical conditions, this property holds even with large $p$ and or if the data are not exactly generated by the assumed regression model – e.g. due to omitted covariates, non-linear effects, or other potentially incorrect parametric assumptions (Rossell 2018). As a result, the BMA point estimate and 95% interval converge to the frequentist MLE and 95% confidence interval obtained under the optimal model (LeCam 1953; Vaart 1998 Chapter 10, Theorem 10.1).[2]

[1] See also Madigan and Raftery (1994) for a description of BMS and Hoeting et al. (1999) for a tutorial on BMA.

[2] Put differently, it is possible to establish the mathematical validity (from a frequentist statistics viewpoint) of the posterior probabilities $p(M_m \mid \text{data})$. Briefly, this follows from the fact that $p(M_m \mid \text{data})$ converges to 1 for the optimal model and to 0 for any other model as $n$ grows, Slutsky's theorem and standard Bernstein-von-Mises theorems.

## 1.4 Inference in BSCA

BSCA uses Bayesian regression and BMA to provide inference (point estimate, 95% posterior probability interval, and a hypothesis test) for each individual treatment-outcome combination, so that one can fully report their heterogeneity. BSCA can also provide point estimates and hypothesis tests on average treatment effects (ATE). In this section we outline how such inference is obtained.

We first discuss the BSCA for an individual treatment/outcome combination. Let $\beta_j$ be the parameter quantifying said association. Inference is based on the posterior distribution $p(\beta_j \mid \text{data})$, shown in the BSCA top left panel. The panel also shows the posterior mean $E(\beta_j \mid \text{data})$ as a point estimate and a 95% posterior interval. This posterior distribution is also used to perform a hypothesis test for $\beta_j = 0$, specifically when the posterior probability $P(\beta_j \neq 0 \mid \text{data})$ exceeds a threshold one rejects $\beta_j = 0$. We recommend the threshold $P(\beta_j \neq 0 \mid \text{data}) > 0.95$ based on the property that, when the expected value of $P(\beta_j \neq 0 \mid \text{data})$ is above 0.95 (across repeated sampling), then the type I error is guaranteed to be below 0.05 (Rossell 2018 Corollary 1). Moreover, our EBIC-based formulation guarantees that, if truly an effect $\beta_j = 0$ is not present, then the expectation of $P(\beta_j \neq 0 \mid \text{data})$ and the type I error rate converge to 0, as $n$ grows. See Section 1.5 for further discussion on false positive control.

The BSCA top right panel gives point estimates and 95% intervals given by the posterior distribution $p(\beta \mid M_m, \text{data})$ for each considered model (or the top 100 models, when the number of models $2^p$ is too large). This panel is analogous to standard SCA, except that we focus attention on the models more supported by the data. The BSCA middle panel gives the model scores, that is the posterior model probabilities $p(M_m \mid \text{data})$, for each configuration of adjustment covariates. Models are sorted decreasingly in terms of their posterior probabilities. The bottom panel mimicks that in SCA and indicates the covariates that correspond to each model.

The BMA estimate and 95% interval take into account the uncertainty arising from the many possible model specifications, and is asymptotically valid from a frequentist point of view, as explained above. We emphasize that a main motivation for the original SCA was that standard errors conditional on a single selected model fail to account for the model selection uncertainty (Simonsohn, Simmons, and Nelson 2020). This issue is naturally resolved in BSCA by using the BMA weights. We neither have to pick one model (possible resulting in outcome reporting bias) nor take a simple average over all models (many of which maybe relatively implausible given the data, leading to potentially large estimation biases).

In addition, researchers might want to calculate an average treatment effect (ATE). We discuss first reporting an ATE across multiple treatments $\beta_1, \ldots, \beta_J$, for a single outcome. It is common to define the ATE as the mean

$$\text{ATE} = \frac{1}{J}\sum_{j=1}^{J} \beta_j,$$

although it is also possible to consider other summaries, such as the median. Given a posterior distribution on the full $\beta$ vector, $p(\beta \mid \text{data})$, one also has an implied posterior distribution $p(\text{ATE} \mid \text{data})$. It is hence straightforward to obtain a point estimate for the ATE via the posterior mean $E(\text{ATE} \mid \text{data})$, intervals with 95% posterior probability under

$p(\text{ATE} \mid \text{data})$, and to reject the null hypothesis that ATE=0 when $P(\text{ATE} \neq 0 \mid \text{data}) > 0.95$. See the next section on how to obtain such inference in R. As described in the main paper, we recommend always reporting the individual treatment effects – at least their signs – when using the ATE.

Suppose now that one wishes to report a global ATE across $L > 1$ outcomes and $J \geq 1$ treatments. Let $\beta_{jl}$ be the regression coefficient associated to treatment $j \in \{1, \dots, J\}$ and outcome $l = \{1, \dots, L\}$. SCA defines the ATE as the median $\beta_{jl}$, here we consider the (perhaps more standard) definition based on the mean

$$\text{ATE} = \frac{1}{JL} \sum_{l=1}^{L} \sum_{j=1}^{J} \beta_{jl}.$$

A point estimate for the global ATE is given by the posterior mean

$$E(\text{ATE} \mid \text{data}) = \frac{1}{JL} \sum_{l=1}^{L} \sum_{j=1}^{J} E(\beta_{jl} \mid \text{data}).$$

Note that for linear regression the global ATE associated to regressing each individual outcome on treatment and controls is mathematically equivalent to the ATE associated to regressing the average outcome on the treatments and controls. The simulation study in our appendix illustrates how to exploit this property.

An important remark, which makes us caution against using the global ATE for statistical inference, is that it assigns equal weight to all outcomes. This may be inappropriate in situations where some of the outcomes are strongly correlated. For instance, suppose that there are $J = 10$ outcomes, 9 of which measure a very similar latent quantity (they are similar items within a questionnaire) whereas the tenth outcome measures an inherently different quantity. The global ATE will be mostly determined by outcomes 1-9, whereas intuitively one might want to discount their weight. Given that defining alternative global ATE's is a potentially contentious issue, in our examples we use the standard ATE definition, and recommend that BSCA users rely on outcome-specific results.

## 1.5 Statistical considerations and false positive control

As discussed, our EBIC-based formulation guarantees that, as the sample size $n$ grows, the total number of false discoveries across all tested treatment-outcome combinations converges to zero. See the simulation study in the next section for an empirical assessment of this property. In this section we discuss alternative Bayesian and non-Bayesian strategies to control false positives such as P-value adjustment and the False Discovery Rate.

Regarding alternative strategies, one could consider other Bayesian formulations that set different priors and still attain good properties (e.g. model selection consistency) as the sample size $n$ grows. For instance, one could set so-called Complexity priors on the model space (Castillo, Schmidt-Hieber, and Vaart 2015) or non-local priors on the regression coefficients (Johnson and Rossell 2010, 2012; Rossell and Telesca 2017). These are recent developments in the statistical literature to further improve the prevention of false positives in settings with many parameters or hypothesis tests, but given the large sample size in our examples we found that these refinements were not needed. It is also possible to refine our formulation for situations where one considers many outcomes. Briefly, the discussed Beta-Binomial/EBIC property that the probability of having any false positive (family-wise error rate, FWER) converges to 0 for large $n$, for each individual outcome, implies that the overall FWER across outcomes also converges to 0. In our experience the Beta-Binomial/EBIC in practice attains a very good FWER control, as long as the number of outcomes is moderate relative to $n$. In situations with a truly large number $L$ of outcomes one can extend the Beta-Binomial prior, by setting uniform prior probabilities to models that select $0, 1, \ldots, (J + q)L$ variables across all outcomes, where $q$ is the number of potential control variables. While these refinements are potentially interesting, we found them to be unnecessary in the considered applications.

We remark that one can also use non-Bayesian false positive control methods. First, note that BMA could be viewed simply as a mechanism to obtain a point estimate, both a global $\widehat{\text{ATE}}$ and $\hat{\beta}_{jl} = E(\beta_{jl} \mid \text{data})$ for individual parameters. One can then use a permutation test akin to that used in SCA to obtain a P-value for the ATE. Permutation tests can also provide P-values for individual parameters, and one could use standard P-value adjustment or False Discovery Rate control methods to prevent false positive inflation due to testing multiple $\beta_j$'s. See Benjamini and Hochberg (1995) and Efron (2007) for discussion on FWER and FDR control. Briefly, these methods ensure that the probability or proportion of false positives is below a pre-specified threshold, the default being the usual 0.05. That is, these methods are willing to admit a fixed positive probability of including regression parameters that are truly zero (similar to a P-value for a single test admitting a 0.05, say, false positive probability). In contrast the Beta-Binomial/EBIC formulation ensures that said probability converges to 0, that is as $n$ grows one discards all truly irrelevant parameters, which intuitively provides a stronger false positive control.

# 2 SCA and BSCA estimator and hypothesis testing properties: a simulation study

We illustrate the use of BMA via a simple simulation study to infer individual treatment effects, as well as the average treatment effect (ATE). We evaluate the bias, root mean squared estimation error (RMSE) associated to the BMA point estimator, and the type I error probability (false positive) and power for testing if an effect is indeed present. We provide the R code so that readers can easily modify the simulation parameters.

We first summarize the results of the results for one outcome. In our settings BMA had near-zero bias and its RMSE was an order of magnitude smaller than SCA estimates. Further, the estimated type I error probability and power for all coefficients were near 0 and 1, respectively. These high-quality results are due to using a relatively large sample size of $n = 1000$. We chose this value because the teenager well-being studies also had such large $n$ (actually larger), but we encourage readers to also assess performance in other settings. All scenarios include one control covariate that truly has an effect and is correlated with the treatment(s). In the multiple treatment case we consider effects of different magnitudes, to illustrate the power to detect smaller versus larger effects.

- Scenario 1 ($\beta_1 = 0$): `BMA bias = -7e-04, RMSE = 0.01, type I error = 0, SCA bias = 0.4979, RMSE = 0.4992`
- Scenario 2 ($\beta_1 = 1$): `BMA bias = -0.0027, RMSE = 0.044, power = 1, SCA bias = 0.4969, RMSE = 0.4984`
- Scenario 3 ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$): `BMA bias = 3e-04, RMSE = 0.0023, type I error = 0, SCA bias = 0.0751, RMSE = 0.0775`
- Scenario 4 ($\beta_1 = \beta_2 = 0, \beta_3 = 0.25, \beta_4 = 0.75, \beta_5 = \beta_6 = 1$): `BMA bias = -0.0027, RMSE = 0.0124, power = 1, SCA bias = 0.0549, RMSE = 0.0717`

Subsection 1 sets up some useful functions to run the simulations. Subsection 2 considers a simplest setting with a single treatment variable of interest. BMA and the SCA median are used to average inference across models, that is across possible specifications defined by each treatment/control having/not having an effect. Subsection 3 considers multiple treatments of interest. We show how to obtain inference for all individual treatment effects, so that one can assess and report on their heterogeneity. We also show how to estimate and test for the presence on an average treatment effect. Finally, Subsection 4 studies the properties of the BMA estimator in the case the ATE is taken over multiple outcomes.

## 2.1 Setup

We use the packages `mombf` and `specr` for BMA and SCA inference, respectively.

```
library(mombf)
library(specr)
```

We also load our auxiliary functions.

```
source('code/functions.R')
```

We create a function that simulates data from a Gaussian linear regression with $n$ observations: the outcome variable $y$, treatment(s) $x$ and control covariates $z$. The true value of the regression parameters for $x$ and $z$ are specified in $\beta$ and $\alpha$, respectively. The treatment(s) and controls are correlated, i.e. this is a situation where it is necessary to identify the relevant controls to avoid parameter estimation bias and increase the power of statistical tests.

```
simdata= function(n, beta, alpha, seed) {
 set.seed(seed)
 J= length(beta); q= length(alpha)
 x = matrix(rnorm(n*J),nrow=n) #simulate the value of treatment(s)
 xnames= paste('x',1:length(beta),sep='')
 if (q>0) { #if there are control covariates
   z = rowMeans(x) + matrix(rnorm(n*q),nrow=n) #correlated with treatments
   y = x %*% matrix(beta,ncol=1) + z %*% matrix(alpha,ncol=1) + rnorm(n)
   znames= paste('z',1:length(alpha),sep='')
   ans= data.frame(y,1,x,z)
   colnames(ans)= c('y','Intercept',xnames,znames)
 } else {   #if there are no control covariates
   y = x %*% matrix(beta,ncol=1) + rnorm(n)
   ans= data.frame(y,1,x)
   colnames(ans)= c('y','Intercept',xnames)
 }
 return(ans)
}
```

## 2.2 Single treatment

We consider two scenarios, both with a single treatment. In Scenario 1 the treatment has a truly zero effect, and a single control with a non-zero effect. This scenario assesses the type I error, that is the frequentist probability that BMA would wrongly claim the treatment effect to exist. In Scenario 2 the treatment has a truly non-zero effect, and we assess the statistical power of BMA to detect said effect.

In Scenario 1 we also illustrate that the type I error of SCA can be very large, despite using a permutation test to control for false positives. The reason is that SCA tests the median effect averaging across all covariate specifications. In our case there are two such specifications, depending on whether one includes the confounder or (wrongly) excludes it from the analysis. Although the median across these two specifications is non-zero, the true treatment effect is zero.

We use the bootstrap-based P-value proposed by Simonsohn, Simmons, and Nelson (2020) for non-experimental data. To facilitate comparisons, we clearly indicate each step laid out in that paper with comments in the code. A standard permutation test gave the same result (100% false positives).

### 2.2.1 Truly no treatment effect

BMA point estimates are stored in bmaest and its posterior probabilities on the presence of an effect in margpp. We also store in scaest the point estimate returned by a standard Specification Curve Analysis, which takes the median across all possible specifications.

```
n= 1000; beta= 0; alpha= 1; nsims= 100; nperm= 500;
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= scapval= double(nsims)
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=300+i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1],verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b= coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    # SCA median bootstrap-based permutation test
    # "(1) Estimate all K specifications with the observed data"
    sca= run_specs(
      df=data, y="y", x=xnames, controls=znames,
      model="lm", keep.results=TRUE
    )
    # remove duplicate models (bug in specr)
    sca= sca[!duplicated(sca[,c("x", "y", "controls")]),]
    scaest[i]= summarise_specs(sca)$median
    # "(2) Generate K different dependent variables under the null"
    yhat = apply(sca, 1, function(r) {
      t(data$y - r$res$coefficients[xnames]*data[xnames])
    })
    colnames(yhat) = paste("y", 1:ncol(yhat), sep="")
    data = cbind(data, yhat)
    medianperm= double(nperm)
    # Generate formulas for step 4 (using "predicted" y)
    formulas = apply(sca, 1, function(r) {r$res$call$formula})
    new_formulas = paste(
      colnames(yhat),
      sapply(formulas, function(f) {substring(f, 2)}, USE.NAMES=FALSE),
      sep=""
    )
    for (b in 1:length(medianperm)) {
      # "(3) Draw at random, and with replacement, N rows from this matrix,
      #       using the same drawn rows of data for all K specifications."
      dataperm = data[sample(1:n, replace=TRUE),]
      # "(4) Estimate the K specifications on the drawn data."
      coefperm = sapply(new_formulas, function(f) {
        lm(f, dataperm)$coefficients[xnames]
      })
      medianperm[b] = median(coefperm)
    }
    # "(6) ... Compute what percentage..."
    scapval[i]= mean(abs(medianperm) > abs(scaest[i]))
}
```

We report the estimated bias and RMSE. The BMA estimate has a bias close to zero, in contrast SCA tends to over-estimate the true parameter value $\beta_1 = 0$. The reason is that $x$

is correlated with the control $z$, which truly has an effect, hence the model including $x$ but not $z$ over-estimates $\beta_1$ (this follows from simple algebra and standard least-squares theory).

We also assess the type I error for testing the null hypothesis $\beta_1 = 0$ versus the alternative $\beta_1 \neq 0$. The BMA test rejects $\beta_1 = 0$ when the posterior probability $P(\beta_1 \neq 0 \mid y)$ is large, specifically $P(\beta \neq 0 \mid y) > 0.95$ guarantees that the type I error is below 0.05 (as the sample size $n \to \infty$). In most simulations $P(\beta \neq 0 \mid y)$ took a pretty small value (run `summary(margpp[,'x1'])` in R) and the null hypothesis was never rejected, i.e. the estimated type I error is 0.

```
bma.bias= mean(bmaest[,'x1'] - beta)
bma.rmse= sqrt(mean((bmaest[,'x1'] - beta)^2))
bma.reject= mean(margpp[,'x1']>0.95)
sca.bias= mean(scaest - beta)
sca.rmse= sqrt(mean((scaest - beta)^2))
sca.reject= mean(scapval< 0.05)
tab1= rbind(c(bma.bias, bma.rmse, bma.reject), c(sca.bias, sca.rmse, sca.reject))
rownames(tab1)= c('BMA','SCA')
colnames(tab1)= c('Bias','RMSE','type I error')
tab1
```

```
##              Bias       RMSE type I error
## BMA -0.0007119209 0.01000548            0
## SCA  0.4978690225 0.49915620            1
```

### 2.2.2 Truly non-zero treatment effect

We repeat the exercise with a non-zero treatment effect $\beta_1 = 1$. The table below repors the estimated bias, RMSE and power to detect that truly $\beta_1 \neq 0$. The null hypothesis is rejected in all simulations, hence the estimated power is 1.

```
n= 1000; beta= 1; alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=100+i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b= coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")[-1,]
    scaest[i]= summarise_specs(sca)$median
}
```

```
bma.bias= mean(bmaest[,'x1'] - beta)
bma.rmse= sqrt(mean((bmaest[,'x1'] - beta)^2))
bma.reject= mean(margpp[,'x1']>0.95)
sca.bias= mean(scaest - beta)
sca.rmse= sqrt(mean((scaest - beta)^2))
tab2= rbind(c(bma.bias, bma.rmse, bma.reject), c(sca.bias, sca.rmse, NA))
rownames(tab2)= c('BMA','SCA'); colnames(tab2)= c('Bias','RMSE','Power')
tab2
```

```
##              Bias       RMSE Power
## BMA -0.002661681 0.04396317     1
## SCA  0.496914120 0.49841051    NA
```

## 2.3 Multiple treatments

An interesting feature of SCA is visualizing the heterogeneity across multiple treatments/outcomes, and providing an averaged (or median) treatment effect over treatments/outcomes (and sets of control covariates). We show how to use BMA to estimate and test individual treatments, so that one can distinguish those with positive/zero/negative effects, as well as estimating an testing for the presence of an average treatment effect.

For simplicity we show an example with a single outcome and 6 treatments. We first consider a setting where all 6 treatments truly have a zero effect, so the ATE=0. This setting helps assess the probability that any individual treatment is falsely declared to have an effect. We then consider a second setting where 3 treatments truly have an effect and the remaining 2 treatments do not, which helps assess the statistical power of BMA tests on individual treatments, and on the ATE. More specifically, the ATE is usually defined as ATE$=(\sum_{j=1}^{6} \beta_j)/6$. We test this null hypothesis by obtaining the posterior probability $P(\text{ATE} \neq 0 \mid y) = P(\sum_j \beta_j \neq 0 \mid y)$, and rejecting the null hypothesis when said probability exceeds the 0.95 threshold.

We remark on an important property of our specific BMS implementation. In classical P-value based hypothesis tests the probability of claiming at least one false positive finding (family-wise type I error) increases as one adds more treatments, unless one uses more stringent P-value thresholds. In our BMS framework false positive inflation is avoided by using a Beta-Binomial(1,1) prior on the model space (or the EBIC approximation to model posterior probabilities, as outlined earlier). This formulation adds a penalization term as one increases the number of treatments (or covariates). The penalization ensures that as the sample size $n$ grows, the family-wise type I error probability converges to 0, see Chen and Chen (2008) or Rossell (2018) (Sections 3-4) for a mathematical proof. It also ensures that BMA parameter estimates and 95% intervals converge to those obtained via maximum likelihood estimation under the model that includes only the subset of treatments and controls truly associated with the outcome Rossell and Telesca (2017) (Proposition 3).

### 2.3.1 Zero average treatment effect

The simulation is as before, with the difference that beta has now 6 elements.

```r
n= 1000; beta= c(0,0,0,0,0,0); alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=200+i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b = coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #ATE
    bma.ate[i,]= getATE(ms, xvars=xnames, fun='mean')[c('estimate','margpp')]
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")
    scaest[i]= summarise_specs(unique(sca))$median
}
```

We first report the bias, RMSE and type I error probabilities for individual treatments. BMA did not declare as significant any individual treatment in any of the simulated datasets, that is both the individual and family-wise type I error probabilities are estimated to be near-zero. The individual treatment effects are not usually considered in SCA analyses.

```r
bias.indiv= rowMeans(t(bmaest[,xnames]) - beta)
rmse.indiv= sqrt(rowMeans((t(bmaest[,xnames]) - beta)^2))
reject.indiv= colSums(margpp[,xnames] > 0.95) / nrow(margpp)
tab3.indiv= rbind(bias.indiv, rmse.indiv, reject.indiv)
rownames(tab3.indiv)= c('Bias','RMSE','type I error')
round(tab3.indiv, 5)
```

```
##                    x1      x2      x3      x4       x5      x6
## Bias         -0.00028 0.00026 0.00069 0.00033 -0.00040 0.00127
## RMSE          0.00311 0.00597 0.00888 0.00319  0.00258 0.00868
## type I error  0.00000 0.00000 0.00000 0.00000  0.00000 0.00000
```

We next estimate the bias, RMSE and type I error associated to the ATE. In this simulation, the absolute bias and RMSE are more than 100 and 10 times larger for the median taken over all specifications, respectively.

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)

sca.biasate= mean(scaest - ate)
sca.rmseate= sqrt(mean((scaest - ate)^2))
tab3= rbind(c(bma.biasate, bma.rmseate, bma.rejectate),
            c(sca.biasate, sca.rmseate, NA))
rownames(tab3)= c('BMA','SCA')
colnames(tab3)= c('Bias','RMSE','type I error')
round(tab3, 4)
```

```
##       Bias   RMSE type I error
## BMA 0.0003 0.0023            0
## SCA 0.0751 0.0775           NA
```

### 2.3.2 Non-zero average treatment effect

Lastly we consider a setting where 2 treatments truly have no effect ($\beta_1 = \beta_2 = 0$) and 4 treatments have heterogeneous effects ($\beta_3 = 0.25$, $\beta_4 = 0.75$, $\beta_4 = 1$, $\beta_5 = 1$). Again, we include one control covariate that is correlated with the treatments and truly has an effect. Note that the true ATE and median treatment effects are both 0.5, to facilitate comparison between the BMA and SCA results. This is to facilitate comparison between BMA and SCA , since in our implementation BMA targets the ATE and SCA the median. One can use BMA to obtain inference on the median by using `getATE(ms, xvars=xnames, fun='median')` below.

```
n= 1000; beta= c(0,0,1/4,3/4,1,1); alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
```

```
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b = coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #ATE
    bma.ate[i,]= getATE(ms, xvars=xnames, fun='mean')[c('estimate','margpp')]
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")
    scaest[i]= summarise_specs(unique(sca))$median
}
```

We first report the bias, RMSE and type I error probabilities for individual treatments. BMA correctly detected that $\beta_1 = \beta_2 = 0$, and that $\beta_3 \neq 0$, $\beta_4 \neq 0$, $\beta_5 \neq 0$ and $\beta_6 \neq 0$ in all simulations.

```
bias.indiv= rowMeans(t(bmaest[,xnames]) - beta)
rmse.indiv= sqrt(rowMeans((t(bmaest[,xnames]) - beta)^2))
reject.indiv= colSums(margpp[,xnames] > 0.95) / nrow(margpp)
tab4.indiv= rbind(bias.indiv, rmse.indiv, reject.indiv)
rownames(tab4.indiv)= c('Bias','RMSE','Proportion rejected')
round(tab4.indiv, 5)
```

```
##                           x1       x2       x3       x4       x5       x6
## Bias                -0.00091 -0.00222 -0.00184 -0.00554 -0.00423 -0.00120
## RMSE                 0.01037  0.01586  0.03005  0.03370  0.03254  0.03484
## Proportion rejected  0.00000  0.01000  1.00000  1.00000  1.00000  1.00000
```

We next estimate the bias, RMSE and type I error associated to the ATE. Here BMA correctly detected that the ATE$\neq$ 0 in all simulations.
Again, the bias and RMSE are magnitudes larger for the median taken over all specifications than for the BMA average.

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)

sca.biasate= mean(scaest - ate)
sca.rmseate= sqrt(mean((scaest - ate)^2))
tab4= rbind(c(bma.biasate, bma.rmseate, bma.rejectate),
            c(sca.biasate, sca.rmseate, NA))
rownames(tab4)= c('BMA','SCA')
colnames(tab4)= c('Bias','RMSE','Power')
round(tab4, 4)
```

```
##        Bias   RMSE Power
## BMA -0.0027 0.0124     1
## SCA  0.0549 0.0717    NA
```

## 2.4 Multiple outcomes

Although we generally recommend running BSCA for each outcome individually and reporting the whole heterogeneity across outcomes, below we illustrate how to perform inference for a global ATE across $L$ outcomes and $J$ treatments. We define the global ATE in terms of the original outcomes $y$,

$$\text{ATE} = \frac{1}{JL} \sum_{l=1}^{L} \sum_{j=1}^{J} \beta_{lj}$$

where $\beta_{jl}$ is the regression parameter for treatment $j$ on outcome $l$.

We consider a setting with $L = 4$ outcomes, $J = 5$ treatments and $q = 1$ control variable. The outcomes are generated from a multivariate linear regression model where the errors are correlated, specifically $\epsilon \sim N(0, \Sigma)$ where $\Sigma$ has unit variances in the diagonal, pairwise correlations equal to 0.9 among outcomes 1-3, and 0.1 correlation with outcome 4. This correlation structure is meant to represent a situation where the first three outcomes can be thought of as measuring one common latent chararacteristic, that is different from that measured by the fourth outcome.

```
L= 4; J= 5; q= 1; n= 1000; nsims= 100
Sigma= diag(L)
Sigma[1,2]= Sigma[1,3]= Sigma[2,3]= Sigma[2,1]= Sigma[3,1]= Sigma[3,2]= 0.9
Sigma[1:3,4]= 0.1; Sigma[4,1:3]= 0.1
```

Function `simmultivdata` (created in the supplementary file `functions.R` ) generates such data $y$.

### 2.4.1 No average treatment effect

Our first simulation considers a setting where no treatment truly has an effect, hence ATE $=$ 0. The bias, RMSE to estimate $\text{ATE}_y$ are reported below, as well as the type I error rate, which is essentially zero.

```
beta= matrix(0,nrow=J,ncol=L); alpha= matrix(c(1,1,1,-1),nrow=q, ncol=L)

bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
xnames= paste('x',1:J,sep='')
for (i in 1:nsims) {
    data= simmultivdata(n=n,beta=beta,alpha=alpha,Sigma=Sigma,seed=300+i)
    y= as.matrix(data[, 1:L])
    m= rowMeans(y)
    ms = modelSelection(
      y=m, x=data[,-1:-L], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    bma.ate[i,]= getATE(ms, xvars=xnames)[c('estimate','margpp')]
}
```

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)
```

```
tab5= c(bma.biasate, bma.rmseate, bma.rejectate)
names(tab5)= c('Bias','RMSE','Type I error')
round(tab5, 4)
```

```
##         Bias         RMSE Type I error
##       0.0000       0.0018       0.0000
```

### 2.4.2 Non-zero average treatment effect

Next we consider a case where 3 treatments truly have an effect, whereas treatments 4-5 do not. The effect of treatments 1-3 on outcomes 1-3 is different to that on outcome 4, again mimicking a situation where outcomes 1-3 measure a common latent characteristic.

```
beta= matrix(0,nrow=J,ncol=L)
beta[,1]= beta[,2]= beta[,3]= c(1,1,1,0,0)
beta[,4]= c(1/4, 1/4, 1/4, 0, 0)
alpha= matrix(c(1,1,1,-1),nrow=q, ncol=L)
```

```
bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
xnames= paste('x',1:J,sep='')
for (i in 1:nsims) {
    data= simmultivdata(n=n,beta=beta,alpha=alpha,Sigma=Sigma,seed=300+i)
    y= as.matrix(data[, 1:L])
    m= rowMeans(y)
    ms = modelSelection(
      y=m, x=data[,-1:-L], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    bma.ate[i,]= getATE(ms, xvars=xnames)[c('estimate','margpp')]
}
```

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)
```

```
tab6= c(bma.biasate, bma.rmseate, bma.rejectate)
names(tab6)= c('Bias','RMSE','Power')
round(tab6, 4)
```

```
##    Bias    RMSE   Power
## -0.0010  0.0089  1.0000
```

# SI References

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society B* 57 (1): 289–300.

Castillo, I., J. Schmidt-Hieber, and A. W. van der Vaart. 2015. "Bayesian Linear Regression with Sparse Priors." *The Annals of Statistics* 43 (5): 1986–2018.

Chen, J., and Z. Chen. 2008. "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces." *Biometrika* 95 (3): 759–71.

Clyde, Merlise, Mine Cetinkaya-Rundel, Colin Rundel, David Banks, Christine Chai, and Lizzy Huang. 2020. *An Introduction to Bayesian Thinking*. https://statswithr.github.io/book/.

Dawid, A. P. 1999. "The Trouble with Bayes Factors." University College London.

Efron, Bradley. 2007. "Size, Power and False Discovery Rates." *The Annals of Statistics* 35 (4): 1351–77.

Friel, Nial, and Jason Wyse. 2012. "Estimating the Evidence–a Review." *Statistica Neerlandica* 66 (3): 288–308.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Boca Raton: Chapman and Hall/CRC. http://www.stat.columbia.edu/~gelman/book/BDA3.pdf.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. “Bayesian Model Averaging: A Tutorial.” *Statistical Science* 14: 382–401.

Johnson, V. E., and D. Rossell. 2010. “On the Use of Non-Local Prior Densities for Default Bayesian Hypothesis Tests.” *Journal of the Royal Statistical Society B* 72: 143–70.

———. 2012. “Bayesian Model Selection in High-Dimensional Settings.” *Journal of the American Statistical Association* 24 (498): 649–60.

Kass, R. E., L. Tierney, and J. B. Kadane. 1990. “The Validity of Posterior Expansions Based on Laplace’s Method.” *Bayesian and Likelihood Methods in Statistics and Econometrics* 7: 473–88.

Kass, Robert E, and Adrian E Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90 (430): 773–95.

LeCam, Lucien. 1953. “On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates.” *Univ. California Pub. Statist.* 1: 277–330.

Madigan, D., and A. E. Raftery. 1994. “Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam’s Window.” *Journal of the American Statistical Association* 89 (428): 1535–46.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. 2nd ed. Boca Raton: Chapman and Hall/CRC.

Rossell, D. 2018. “A Framework for Posterior Consistency in Model Selection.” *arXiv* 1806.04071: 1–58.

Rossell, D., and D. Telesca. 2017. “Non-Local Priors for High-Dimensional Estimation.” *Journal of the American Statistical Association* 112: 254–65.

Schwarz, G. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6: 461–64.

Scott, J. G., and J. O Berger. 2010. “Bayes and Empirical Bayes Multiplicity Adjustment in the Variable Selection Problem.” *The Annals of Statistics* 38 (5): 2587–2619.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. “Specification Curve Analysis.” *Nature Human Behaviour*, July. https://doi.org/10.1038/s41562-020-0912-z.

Vaart, A. W. van der. 1998. *Asymptotic Statistics*. New York: Cambridge University Press.