bayesselect — Bayesian variable selection for linear regression

Description

`bayesselect` implements Bayesian variable selection for linear regression. Bayesian variable selection uses special priors, global–local shrinkage or spike-and-slab priors, for regression coefficients to "select" variables. Unlike traditional variable-selection approaches, where each potential predictor is either included or not, `bayesselect` considers all predictors, but their impact in the full regression is controlled by the magnitudes of their random coefficients. `bayesselect` produces posterior summaries of regression coefficients and other model parameters using efficient Gibbs sampling. All Bayesian postestimation features (see [BAYES] **Bayesian postestimation**), including Bayesian predictions, are available after `bayesselect`.

Quick start

Bayesian variable selection for a linear regression with outcome `y` and potential predictors `x1` through `x10` using the default horseshoe prior for regression coefficients

```
bayesselect y x1-x10
```

Same as above, but use the Bayesian lasso prior for regression coefficients and display coefficients with inclusion values of 0.5 or above instead of the default of 0.1

```
bayesselect y x1-x10, blasso cutoff(0.5)
```

Variable selection using the Laplace spike-and-slab prior with scales of 0.1 and 10

```
bayesselect y x1-x10, sslaplace(0.1 10)
```

Variable selection using the normal spike-and-slab prior with default standard deviations of 0.01 and 1 and using the conjugate form of the prior

```
bayesselect y x1-x10, ssnormal conjugate
```

Show all 10 regression coefficients on replay

```
bayesselect, allcoef
```

Save current simulation results in external dataset `sim1.dta`

```
bayesselect, saving(sim1)
```

Menu

Statistics > Linear models and related > Bayesian regression > Variable selection for linear regression

Syntax

`bayesselect` *depvar indepvars* [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
noconstant	suppress constant term
Global–local shrinkage priors:	
hshoe	horseshoe prior with scale 1; the default
hshoe(#)	horseshoe prior with scale #
blasso	Bayesian lasso prior with scale 1
blasso(#)	Bayesian lasso prior with scale #
Spike-and-slab priors:	
ssnormal	mixture of normal priors with standard deviations 0.01 and 1
ssnormal(*#1* [*#2*])	mixture of normal priors with standard deviations *#1* and *#2*
sslaplace	mixture of Laplace priors with scales 0.01 and 1
sslaplace(*#1* [*#2*])	mixture of Laplace priors with scales *#1* and *#2*
betaprior(*#1* [*#2*])	beta prior with shapes *#1* and *#2* for hyperparameter θ of spike-and-slab priors; default is betaprior(1 1); requires ssnormal() or sslaplace()
conjugate	use conjugate form of priors for regression coefficients
normalprior(#)	specify standard deviation of default normal prior for constant term; default is normalprior(100)
prior(*priorspec*)	prior for some model parameters; this option may be repeated; not allowed for regression coefficients and latent parameters
dryrun	show model summary without estimation
Simulation	
nchains(#)	number of chains; default is to simulate one chain
mcmcsize(#)	MCMC sample size; default is mcmcsize(10000)
burnin(#)	burn-in period; default is burnin(2500)
thinning(#)	thinning interval; default is thinning(1)
rseed(#)	random-number seed
Blocking	
block(*paramref* [, *blockopts*])	specify a block of model parameters; this option may be repeated
blocksummary	display block summary
Initialization	
initial(*initspec*)	specify initial values for model parameters with a single chain
init#(*initspec*)	specify initial values for #th chain; requires nchains()
initall(*initspec*)	specify initial values for all chains; requires nchains()
nomleinitial	suppress the use of maximum likelihood estimates as starting values
initrandom	specify random initial values
initsummary	display initial values used for simulation

Reporting	
clevel(#)	set credible interval level; default is clevel(95)
hpd	display HPD credible intervals instead of the default equal-tailed credible intervals
cutoff(#)	specify cutoff inclusion value; default is cutoff(.1)
allcoef	display all coefficients; synonym for cutoff(0)
batch(#)	specify length of block for batch-means calculations; default is batch(0)
saving(*filename*[, replace])	save simulation results to *filename*.dta
nomodelsummary	suppress model summary
chainsdetail	display detailed simulation summary for each chain
[no]dots	suppress dots or display dots every 100 iterations and iteration numbers every 1,000 iterations; default is nodots
dots(#[, every(#)])	display dots as simulation is performed
notable	suppress estimation table
noheader	suppress output header
title(*string*)	display *string* as title above the table of parameter estimates
display_options	control spacing, line width, and base and empty cells
Advanced	
search(*search_options*)	control the search for feasible initial values
corrlag(#)	specify maximum autocorrelation lag; default varies
corrtol(#)	specify autocorrelation tolerance; default is corrtol(0.01)

indepvars and *paramref* may contain factor variables; see **[U] 11.4.3 Factor variables**.

indepvars may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

Only fweights are allowed; see **[U] 11.1.6 weight**.

Options noconstant and normalprior() may not be combined.

Options hshoe(), blasso(), ssnormal(), and sslaplace() may not be combined.

Options prior() and block() may be repeated.

priorspec and *paramref* are defined in [BAYES] **bayesmh**.

collect is allowed; see **[U] 11.1.10 Prefix commands**.

See **[U] 20 Estimation and postestimation commands** for more capabilities of estimation commands.

Model parameters are regression coefficients {*depvar*:*indepvars*} and error variance {sigma2}. For global–local shrinkage models, additional parameters are global shrinkage {tau} and latent predictor-specific local shrinkages {lambdas:*indepvars*}. For spike-and-slab models, additional parameters are latent predictor-specific Bernoulli inclusion indicators {gammas:*indepvars*} with success probability hyperparameter {theta}.

Options

Model

noconstant suppresses the constant term. This option may not be combined with option normalprior().

hshoe and hshoe(#) specify a horseshoe prior with respective scales of 1 and # for regression coefficients (excluding the intercept). hshoe is the default. The horseshoe prior belongs to the class of global–local shrinkage priors. Only one of options hshoe(), blasso(), ssnormal(), and sslaplace() may be specified. See *Global–local shrinkage priors* in *Methods and formulas*.

`blasso` and `blasso(`*#*`)` specify a Bayesian lasso prior with respective scales of 1 and # for regression coefficients (excluding the intercept). The Bayesian lasso prior belongs to the class of global–local shrinkage priors. Only one of options `hshoe()`, `blasso()`, `ssnormal()`, and `sslaplace()` may be specified. See *Global–local shrinkage priors* in *Methods and formulas*.

`ssnormal` and `ssnormal(`*#1* [*#2*]`)` specify a spike-and-slab mixture of two normal priors with respective standard deviations of 0.01 and 1 and of *#1* and *#2* for regression coefficients (excluding the intercept). Only one of options `hshoe()`, `blasso()`, `ssnormal()`, and `sslaplace()` may be specified. See *Spike-and-slab priors* in *Methods and formulas*.

`sslaplace` and `sslaplace(`*#1* [*#2*]`)` specify a spike-and-slab mixture of two Laplace priors with respective scales of 0.01 and 1 and of *#1* and *#2* for regression coefficients (excluding the intercept). Only one of options `hshoe()`, `blasso()`, `ssnormal()`, and `sslaplace()` may be specified. See *Spike-and-slab priors* in *Methods and formulas*.

`betaprior(`*#1* [*#2*]`)` specifies a beta prior with shapes *#1* and *#2* for the hyperparameter θ of spike-and-slab priors. The default is `betaprior(1 1)`, which is equivalent to a uniform prior on $[0, 1]$. This option requires one of option `ssnormal()` or `sslaplace()`. Option `betaprior()` can be used to control the sparsity of the regression model.

If you want to explore the effects of different `ssnormal()`, `sslaplace()`, and `betaprior()` priors on your results, it may be more convenient to specify only the first parameter value (and leave the second parameter value at the default 1), because the shapes of these priors are mainly controlled by the relative proportion between their two parameter values.

`conjugate` specifies a conjugate form of priors for regression coefficients. For global–local shrinkage and normal spike-and-slab priors, it includes the error variance parameter as a factor in the prior variances. For Laplace spike-and-slab priors, it includes the error standard deviation as a factor in the prior scale parameters. By default, `bayesselect` uses nonconjugate priors.

`normalprior(`*#*`)` specifies the standard deviation of the default normal prior for the constant term, the regression intercept. The default is `normalprior(100)`. This option may not be combined with option `noconstant`.

`prior(`*priorspec*`)` specifies a prior distribution for model parameters. For the syntax of *priorspec*, see *priorspec* in [BAYES] **bayesmh**. This option may be repeated. A prior may be specified for any of the model parameters, except the regression coefficients and latent parameters λ's and γ's, which use specialized priors. Model parameters that are not included in prior specifications are assigned default priors; see *Methods and formulas*. Model parameters with user-specified priors are not subjected to default blocking, which may cause suboptimal sampling efficiency. The block structure of model parameters can be inspected by using option `blocksummary`.

`dryrun` specifies to show the summary of the model that would be fit without actually fitting the model. This option is recommended for checking specifications of the model before fitting the model. The model summary reports the information about the likelihood model and about priors for all model parameters.

Simulation

`nchains()`, `mcmcsize()`, `burnin()`, `thinning()`, and `rseed()`; see *Options* in [BAYES] **bayesmh**.

Blocking

`block(`*paramref* [, *blockopts*]`)` and `blocksummary`; see *Options* in [BAYES] **bayesmh**. *blockopts* include `gibbs`, `split`, `scale()`, `covariance()`, and `adaptation()`.

Initialization

initial(), init#(), initall(), nomleinitial, initrandom, and initsummary; see *Options* in [BAYES] **bayesmh**.

Reporting

clevel() and hpd; see *Options* in [BAYES] **bayesmh**.

cutoff(#) specifies a cutoff inclusion value for regression coefficients. The default is cutoff(.1). Coefficients with inclusion values less than # are not shown in the coefficient table. The default is an arbitrary choice that allows you to see more predictors. In practice, a cutoff of 0.5 is often used to determine important predictors. The rationale behind the 0.5 cutoff is that it corresponds to the mean of the default prior distributions used for parameters that control the shrinkage. In general, a different cutoff may be considered whenever these default priors change; see *Remarks and examples* for details.

allcoef specifies that all regression coefficients be displayed in the coefficient table. This option is a synonym for cutoff(0).

batch(), saving(), nomodelsummary, chainsdetail, nodots, dots, dots(), notable, noheader, and title(); see *Options* in [BAYES] **bayesmh**.

display_options: vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), fvwrapon(*style*), and nolstretch; see [R] **Estimation options**.

Advanced

search(), corrlag(), and corrtol(); see *Options* in [BAYES] **bayesmh**.

Remarks and examples

Remarks are presented under the following headings:

Introductory examples
Diabetes progression study

Regression analysis, which models an outcome as a function of potential predictors, is one of the most popular methods in statistics. Variable selection can be viewed as a so-called sparse regression, in which only a small subset of predictors is relevant to the outcome. Identifying a subset of relevant predictors is important for multiple reasons. The first one is methodological. Variable selection provides a researcher with meaningful predictors, which improves interpretability of a model and helps pose more relevant causal hypotheses for a future study. Another benefit is inferential. Variable selection provides a more stable analysis that, as a result, improves the prediction power of the model. Finally, variable selection may also increase computational efficiency.

Consider a linear regression with outcome y and potential predictors $x_1, x_2, \dots, x_p$,

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \alpha + \epsilon$$

with a normal error term $\epsilon \sim N(0, \sigma^2)$ and error variance σ^2.

In a sparse linear regression, the majority of regression coefficients β_i's from the data-generating process are zeros. Identifying the nonzero coefficients is the primary problem of variable selection.

Let $\{y_i, x_{1i}, x_{2i}, \dots, x_{pi}\}$, $i = 1, 2, \dots, n$, be a data sample. A standard approach to variable selection is a penalized least-squares method. It involves minimizing a quantity of the form

$$l(\beta_1, \dots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{p} \phi(\beta_j)$$

where $\phi(\cdot)$ is a regularization function that penalizes deviation of regression coefficients from zero and λ is a penalty parameter. In lasso, Tibshirani (1996) uses $\phi(\beta_j) = |\beta_j|$ (l_1-penalization), and the irrelevant predictors are identified by coefficient estimates $\widehat{\beta_j}$'s that are strictly zero. Common difficulties in applying penalized least squares in practice are the choice of λ and obtaining valid standard errors for coefficient estimates.

In what follows, we assume basic knowledge of Bayesian analysis; see [BAYES] **Intro**.

A Bayesian variable-selection model is one that treats all regression coefficients as random variables with prior distributions designed to distinguish the importance of the corresponding predictor variables with respect to the observed data. For example, some suitable priors include penalty parameters that directly control the a priori assumed sparsity of the model. What makes the Bayesian approach to variable selection attractive is that it treats all regression and other model parameters, including penalty parameters, on an equal footing, as random quantities in one overall model, and controls them systematically through their prior distributions.

The Bayesian approach to variable selection is general and includes existing penalization-based methods as special cases. For example, a Bayesian formulation of the penalized least squares corresponds to finding the posterior mode for a model with independent regression coefficient priors of the form $\pi(\beta_i|\lambda) \propto \exp\{-\lambda\phi(\beta_i)\}$. But the mode is only one aspect of the posterior distribution, and the potential for full exploration of the available posterior distribution of parameters is one of the main strengths of Bayesian analysis.

Let's consider some of the priors for regression coefficients used in Bayesian variable selection. Regression coefficients are assumed to be continuous random parameters and are usually assigned continuous prior distributions. Thus, the prior probability for β_j to be zero is assumed to be zero, $\Pr(\beta_j = 0) = 0$. There are prior models that assign positive prior probabilities at zero, but because of estimation difficulties, these are rarely considered in practice. Continuous prior distributions for coefficients imply continuous posterior distributions. We thus have that the posterior probability for β_j to be zero is zero, $P(\beta_j = 0|y) = 0$. In contrast to solutions of some penalized least-squares approaches, where a coefficient is either zero or not, that is, the corresponding predictor is either included or not included, the inferential results of Bayesian variable selection provide degrees of inclusion for all predictors. This is similar to Bayesian model averaging (BMA; see [BMA] **Intro**), where the posterior probabilities of inclusion are reported and used to judge the importance of predictors.

There are two main classes of prior models for regression coefficients in Bayesian variable selection. One includes the global–local shrinkage priors (Carvalho, Polson, and Scott 2009; Griffin and Brown 2010; and Polson and Scott 2011). The other one includes the spike-and-slab priors, also known as two-group models (Johnstone and Silverman 2004; Efron 2008; and Castillo and van der Vaart 2012).

All the prior models under consideration introduce a set of latent (unobserved) parameters (λ's in global–local shrinkage priors and γ_j's in spike-and-slab priors), one for each coefficient β_j. Each latent parameter takes values between zero and one and describes the degree of inclusion of the predictor x_j. These latent parameters help interpret Bayesian variable-selection results. For example, with spike-and-slab priors, the prior for each regression coefficient is a mixture of two distributions,

$$\beta_j|\gamma_j \sim (1-\gamma_j)\phi_0(\beta_j) + \gamma_j\phi_1(\beta_j)$$

where $\phi_0(\cdot)$ and $\phi_1(\cdot)$ are two continuous distributions. Here γ_j's are random binary indicators and the degree of inclusion of x_j is measured by the marginal posterior probability $P(\gamma_j = 1|y)$. We refer to γ_j's as inclusion probabilities. See *Spike-and-slab priors* in *Methods and formulas*.

With the global–local shrinkage priors, normal priors are assumed for regression coefficients, and λ_j's are used to define the prior variances of coefficients,

$$\beta_j|\lambda_j, \tau^2 \sim N\left(0, \frac{\lambda_j \tau^2}{1-\lambda_j}\right)$$

where (random) hyperparameter τ controls global shrinkage and random λ_j's control local shrinkage. λ_j's cannot be interpreted as probabilities similarly to γ_j's in spike-and-slab priors, but each λ_j still controls the degree of inclusion of x_j in the following sense. For values of λ_j close to zero, the prior variance of β_j is shrunk to zero, and x_j is "excluded" or, more precisely, provides less contribution to the regression. For values of λ_j close to one, the prior variance of β_j gets closer to infinity so that the coefficient is unconstrained and x_j is "included" or rather provides more contribution to the regression model. λ_j's are used to define what we call inclusion coefficients; see *Global–local shrinkage priors* in *Methods and formulas*.

Interpretation of coefficient estimates is an important aspect of variable selection. Ideally, we want inferential methods that recover the data-generating model consistently. In classical approaches, such as penalized least squares, the estimates are predicated on the selected predictors to be included in the model. Such approaches do not account for the selection uncertainty. In model averaging approaches, such as BMA, the estimates are aggregated over many models, which can make interpretation difficult. In Bayesian variable selection, the two steps, variable selection and coefficient estimation, go hand in hand and are performed simultaneously, which inherently accounts for selection uncertainty during estimation. If, for example, the posterior mean estimate $\widehat{\gamma}_j$ of the inclusion indicator γ_j is close to zero, we can expect the corresponding coefficient estimate $\widehat{\beta}_j$ to be close to zero as well. The Bayesian model accounts for both possibilities, inclusion and exclusion of x_j as a predictor, and this is reflected in the posterior coefficient estimate $\widehat{\beta}_j$. We should not, however, judge the importance of x_j based on how close $\widehat{\beta}_j$ is to zero. We should use estimates $\widehat{\gamma}_j$'s (or $\widehat{\lambda}_j$'s with global–local shrinkage priors) to interpret the importance of predictors and estimates $\widehat{\beta}_j$'s to describe the effect sizes associated with predictors. Under certain conditions, $\widehat{\beta}_j$'s are consistent estimates of the true effect sizes, and the data-generating model can be recovered assuming all true predictors are included in the model. See *Methods and formulas* for details.

Introductory examples

In the following series of examples, we will demonstrate how to use the `bayesselect` command and interpret its output. We consider the simulated dataset `bmaintro` from *Motivating examples* in [BMA] **Intro**.

```
. use https://www.stata-press.com/data/r19/bmaintro
(Simulated data for BMA example)
```

There are 10 potential predictors, `x1` through `x10`, for the response variable `y`. By design, only `x2` and `x10` are true predictors, and the rest of the variables are unrelated to `y`.

We will model `y` using `x1` through `x10` as predictors and apply four different priors for regression coefficients. We will then compare the models.

▷ Example 1: Variable selection using the default horseshoe global–local shrinkage prior

We start by using the default prior for regression coefficients in `bayesselect`. It is the horseshoe prior with the scale of 1, which also corresponds to the `hshoe` option. To specify a different scale value, we can use the `hshoe(#)` option. This prior is one of the global–local shrinkage priors.

The syntax of `bayesselect` is similar to that of any other regression command in Stata, a dependent variable, `y`, followed by a list of predictors, `x1-x10` in this case. The only option we add is a random-number seed for reproducibility.

```
. bayesselect y x1-x10, rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,{sigma2})

Priors:
  {y:x1 ... x10} ~ glshrinkage(1,{tau},{lambdas})                          (1)
       {y:_cons} ~ normal(0,10000)                                         (1)
        {sigma2} ~ jeffreys

Hyperprior:
  {tau lambdas} ~ halfcauchy(0,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.

Bayesian variable selection                       MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling            Burn-in          =      2,500
                                                  MCMC sample size =     10,000
Global-local shrinkage coefficient prior:         Number of obs    =        200
  Horseshoe(1)                                    Acceptance rate  =      .8628
                                                  Efficiency:  min =      .1384
                                                               avg =      .6807
Log marginal-likelihood = -296.17324                           max =          1

------------------------------------------------------------------------------
             |                                     Equal-tailed      Inclusion
           y |      Mean   Std. dev.     MCSE   [95% cred. interval]      coef.
-------------+----------------------------------------------------------------
         x10 |  5.118244   .0870914   .0008709   4.950923    5.29129       1.00
          x2 |   1.18836   .0717654   .0007421   1.048757   1.328171       0.95
          x3 |  -.119698   .0842116   .0022636  -.2889022   .0135837       0.48
          x9 |  .0456459   .0657175   .0013671  -.0584361   .1970286       0.34
          x1 |  .0351392   .0595862   .0010334  -.0620478   .1757773       0.31
          x4 |  -.022399   .0557828   .0007517  -.1531457    .080328       0.30
          x5 |  .0124905   .0539176   .0006082  -.0931158   .1348377       0.29
          x7 |  .0016312   .0543838   .0005438  -.1126321   .1209322       0.29
          x8 | -.0113579   .0546242     .00059  -.1352596   .0968524       0.28
          x6 | -.0053055    .050503    .000511  -.1189606   .0979294       0.28
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. dev.     MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
y            |
       _cons |   .603351   .0788468    .000809   .6033972   .4488462   .7566242
-------------+----------------------------------------------------------------
      sigma2 |   1.16503   .1206306    .002689   1.160276   .9471227    1.41593
         tau |  .1923435   .1571121    .008269   .1476418   .0305629   .6212223
------------------------------------------------------------------------------
```

The output of bayesselect includes a model summary, a header, and two estimation tables. The first one is a table of regression coefficient summaries. The second one is a standard Markov chain Monte Carlo (MCMC) summary table for additional model parameters such as the constant term, {y:_cons}, error variance {sigma2}, and hyperparameters, {tau} in this case.

The regression coefficient table is similar to the standard MCMC table (see [BAYES] **bayesmh**), but instead of a column for the estimated medians, it includes a column for the estimated inclusion coefficients. The inclusion coefficients are measures of predictor importance. By default, only predictors with inclusion coefficients of 0.1 or above are reported, which is all predictors in our example. Only two predictors, x10 and x2, have inclusion coefficients above 0.5. These are the true predictors of y by design. The actual coefficient values for x10 and x2 used to simulate the data were 5 and 1.2, and the error variance was 1. The estimated posterior means for the coefficients, 5.12 and 1.19, and the error variance, 1.17, are very close to the true values.

The coefficient estimates for all predictors with inclusion coefficients less than 0.5, except x3, are close to 0. Moreover, their respective credible intervals, including those for x3, contain zero. In this simulation example, there is a clear distinction between important and unimportant predictors, which, of course, may not be the case with real datasets. You should not be concerned because bayesselect does not exclude any of the potential predictors from the regression model but simply controls their effect according to their relevance in predicting the outcome.

As we mentioned in the introduction, bayesselect regulates the effects of predictors by specifying a prior for regression coefficients that shrinks them toward zero based on how well the predictors explain the outcome. The regression coefficients of weak predictors are shrunk more toward zero. The default prior for coefficients is a horseshoe prior with the scale of 1, as we can see in the header. From the model summary output, a horseshoe prior is a global–local shrinkage prior with hyperparameter {tau} (global shrinkage) and latent parameters {lambdas:} (local shrinkage), one for each coefficient, all distributed as half-Cauchy with location of 0 and scale of 1. A global–local shrinkage prior assumes a normal prior for each regression coefficient with mean 0 and standard deviation controlled by {tau} and the corresponding parameter in {lambdas:}. The smaller these parameters, the closer the coefficient is to zero. See *Global–local shrinkage priors* in *Methods and formulas* for details.

Although the {lambdas:} parameters are not shown by bayesselect, they can be summarized by using the bayesstats summary command (see [BAYES] **bayesstats summary**).

```
. bayesstats summary {lambdas:}
```

Posterior summary statistics MCMC sample size = 10,000

lambdas	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
x1	.9367181	1.866672	.031428	.5139979	.0308739	4.345145
x2	13.94609	27.81457	.688323	7.523173	1.452113	66.45129
x3	1.702801	4.462234	.059901	.9696198	.0564161	7.329253
x4	.9358786	2.41981	.037191	.4985472	.0132313	4.053352
x5	.8772942	2.198556	.034558	.4730888	.0149907	3.95497
x6	.8135167	1.794375	.034068	.4435154	.0135487	3.642703
x7	.8537399	1.768146	.032387	.4734345	.020679	3.825944
x8	.8606228	1.840859	.033986	.4585138	.0163238	4.008136
x9	1.009114	1.758922	.033081	.5741607	.024573	4.654922
x10	59.46404	118.0737	3.21056	31.47493	5.482909	285.2516

All {lambdas:} parameters are positive, and the magnitudes of those corresponding to the important predictors x2 and x10 are much larger than the rest. The difference between magnitudes is a relative measure; this is why the inclusion coefficients, with values between 0 and 1, are introduced as a more convenient measure of predictor importance than the posterior mean estimates of {lambdas:}.

The inclusion coefficients reported by bayesselect in the last column of the coefficient table are the posterior mean estimates of {lambdas:} after the latter are transformed to take values in the [0,1] range. Specifically, from *Methods and formulas*, an inclusion coefficient for a predictor x_j is defined as $\gamma_j = 1 - \kappa_j = 1 - 1/(1 + \lambda_j^2/\lambda_0^2)$, where κ_j is known as a shrinkage coefficient and λ_0 is a scale parameter specified with a global–local shrinkage prior. In our example, the scale of the horseshoe prior is one, $\lambda_0 = 1$. For instance, we can estimate the inclusion coefficient for x2, γ_2, reported to be 0.95 by bayesselect, as follows:

```
. bayesstats summary (gamma2: (1-1/(1+{lambdas:x2}^2)))
Posterior summary statistics                      MCMC sample size =     10,000
      gamma2 : 1-1/(1+{lambdas:x2}^2)
```

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
gamma2	.9477335	.0891366	.003519	.9826383	.6783145	.9997736

In this example, we used 0.5 as an inclusion cutoff to determine which predictors are important. This may be justified because the mean of the default prior distribution used for the local shrinkage coefficients κ_j's, and consequently γ_j's, is 0.5. Specifically, the default HalfCauchy$(0, 1)$ prior for λ_j's leads to the default Beta$(0.5, 0.5)$ prior for κ_j's, which has a mean of 0.5. In general, if we change the default prior, we may consider a different inclusion cutoff value.

◁

▷ Example 2: Bayesian lasso global–local shrinkage prior

Bayesian lasso (Park and Casella 2008) is a Bayesian analog of the l_1-penalized least-squares approach to variable selection. It uses a global–local shrinkage prior for regression coefficients that assumes a Rayleigh distribution for local shrinkage latent parameters λ_j's instead of a half-Cauchy distribution as in example 1. This is also equivalent to using Laplace priors as marginal priors for regression coefficients β_j's.

To request a Bayesian lasso with a scale of 1, we use the blasso option. The blasso(#) option allows us to specify any other positive scale value.

We refit our model from example 1 using Bayesian lasso.

```
. bayesselect y x1-x10, blasso rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,{sigma2})
Priors:
  {y:x1 ... x10} ~ glshrinkage(1,{tau},{lambdas})                          (1)
       {y:_cons} ~ normal(0,10000)                                         (1)
        {sigma2} ~ jeffreys
Hyperpriors:
      {tau} ~ halfcauchy(0,1)
  {lambdas} ~ rayleigh(1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.
Bayesian variable selection                      MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling           Burn-in          =      2,500
                                                 MCMC sample size =     10,000
Global-local shrinkage coefficient prior:        Number of obs    =        200
  Bayesian lasso(1)                              Acceptance rate  =      .8597
                                                 Efficiency:  min =      .8911
                                                              avg =      .9731
Log marginal-likelihood = -333.53826                          max =          1
```

y	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion coef.
x10	5.120569	.0875861	.0008759	4.950711	5.294459	0.87
x2	1.182651	.0719754	.0007198	1.039568	1.323594	0.65
x3	-.1771405	.0797991	.0008454	-.3355561	-.0213421	0.41
x9	.0891755	.0795337	.0008133	-.0649558	.2444695	0.39
x5	.0327607	.0761729	.0007617	-.1131709	.1846671	0.38
x4	-.041633	.0765783	.0007789	-.1936397	.1045709	0.38
x1	.0689381	.0753258	.0007865	-.0752699	.2188716	0.38
x8	-.0323204	.0770683	.0007707	-.184323	.1217865	0.37
x6	-.0132317	.0749707	.0007497	-.1599103	.1358485	0.37
x7	.0081383	.0804661	.0008047	-.1498234	.1664523	0.37

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
y						
_cons	.6178375	.0801636	.000812	.6184826	.4568188	.7739675
sigma2	1.176275	.120801	.002555	1.171413	.9596697	1.436654
tau	.7903534	.2686312	.005125	.7395585	.4237649	1.447437

The posterior summary results are very similar to those using the horseshoe prior. Because a different prior is assumed for local shrinkage parameters `{lambdas:}`, the estimates for the global shrinkage `{tau}` are different.

The inclusion coefficients are between 0.37 and 0.87 and are less spread out than those for the horseshoe prior. And the inclusion coefficients for `x10` and `x2`, 0.87 and 0.65, are somewhat smaller than those for the horseshoe prior. The Bayesian lasso thus tends to apply less shrinkage to the coefficients, resulting in less distinction between important and unimportant predictors. For example, the posterior mean estimate for the x3 coefficient is −0.18, and the 95% credible interval does not include 0, in contrast to the estimates for the horseshoe prior.

For comparison, let's also inspect the {lambdas:} parameters.

```
. bayesstats summary {lambdas:}
Posterior summary statistics                      MCMC sample size =    10,000
```

lambdas	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
x1	.8657674	.586799	.006028	.7458889	.086964	2.244439
x2	1.531099	.5632616	.006642	1.45914	.630293	2.831665
x3	.9452156	.5800462	.005875	.8353631	.1587102	2.356409
x4	.8677547	.5924273	.005924	.7515854	.088186	2.271446
x5	.8766478	.6006823	.006007	.7546303	.0899792	2.31136
x6	.8613335	.5913605	.005992	.7469369	.0868421	2.275777
x7	.8540772	.5932745	.005933	.7341128	.0844432	2.282867
x8	.8639806	.5946548	.006198	.741773	.0877804	2.308843
x9	.8930035	.586916	.005793	.7759418	.1063435	2.287071
x10	2.786343	.6363102	.010189	2.749506	1.645258	4.121829

The posterior mean estimates are between 0.85 and 2.79. The differences between magnitudes of {lambdas:x2} and {lambdas:x10} and the less important predictors are much smaller than with the horseshoe prior, which confirms the smaller shrinkage effect of Bayesian lasso. From the point of view of classical model selection, we can say that Bayesian lasso prefers more complex models than the horseshoe prior.

◁

▷ Example 3: Normal spike-and-slab prior

In the next two examples, we demonstrate the other important class of priors for variable selection, the spike-and-slab priors. We first show a normal spike-and-slab prior. The regression coefficient priors in this case are mixtures of two normal distributions.

We fit the same regression model as in the previous examples, but now we use the `ssnormal` option, which specifies a normal spike-and-slab prior with the default values of 0.01 and 1 for the two standard deviation parameters. We can specify different values for standard deviations by using the `ssnormal(`*#1* *#2*`)` option.

```
. bayesselect y x1-x10, ssnormal rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,{sigma2})
Priors:
  {y:x1 ... x10} ~ mixnormal0(1,.01,1,{gammas})                            (1)
       {y:_cons} ~ normal(0,10000)                                          (1)
        {sigma2} ~ jeffreys
Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(1,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.
Bayesian variable selection                      MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling           Burn-in          =      2,500
                                                 MCMC sample size =     10,000
Spike-and-slab coefficient prior:                Number of obs    =        200
  Normal mixture: N(0,.01) and N(0,1)            Acceptance rate  =      .8638
  Beta(1,1) for {theta}                          Efficiency:  min =     .02048
                                                              avg =      .5557
Log marginal-likelihood = -313.24428                          max =          1
```

y	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion prob.
x2	1.184036	.0715031	.000715	1.044366	1.324463	1.00
x10	5.100833	.0883483	.0008835	4.928953	5.27378	1.00
x3	-.0798283	.1059473	.0074037	-.3104386	.0203455	0.44
x8	.0038787	.0393615	.0005284	-.1223508	.0550395	0.18
x7	.0098883	.0309695	.0003097	-.0516427	.0802481	0.12
x9	.0140702	.0430647	.0012918	-.0194029	.1649108	0.12
x1	.002177	.0365315	.0008101	-.0292478	.1265267	0.11

Note: 3 coefficients with inclusion values less than .1 not shown.

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
y						
_cons	.6209303	.0791626	.000792	.6216375	.4674763	.7745341
sigma2	1.171751	.1201083	.002683	1.161649	.9620094	1.429011
theta	.3491553	.1607552	.004354	.3323494	.0880766	.6986263

Compared with the global–local shrinkage priors from the previous two examples, the estimated coefficients of unimportant predictors are closer to zero with this normal spike-and-slab prior. Three regression coefficients are not reported because their inclusion values are below the default cutoff of 0.1.

The spike-and-slab priors introduce latent parameters `{gammas:}`. These are random binary indicators for the mixture distributions; see *Spike-and-slab priors* in *Methods and formulas* for details. From the model summary output, `{gammas:}` are distributed as Bernoulli with hyperparameter (success probability) `{theta}`. And `{theta}` is assumed to have a beta distribution with shape parameters of 1s, which is equivalent to a uniform distribution on [0,1]. We can specify other shape values by using the `betaprior(`*#1 #2*`)` option.

The inclusion values reported in the table are the posterior means of `{gammas:}` and thus can be interpreted as mixing probabilities between the spike and slab portions of the coefficient priors. In our case, the posterior mean estimates for `{gammas:x2}` and `{gammas:x10}` are perfect ones and so are their inclusion probabilities. This means that for `x2` and `x10` the model always chooses the slab, flat, portion of the priors.

`{theta}` is the probability parameter of the Bernoulli hyperpriors for `{gammas:}`. Its posterior mean estimate, 0.35 in our case, can be interpreted as an indication of the overall sparsity of the model and can be used for comparing one spike-and-slab model with another.

In the default output, several predictors are not reported because their inclusion probabilities are below 0.1. We can use the `allcoef` option to see the summary for all coefficients. To avoid repetition, we also suppress the model summary and the header.

```
. bayesselect, allcoef nomodelsummary noheader
```

y	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion prob.
x2	1.184036	.0715031	.000715	1.044366	1.324463	1.00
x10	5.100833	.0883483	.0008835	4.928953	5.27378	1.00
x3	-.0798283	.1059473	.0074037	-.3104386	.0203455	0.44
x8	.0038787	.0393615	.0005284	-.1223508	.0550395	0.18
x7	.0098883	.0309695	.0003097	-.0516427	.0802481	0.12
x9	.0140702	.0430647	.0012918	-.0194029	.1649108	0.12
x1	.002177	.0365315	.0008101	-.0292478	.1265267	0.11
x5	.0071316	.0263058	.0003278	-.0224294	.0780056	0.08
x6	-.0008068	.0235381	.0002354	-.0421222	.0292265	0.07
x4	-.00223	.0252274	.0003786	-.0614174	.0240777	0.06

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
y						
_cons	.6209303	.0791626	.000792	.6216375	.4674763	.7745341
sigma2	1.171751	.1201083	.002683	1.161649	.9620094	1.429011
theta	.3491553	.1607552	.004354	.3323494	.0880766	.6986263

After `x10` and `x2`, the predictor with the next highest inclusion probability of 0.44 is `x3`.

Similarly to {lambdas:} of global–local shrinkage priors, {gammas:} are not reported by bayesselect, but we can use bayesstats summary to inspect these mixing probability parameters.

```
. bayesstats summary {gammas:}
Posterior summary statistics                      MCMC sample size =    10,000
```

gammas	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
x1	.1115	.314766	.008982	0	0	1
x2	1	0	0	1	1	1
x3	.4366	.495989	.040969	0	0	1
x4	.0634	.2436932	.00592	0	0	1
x5	.0832	.2761981	.006551	0	0	1
x6	.0702	.2554966	.006326	0	0	1
x7	.1247	.330395	.007687	0	0	1
x8	.1752	.3801571	.008953	0	0	1
x9	.1167	.3210785	.011932	0	0	1
x10	1	0	0	1	1	1

Because {gammas:} are binary indicators, the medians and the endpoints of credible intervals are always 0 or 1. The medians indicate which of the two values dominate in the MCMC sample. Given perfect inclusion of x2 and x10, {gammas:x2} and {gammas:x10} have a constant value of one in the entire MCMC sample. This gives us high confidence in the importance of predictors x2 and x10.

◁

▷ Example 4: Laplace spike-and-slab prior

The second type of a spike-and-slab prior uses a mixture of Laplace distributions. That is, the spike and slab portions of the coefficient priors are Laplace distributions instead of normal distributions as in the previous example.

We request this prior by using the `sslaplace` option. The `sslaplace` prior uses the default values of 0.01 and 1 for the two scale parameters, but we can specify different values by using the `sslaplace(`*#1* *#2*`)` option.

```
. bayesselect y x1-x10, sslaplace rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,{sigma2})
Priors:
  {y:x1 ... x10} ~ mixlaplace(1,.01,1,{gammas})                           (1)
       {y:_cons} ~ normal(0,10000)                                        (1)
        {sigma2} ~ jeffreys
Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(1,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.
Bayesian variable selection                      MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling           Burn-in          =      2,500
                                                 MCMC sample size =     10,000
Spike-and-slab coefficient prior:                Number of obs    =        200
  Laplace mixture: L(0,.01) and L(0,1)           Acceptance rate  =      .8635
  Beta(1,1) for {theta}                          Efficiency:  min =     .04937
                                                              avg =      .6597
Log marginal-likelihood = -294.02003                          max =      .9705

------------------------------------------------------------------------------
             |                                     Equal-tailed     Inclusion
           y |      Mean   Std. dev.     MCSE   [95% cred. interval]    prob.
-------------+----------------------------------------------------------------
          x2 |  1.185791   .0715964   .000731   1.045868   1.324387      1.00
         x10 |  5.122913   .0860102  .0008731   4.951631   5.291972      1.00
          x3 | -.0595752    .091769    .00413  -.2895237   .0187028      0.31
------------------------------------------------------------------------------
Note: 7 coefficients with inclusion values less than .1 not shown.

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. dev.     MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
y            |
       _cons |  .6148945   .0800458     .0008   .6153598   .4574479   .7699493
-------------+----------------------------------------------------------------
      sigma2 |  1.166491   .1200866   .002618   1.158892   .9575422    1.42881
       theta |  .3087888   .1438559   .002711   .2943327   .0776807   .6266575
------------------------------------------------------------------------------
```

The coefficient estimates of the important predictors are similar to those of the normal-mixture prior model from example 3. But now 7 (compared with 3 before) predictors have inclusion probabilities below 0.1. And the posterior mean estimate for `{theta}`, 0.31, is lower, which suggests that the Laplace-mixture model is sparser. Indeed, if we inspect all inclusion probabilities (see below), we will see that all, except the top 3, are between 0.05 and 0.07, whereas those for the normal-mixture prior are between 0.06 and 0.18.

```
. bayesselect, allcoef nomodelsummary noheader
```

y	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion prob.
x2	1.185791	.0715964	.000731	1.045868	1.324387	1.00
x10	5.122913	.0860102	.0008731	4.951631	5.291972	1.00
x3	-.0595752	.091769	.00413	-.2895237	.0187028	0.31
x9	.0096531	.0341256	.0006481	-.0250025	.120656	0.07
x1	.004208	.0274389	.0004817	-.0302802	.0810768	0.06
x8	.000295	.0245224	.0002962	-.0489182	.0391258	0.05
x7	.0020103	.0226965	.000227	-.036048	.0436753	0.05
x4	-.0029021	.0239831	.0003345	-.0534428	.0300279	0.05
x5	.0033791	.0227582	.0002686	-.0284755	.0500044	0.05
x6	-.0008044	.0206703	.0002005	-.0374593	.0346636	0.05

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
y						
_cons	.6148945	.0800458	.0008	.6153598	.4574479	.7699493
sigma2	1.166491	.1200866	.002618	1.158892	.9575422	1.42881
theta	.3087888	.1438559	.002711	.2943327	.0776807	.6266575

The fact that we obtain very similar results with different priors from examples 1, 2, and 3 and from this example suggests that our results are not sensitive to the choice of priors and we can be confident in our conclusions about the importance of predictors x2 and x10.

◁

▷ Example 5: Sparsity control

In spike-and-slab models, we can control model sparsity through the prior of the hyperparameter {theta}. The default prior for {theta} is $\text{Beta}(1,1)$, which is equivalent to the uniform distribution on [0,1]. That is, by default, we have no preference for the degree of sparsity of the regression model. By providing an informative prior for {theta}, we can make models sparser or denser.

For example, by specifying a $\text{Beta}(1,9)$ prior for {theta}, we favor sparser models. The mean of $\text{Beta}(1,9)$ is 0.1 and so is the prior mean of {theta}. In other words, a priori, we expect only one important predictor of y out of the potential 10. In the process of Bayesian variable selection, this expectation is weighted by the evidence from the data to provide its posterior estimate.

Continuing with the Laplace model from example 4, let's use this beta prior for theta. We specify the allcoef option to see all regression coefficients.

```
. bayesselect y x1-x10, sslaplace betaprior(1 9) allcoef rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,{sigma2})
Priors:
  {y:x1 ... x10} ~ mixlaplace(1,.01,1,{gammas})                           (1)
       {y:_cons} ~ normal(0,10000)                                        (1)
        {sigma2} ~ jeffreys
Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(1,9)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.
Bayesian variable selection                      MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling           Burn-in          =      2,500
                                                 MCMC sample size =     10,000
Spike-and-slab coefficient prior:                Number of obs    =        200
  Laplace mixture: L(0,.01) and L(0,1)           Acceptance rate  =      .8649
  Beta(1,9) for {theta}                          Efficiency:  min =     .04154
                                                              avg =      .6557
Log marginal-likelihood = -322.15504                          max =          1
```

y	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion prob.
x2	1.185836	.0723118	.0007478	1.043249	1.326291	1.00
x10	5.123372	.0877025	.000877	4.952162	5.298158	1.00
x3	-.0431981	.0814627	.003997	-.2766626	.0189029	0.22
x9	.0073567	.0292032	.0005792	-.0246281	.0894012	0.05
x1	.0026981	.0231029	.0003779	-.0283663	.0443341	0.03
x7	.0021759	.0184902	.0001913	-.0288422	.0379245	0.02
x5	.0028945	.0178557	.0001985	-.0262179	.0387407	0.02
x8	.001304	.0186369	.0002192	-.0293738	.0339263	0.02
x6	-.0011907	.0171051	.0001862	-.0334828	.0286884	0.02
x4	-.0014873	.0180464	.0002251	-.0350797	.0278281	0.02

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
y						
_cons	.6151788	.0789083	.00078	.6159789	.4614939	.7671695
sigma2	1.169404	.1199563	.002649	1.161906	.9594109	1.428383
theta	.1704879	.088016	.001268	.1572679	.0376195	.3730968

The resulting posterior mean estimate for {theta} is now 0.17, down from 0.31 for the Laplace spike-and-slab model with the default beta prior. x10 and x2 remain to be the two important predictors, but the rest of the predictors (ignoring x3) now have lower inclusion probabilities, all between 0.02 and 0.05. The separation between important and unimportant predictors is more prominent.

Let's see what happens when we use a denser model. A Beta$(9, 1)$ prior for {theta} sets the prior mean to 0.9, which means we expect to have 9 important predictors in the model.

```
. bayesselect y x1-x10, sslaplace betaprior(9 1) allcoef rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,{sigma2})
Priors:
  {y:x1 ... x10} ~ mixlaplace(1,.01,1,{gammas})                            (1)
       {y:_cons} ~ normal(0,10000)                                         (1)
        {sigma2} ~ jeffreys
Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(9,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.
Bayesian variable selection                       MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling            Burn-in          =      2,500
                                                  MCMC sample size =     10,000
Spike-and-slab coefficient prior:                 Number of obs    =        200
  Laplace mixture: L(0,.01) and L(0,1)            Acceptance rate  =      .8647
  Beta(9,1) for {theta}                           Efficiency:  min =     .09248
                                                               avg =      .6329
Log marginal-likelihood = -316.37911                           max =          1
```

y	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion prob.
x2	1.18411	.0718673	.0007187	1.042284	1.326711	1.00
x10	5.123829	.0874442	.0008744	4.951699	5.295111	1.00
x3	-.1261931	.104562	.0034383	-.3224889	.0132646	0.69
x9	.0296768	.0610811	.001271	-.0312488	.2024558	0.30
x1	.0176132	.0493214	.0009162	-.0364677	.166006	0.24
x8	-.0051904	.041614	.0005622	-.1283518	.0606106	0.20
x5	.0082778	.0398273	.0004957	-.0596965	.124567	0.20
x4	-.0090804	.0403741	.0005142	-.1345912	.0475379	0.20
x7	.0032438	.0378493	.0003976	-.082511	.101217	0.20
x6	-.0024044	.0344595	.0003446	-.0901232	.068009	0.18

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
y						
_cons	.6153574	.0787261	.000787	.6161907	.4579815	.7685391
sigma2	1.164687	.1207741	.002705	1.157213	.9461043	1.424447
theta	.6610358	.124062	.002632	.6642454	.411753	.8930941

The posterior mean of {theta} is now estimated to be 0.66, much higher than 0.31 from the model with the default beta prior. Moreover, the inclusion probability for x3 increases to 0.69. Inclusion probabilities for all other predictors also increase. If we apply the 0.5 threshold of importance, we now have

3 important predictors in the model, x10, x2, and x3. However, as we commented in example 1, with a prior mean of 0.9 for {theta}, we may consider a higher inclusion cutoff value than 0.5 to determine importance of predictors.

The model with the default beta prior provides a better fit than both models with informative priors for {theta}, in terms of the log-marginal likelihood, −294 versus −322 and −316. Specifying strong sparsity information a priori thus should be carefully justified.

◁

Diabetes progression study

In the following examples, we use the diabetes dataset from Efron et al. (2004). The dataset is from a study on disease progression of 442 diabetes patients. At the beginning of the study, age, sex, body mass index (bmi), and blood pressure (bp) are collected for each patient, along with six measurements of their blood serum (serum1 through serum6). The response variable diabetes quantifies disease progression one year after the baseline variables are obtained.

Here is a short description of the dataset.

```
. use https://www.stata-press.com/data/r19/diabetes
(2004 Diabetes progression data)

. describe

Contains data from https://www.stata-press.com/data/r19/diabetes.dta
 Observations:           442                  2004 Diabetes progression data
    Variables:            11                  14 Aug 2024 11:39
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
diabetes        float   %9.0g                 Progression of diabetes after one
                                                year (std.)
age             float   %9.0g                 Age (std.)
sex             float   %9.0g                 Sex (std.)
bmi             float   %9.0g                 Body mass index (std.)
bp              float   %9.0g                 Blood pressure (std.)
serum1          float   %9.0g                 Blood serum measurement 1 (std.)
serum2          float   %9.0g                 Blood serum measurement 2 (std.)
serum3          float   %9.0g                 Blood serum measurement 3 (std.)
serum4          float   %9.0g                 Blood serum measurement 4 (std.)
serum5          float   %9.0g                 Blood serum measurement 5 (std.)
serum6          float   %9.0g                 Blood serum measurement 6 (std.)
-------------------------------------------------------------------------------
Sorted by:
```

The variables in the original dataset were standardized to have sample means of zero and sample standard deviations of one. This ensures optimal performance for all variable-selection models in bayesselect.

To compare the predictive performance of different variable-selection models later, we split the sample into subsamples for training and testing.

```
. splitsample, generate(sample) split(1 1) rseed(19)
```

The newly generated variable sample records the subsample.

▷ Example 6: Performing variable selection for the diabetes study

We fit the default variable-selection model of `bayesselect`. It uses a horseshoe global–local shrinkage prior with the scale of one for regression coefficients. We use the training subsample to fit the model and specify a random-number seed for reproducibility. And we will use the testing subsample to compute predictions for later comparison of model performances.

```
. bayesselect diabetes age sex bmi bp serum1-serum6 if sample == 1, rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  diabetes ~ normal(xb_diabetes,{sigma2})
Priors:
  {diabetes:age ... serum6} ~ glshrinkage(1,{tau},{lambdas})              (1)
           {diabetes:_cons} ~ normal(0,10000)                             (1)
                   {sigma2} ~ jeffreys
Hyperprior:
  {tau lambdas} ~ halfcauchy(0,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_diabetes.
Bayesian variable selection                       MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling            Burn-in          =      2,500
                                                  MCMC sample size =     10,000
Global-local shrinkage coefficient prior:         Number of obs    =        221
  Horseshoe(1)                                    Acceptance rate  =      .8587
                                                  Efficiency:  min =      .2055
                                                               avg =      .3858
Log marginal-likelihood = -228.01981                           max =      .8596
```

diabetes	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion coef.
bmi	.3251239	.0605027	.0007609	.2074427	.4405296	0.74
serum5	.3190135	.0774733	.0012965	.1741643	.480524	0.73
bp	.1820939	.0583262	.0009469	.0646499	.2973787	0.59
serum3	-.1483656	.0902192	.001974	-.3278771	.0116305	0.53
serum1	-.0673476	.1158495	.0025556	-.3630464	.1077651	0.38
sex	-.06953	.0536515	.0011804	-.1792851	.0170811	0.37
serum2	-.0025097	.0930945	.0016917	-.171356	.2276703	0.31
serum4	-.0045453	.0771996	.0011999	-.1771875	.1556561	0.31
age	-.0331836	.0446677	.0008272	-.132988	.0410595	0.28
serum6	-.0098386	.0401496	.0004331	-.0970883	.0706108	0.25

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
diabetes						
_cons	-.008172	.0461672	.000469	-.0081996	-.0985132	.0822581
sigma2	.4639809	.0454678	.001003	.4613123	.3833852	.5611066
tau	.1984424	.1206534	.00484	.1679971	.0532921	.5104429

Four predictors have inclusion coefficients greater than 0.5: bmi, serum5, bp, and serum3. This is in agreement with lasso regression results from Efron et al. (2004), who report these same predictors in the same order of importance to be the top predictors of diabetes.

To generate predictions, we save the MCMC simulation sample. We also store the estimation results as model1.

```
. bayesselect, saving(model1sim)
note: file model1sim.dta saved.
. estimates store model1
```

We compute the predictive posterior means for the testing subsample using bayespredict. We store the predictions in the new pmean1 variable. Using the predicted means, we compute the squared prediction error over the testing subsample and save it in the sqerr1 variable. We then drop the pmean1 variable.

```
. bayespredict double pmean1 if sample == 2, mean
Computing predictions ...
. generate double sqerr1 = (diabetes-pmean1)^2
(221 missing values generated)
. drop pmean1
```

We fit a Bayesian lasso model and store its estimation results in `model2`. This is the other global–local shrinkage model available in `bayesselect`. We also specify the cutoff inclusion value of 0.5 to focus on our top predictors of interest.

```
. bayesselect diabetes age sex bmi bp serum1-serum6 if sample == 1, blasso
> cutoff(0.5) rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  diabetes ~ normal(xb_diabetes,{sigma2})
Priors:
  {diabetes:age ... serum6} ~ glshrinkage(1,{tau},{lambdas})              (1)
           {diabetes:_cons} ~ normal(0,10000)                              (1)
                   {sigma2} ~ jeffreys
Hyperpriors:
      {tau} ~ halfcauchy(0,1)
  {lambdas} ~ rayleigh(1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_diabetes.
Bayesian variable selection                       MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling            Burn-in          =      2,500
                                                  MCMC sample size =     10,000
Global-local shrinkage coefficient prior:         Number of obs    =        221
  Bayesian lasso(1)                               Acceptance rate  =      .8588
                                                  Efficiency:  min =      .6102
                                                               avg =      .7203
Log marginal-likelihood = -240.89592                           max =      .8076
```

diabetes	Mean	Std. dev.	MCSE	Equal-tailed [95% cred.	interval]	Inclusion coef.
bmi	.3168865	.0591047	.0006577	.200824	.4324047	0.68
serum5	.3153706	.079797	.0009319	.163874	.4748501	0.68
bp	.194158	.0557217	.0006521	.0846107	.3028491	0.60
serum3	-.1598567	.0932792	.0011941	-.3477528	.0145296	0.56

Note: 6 coefficients with inclusion values less than .5 not shown.

	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred.	interval]
diabetes						
_cons	-.0077397	.0465436	.000471	-.0073796	-.0997193	.0832937
sigma2	.4639159	.0445557	.000939	.4620219	.3849446	.5569725
tau	.1773416	.0743409	.001842	.1610344	.0831944	.3620409

```
. bayesselect, saving(model2sim)
note: file model2sim.dta saved.
. estimates store model2
```

Again, the top four most important predictors are `bmi`, `serum5`, `bp`, and `serum3`. Overall, the estimates of regression coefficients and other model parameters are very close to those of the default horseshoe model. Although the inclusion coefficient for `serum3` is 0.56, its 95% credible interval includes 0. This is another indicator of lesser importance of `serum3` in comparison with the top three predictors.

We use the testing subsample to compute and store in the `sqerr2` variable the squared prediction error for the fitted Bayesian lasso model.

```
. bayespredict double pmean2 if sample == 2, mean
Computing predictions ...
. generate double sqerr2 = (diabetes-pmean2)^2
(221 missing values generated)
. drop pmean2
```

We fit a Laplace spike-and-slab model and store its estimation results in `model3`.

```
. bayesselect diabetes age sex bmi bp serum1-serum6 if sample == 1, sslaplace
> cutoff(0.5) rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  diabetes ~ normal(xb_diabetes,{sigma2})

Priors:
  {diabetes:age ... serum6} ~ mixlaplace(1,.01,1,{gammas})                 (1)
           {diabetes:_cons} ~ normal(0,10000)                              (1)
                   {sigma2} ~ jeffreys

Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(1,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_diabetes.

Bayesian variable selection                        MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling             Burn-in          =      2,500
                                                   MCMC sample size =     10,000
Spike-and-slab coefficient prior:                  Number of obs    =        221
  Laplace mixture: L(0,.01) and L(0,1)             Acceptance rate  =       .862
  Beta(1,1) for {theta}                            Efficiency:  min =      .3024
                                                                avg =      .6477
Log marginal-likelihood =  -231.1353                            max =          1

------------------------------------------------------------------------------
             |                                     Equal-tailed     Inclusion
    diabetes |      Mean   Std. dev.     MCSE   [95% cred. interval]     prob.
-------------+----------------------------------------------------------------
         bmi |  .3191587   .0625475   .0006255   .1975578   .4430021      1.00
      serum5 |  .3708366   .1214061   .0015167   .1270256   .6200841      0.99
          bp |   .204166   .0650311   .0011827    .049531    .324994      0.97
------------------------------------------------------------------------------
Note: 7 coefficients with inclusion values less than .5 not shown.

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. dev.     MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
diabetes     |
       _cons | -.0064705   .0491249    .000491  -.0060638  -.1036942   .0883607
-------------+----------------------------------------------------------------
      sigma2 |  .5160015   .0668118    .001711   .5077463   .4113021   .6680642
       theta |  .4484548   .1693087    .003635   .4377134   .1480547   .7985162
------------------------------------------------------------------------------

. bayesselect, saving(model3sim)
note: file model3sim.dta saved.
. estimates store model3
```

The inclusion probability of serum3 is lower than 0.5, so it is not listed in the regression coefficient table. On the other hand, the inclusion probabilities of bmi, serum5, and bp are very high, above 0.97. The estimate of the serum5 coefficient is also somewhat higher than those from the global–local shrinkage models. We observe a stronger separation between predictors than in the previous two models.

We again use the testing subsample to compute the squared prediction error for this model and store it in the sqerr3 variable.

```
. bayespredict double pmean3 if sample == 2, mean
Computing predictions ...
. generate double sqerr3 = (diabetes-pmean3)^2
(221 missing values generated)
. drop pmean3
```

We fit a normal spike-and-slab model and store its estimation results in `model4`.

```
. bayesselect diabetes age sex bmi bp serum1-serum6 if sample == 1, ssnormal
> cutoff(0.5) rseed(19)
Burn-in ...
Simulation ...
Model summary
------------------------------------------------------------------------------
Likelihood:
  diabetes ~ normal(xb_diabetes,{sigma2})
Priors:
  {diabetes:age ... serum6} ~ mixnormal0(1,.01,1,{gammas})                 (1)
           {diabetes:_cons} ~ normal(0,10000)                              (1)
                   {sigma2} ~ jeffreys
Hyperpriors:
  {gammas} ~ bernoulli({theta})
   {theta} ~ beta(1,1)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_diabetes.
Bayesian variable selection                      MCMC iterations  =     12,500
Metropolis-Hastings and Gibbs sampling           Burn-in          =      2,500
                                                 MCMC sample size =     10,000
Spike-and-slab coefficient prior:                Number of obs    =        221
  Normal mixture: N(0,.01) and N(0,1)            Acceptance rate  =      .8552
  Beta(1,1) for {theta}                          Efficiency:  min =     .01052
                                                              avg =      .3413
Log marginal-likelihood = -228.80453                          max =          1
------------------------------------------------------------------------------
             |                                      Equal-tailed     Inclusion
    diabetes |      Mean   Std. dev.     MCSE   [95% cred. interval]     prob.
-------------+----------------------------------------------------------------
         bmi |   .315811   .0642621   .0006426   .1883758   .4417061      1.00
          bp |  .1892194   .0872727    .008507   .0027252   .3294809      0.88
      serum5 |  .3448803   .1735588   .0150721   .0055886   .6334824      0.88
------------------------------------------------------------------------------
Note: 7 coefficients with inclusion values less than .5 not shown.
------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. dev.     MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
diabetes     |
       _cons | -.0047216   .0499829      .0005  -.0048385  -.1040443   .0929065
-------------+----------------------------------------------------------------
      sigma2 |  .5448409   .0914339    .004062   .5290122   .4161136   .7791847
       theta |  .3732414   .1600284    .006682   .3623792   .0998876   .7070753
------------------------------------------------------------------------------
. bayesselect, saving(model4sim)
note: file model4sim.dta saved.
. estimates store model4
```

The posterior estimates are similar to those of the Laplace model. The `bp` and `serum5` predictors have somewhat lower inclusion probabilities of 0.88. The posterior mean estimate of `{theta}` is also lower, 0.37 versus 0.45, which indicates that the normal model is slightly more sparse than the Laplace model.

We store the squared prediction error for this model in the sqerr4 variable.

```
. bayespredict double pmean4 if sample == 2, mean
Computing predictions ...
. generate double sqerr4 = (diabetes-pmean4)^2
(221 missing values generated)
. drop pmean4
```

The results from all four models are more or less consistent, which makes it difficult to choose between them. We need to use a more formal model-selection criterion to make a decision.

◁

▷ Example 7: Model comparison using goodness of fit

The standard statistic for assessing goodness of fit of Bayesian models is the marginal likelihood. We can use the bayestest model command (see [BAYES] **bayestest model**) to compare the goodness of fit of the previous four variable-selection models. The command uses estimated marginal likelihoods and prior model probabilities to compute and report posterior model probabilities. By default, all four models are assumed equally likely a priori.

```
. bayestest model model1 model2 model3 model4
Bayesian model tests
```

	log(ML)	P(M)	P(M|y)
model1	-228.0198	0.2500	0.6664
model2	-240.8959	0.2500	0.0000
model3	-231.1353	0.2500	0.0296
model4	-228.8045	0.2500	0.3040

```
Note: Marginal likelihood (ML) is computed using
      Laplace-Metropolis approximation.
```

The horseshoe model, model1, has the highest marginal likelihood, −228, and thus the highest posterior probability, 0.67. This model comparison, however, is based only on the training data goodness of fit and may not reflect the actual predictive performance of the models.

◁

▷ Example 8: Model comparison using predictive performance

Here, for comparison, we also fit a BMA regression by using bmaregress (see [BMA] **bmaregress**) with default settings.

```
. bmaregress diabetes age sex bmi bp serum1-serum6 if sample == 1
Enumerating models ...
Computing model probabilities ...

Bayesian model averaging                         No. of obs         =    221
Linear regression                                No. of predictors  =     10
Model enumeration                                           Groups =     10
                                                            Always =      0
Priors:                                          No. of models      =  1,024
  Models: Beta-binomial(1, 1)                        For CPMP >= .9 =     49
   Cons.: Noninformative                         Mean model size    =  4.878
   Coef.: Zellner's g
       g: Benchmark, g = 221                     Shrinkage, g/(1+g) = 0.9955
  sigma2: Noninformative                         Mean sigma2        =  0.464

------------------------------------------------------------------------------
    diabetes |      Mean   Std. dev.                          Group        PIP
-------------+----------------------------------------------------------------
         bmi |  .3383962   .0623547                              3          1
      serum5 |  .3312051   .0893712                              9     .99817
          bp |  .1567729   .0748241                              4     .89364
      serum3 | -.1128554   .1054237                              7     .63432
      serum1 |  -.083198   .1591955                              5      .3693
         sex | -.0377099   .0603943                              2     .34986
      serum2 |  .0254135   .1165365                              6     .22639
      serum4 | -.0092812   .0594432                              8     .16416
         age | -.0099652   .0313711                              1     .15038
      serum6 | -.0030427   .0192269                             10    .091849
-------------+----------------------------------------------------------------
Always       |
       _cons | -.0082592   .0459976                              0          1
------------------------------------------------------------------------------
Note: Coefficient posterior means and std. dev. estimated from 1,024 models.
Note: Default priors are used for models and parameter g.
```

BMA also identifies bmi, serum5, and bp as the top three predictors.

We compute the squared prediction error for BMA and store it in the sqerrbma variable.

```
. bmapredict double pbmamean if sample == 2, mean
note: computing analytical posterior predictive means.
. generate double sqerrbma = (diabetes-pbmamean)^2
(221 missing values generated)
. drop pbmamean
```

To compare the predictive performance of the five models, we summarize the squared errors of their predicted posterior means.

```
. summarize sqerr1 sqerr2 sqerr3 sqerr4 sqerrbma
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      sqerr1 |        221    .5454139    .6604434   .0001013   3.788343
      sqerr2 |        221    .5458172    .6655874   .0000799   3.735547
      sqerr3 |        221    .5471472    .6654343   6.97e-06   3.828219
      sqerr4 |        221    .5559538    .6602583   .0000311   3.620555
    sqerrbma |        221    .5458022     .650325   .0002141   3.846061
```

The horseshoe model has the lowest mean squared error of 0.545 (variable `sqerr1`), followed by BMA (variable `sqerrbma`) and Bayesian lasso (variable `sqerr2`). Overall, the differences between the models are rather small. In this example, it appears that both the goodness-of-fit and out-of-sample prediction criteria slightly favor the horseshoe model.

Now that we are finished with our analysis, we delete the simulation datasets and extra variables we have created.

```
. rm model1sim.dta
. rm model2sim.dta
. rm model3sim.dta
. rm model4sim.dta
. drop sqerr1 sqerr2 sqerr3 sqerr4 sqerrbma sample
```

◁

Stored results

See *Stored results* in [BAYES] **bayesmh**, except the `e(exclude)` result, which is not applicable to `bayesselect`.

In addition, `bayesselect` stores the following in `e()`:

Scalars

`e(ssprior_scale1)`	first scale parameter of spike-and-slab prior
`e(ssprior_scale2)`	second scale parameter of spike-and-slab prior
`e(ssprior_sd1)`	first standard deviation parameter of spike-and-slab prior
`e(ssprior_sd2)`	second standard deviation parameter of spike-and-slab prior
`e(betaprior_shape1)`	first shape parameter of beta prior for spike-and-slab hyperparameter
`e(betaprior_shape2)`	second shape parameter of beta prior for spike-and-slab hyperparameter
`e(priorsigma)`	standard deviation of normal prior for the intercept
`e(glprior_scale)`	scale for global–local shrinkage prior
`e(conjugate)`	`1` if `conjugate` is specified, `0` otherwise
`e(cutoff)`	cutoff inclusion value

Macros

`e(glprior)`	type of global–local shrinkage prior
`e(ssprior)`	type of spike-and-slab prior

Matrices

`e(inclusion)`	MCMC inclusion values
`e(summary)`	MCMC summary matrix for model parameters other than regression coefficients

Methods and formulas

Methods and formulas are presented under the following headings:

Global–local shrinkage priors
Spike-and-slab priors

We consider a linear regression of a continuous response y with p potential predictors $x_1, x_2, \ldots, x_p$. Specifically,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha + \epsilon_i$$

where for an observation $i = 1, 2, \ldots, n$, y_i is the observed response value, $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ is the observed vector of predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of unknown regression coefficients, α is an unknown intercept, $\epsilon_i \sim N(0, \sigma^2)$ are i.i.d. errors, and σ^2 is the error variance.

The importance of different predictors in modeling y may vary. Variable selection identifies more important predictors of y for more efficient estimation and better prediction performance.

In contrast to model-selection methodologies that rely on inclusion or exclusion of predictors, Bayesian variable selection considers all potential predictors simultaneously and provides a variety of prior distributions for the vector of coefficients $\boldsymbol{\beta}$ to account for the importance of predictors.

The `bayesselect` command supports two main classes of priors for regression coefficients $\boldsymbol{\beta}$: global–local shrinkage priors and spike-and-slab priors.

The default prior for the intercept α is normal,

$$\alpha \sim N(0, \sigma_0^2)$$

where the prior standard deviation σ_0 is controlled by the `normalprior()` option. The default value for σ_0 is 100, the same as the one used by [BAYES] **bayes: regress** and other Bayes prefix commands. This is typically a fairly uninformative prior for α.

The default prior for σ^2 is the Jeffreys prior,

$$\sigma^2 \sim 1/\sigma^2$$

which can be changed by using the `prior()` option.

Global–local shrinkage priors

Global–local shrinkage priors are normal distributions that come in two forms: the nonconjugate form,

$$\beta_j | \lambda_j^2, \tau^2, \sigma^2 \sim N(0, \lambda_j^2 \tau^2) \tag{1}$$

or the conjugate form,

$$\beta_j | \lambda_j^2, \tau^2, \sigma^2 \sim N(0, \lambda_j^2 \tau^2 \sigma^2) \tag{2}$$

where τ is a global scale parameter and λ_j's are independent local scale parameters with prior distributions,

$$\tau \sim \psi(\tau)$$
$$\lambda_j \sim \phi(\lambda_j)$$

For the purpose of shrinkage, prior distribution $\psi(\cdot)$ should have a substantial mass near zero, and $\phi(\cdot)$ should have heavy tails (Polson and Scott 2011). The ability of global–local shrinkage priors to discriminate a signal from a noise is due to the combination of the global shrinkage τ and heavy-tailed local shrinkages λ_j's.

Carvalho, Polson, and Scott (2009) introduced a shrinkage coefficient $\kappa_j = (1 + \lambda_j^2/\lambda_0^2)^{-1}$, where λ_0 is a scale constant (to be defined later), and Cadonna, Frühwirth-Schnatter, and Knaus (2020) proposed to use them to determine variable inclusion: the jth variable is considered to be included if $\kappa_j < 0.5$. This notion of inclusion is used only for reporting and interpretation. The Bayesian variable selection accounts for all potential predictors and does not discard any of them during estimation.

For the global–local shrinkage prior models, we define a more convenient statistic, what we call an inclusion coefficient, $\gamma_j = 1 - \kappa_j$, to be used as a criterion for variable inclusion. Because γ_j's are random parameters, `bayesselect` computes their posterior means and reports those coefficients for which the means are above a given threshold, 0.1 by default. We can use the `cutoff(#)` option to change the default value.

Prior (2) is a standard conjugate prior for coefficients in a Bayesian linear regression. However, some researchers (Moran, Ročková, and George 2019) argue that using (2) leads to underestimation of error variance σ^2 and give preference to prior (1), which is the default in `bayesselect`. You can specify prior (2) by using the `conjugate` option.

The default prior for the hyperparameter τ is

$$\tau \sim \text{HalfCauchy}(0, 1)$$

You can use the `prior()` option to specify a different prior for τ.

There are two common choices for the prior distribution $\phi(\cdot)$.

1. The horseshoe prior (Carvalho, Polson, and Scott 2009) is a special form of a global–local shrinkage prior with

   $$\lambda_j \sim \text{HalfCauchy}(0, \lambda_0)$$

   where λ_0 is a scale parameter. HalfCauchy$(0, \lambda_0)$ distribution has heavier tails than the normal distribution and is simply a truncated Cauchy distribution. By default, $\lambda_0 = 1$, but you can change this by using the `hshoe(#)` option.

   It can be shown that the prior distribution for the shrinkage coefficient $\kappa_j = (1 + \lambda_j^2/\lambda_0^2)^{-1}$ is Beta$(0.5, 0.5)$, which resembles a horseshoe and thus gives the prior its name.

2. The Bayesian lasso (Park and Casella 2008) is another special case of a global–local shrinkage prior with

   $$\lambda_j \sim \text{Rayleigh}(\lambda_0)$$

   which is equivalent to

   $$\lambda_j^2 \sim \text{Exponential}(2\lambda_0^2)$$

   where λ_0 is a scale parameter. The default is $\lambda_0 = 1$, which can be changed by using the `blasso(#)` option.

It can be shown that in the nonconjugate case 1, the marginal prior distribution of β_j is Laplace$(\lambda_0\tau)$ and that in the conjugate case 2, the marginal prior distribution of β_j is Laplace$(\lambda_0\tau\sigma)$. The marginal prior log-density of β_j is thus proportional to $-|\beta_j|$, which is precisely the l_1-penalty term in standard lasso.

Spike-and-slab priors

The original version of this prior was proposed by Mitchell and Beauchamp (1988),

$$\beta_j|\gamma_j \sim (1 - \gamma_j)\delta_0(\beta_j) + \gamma_j\phi_1(\beta_j) \tag{3}$$

where γ_j's are independent binary indicators, $\delta_0(\cdot)$ is the delta function (with a mass concentrated only at zero), and $\phi_1(\cdot)$ is a continuous density. $\delta_0(\cdot)$ is the spike and $\phi_1(\cdot)$ is the slab component of the prior. Difficulties in implementing an efficient sampling for this prior led to the development of various alternatives.

Following the terminology of global–local shrinkage models, we call γ_j an inclusion coefficient and $\kappa_j = 1-\gamma_j$ a shrinkage coefficient. Unlike global–local shrinkage models, inclusion coefficients γ_j's can be interpreted as actual inclusion probabilities. The `bayesselect` command computes their posterior means and reports those coefficients for which the posterior mean is above a given threshold, 0.1 by default. You can use the `cutoff(#)` option to change this value.

The variable-selection effect of the spike-and-slab priors is sensitive to the distribution of the predictors. It is recommended that predictors x_1 through x_p be centered before estimation such that $n\overline{x}_j = \sum_{i=1}^n x_{ji} = 0$, for $j = 1, 2, \dots, p$. If predictors are distributed away from zero, spike-and-slab priors may not be effective in distinguishing between important and unimportant predictors. In this regard, the normal-mixture spike-and-slab priors are more robust than the Laplace-mixture spike-and-slab priors. There is no threshold for $|\overline{x}_j|$ beyond which we should not use spike-and-slab priors—the diminishing effect of the priors is gradual. Ishwaran and Rao (2005) derive consistency properties of spike-and-slab priors under the orthogonality of the design matrix assumption, $\mathbf{X}'\mathbf{X} = n\mathbf{I_n}$, which implies that $\overline{x}_j^2 \leq 1$, for $j = 1, 2, \dots, p$. There is also the so-called vanishing effect of the priors as the sample increases, where the data dominate the specified prior information, which is a general problem in Bayesian analysis. To counteract the vanishing effect of spike-and-slab priors, Ishwaran and Rao (2005) recommend centering the outcome y and rescaling it by a factor of $\sqrt{n}$.

Below, we describe two variations of the spike-and-slab priors.

1. George and McCulloch (1993) proposed an alternative to (3), which is more tractable computationally, using normal distributions in place of the original $\delta_0(\cdot)$ and $\phi_1(\cdot)$ densities:

   $$\beta_j|\gamma_j \sim (1-\gamma_j)\phi_0(\beta_j) + \gamma_j\phi_1(\beta_j)$$

   The $\phi(\cdot)$ distributions are normal with the default forms of

   $$\phi_0(\cdot)\text{: } N(0, \tau_0^2);\;\; \phi_1(\cdot)\text{: } N(0, \tau_1^2)$$

   where $0 < \tau_0^2 \ll \tau_1^2$.

   Alternatively, when the `conjugate` option is specified, `bayesselect` uses the conjugate forms

   $$\phi_0(\cdot)\text{: } N(0, \sigma^2\tau_0^2);\;\; \phi_1(\cdot)\text{: } N(0, \sigma^2\tau_1^2)$$

   The defaults for the standard deviations are $\tau_0 = 0.01$ and $\tau_1 = 1$. These can be changed by using the `ssnormal(`*#1 #2*`)` option.

2. The spike-and-slab lasso model (Ročková and George 2018) uses a mixture of Laplace distributions:

   $$\beta_j|\gamma_j \sim (1-\gamma_j)\phi_0(\beta_j) + \gamma_j\phi_1(\beta_j)$$

   The $\phi(\cdot)$ distributions are Laplace with the default forms of

   $$\phi_0(\cdot)\text{: Laplace}(\lambda_0);\;\; \phi_1(\cdot)\text{: Laplace}(\lambda_1)$$

   where λ_0 and λ_1 are the scale parameters.

   When the `conjugate` option is specified, `bayesselect` uses the conjugate forms,

   $$\phi_0(\cdot)\text{: Laplace}(\sigma\lambda_0);\;\; \phi_1(\cdot)\text{: Laplace}(\sigma\lambda_1)$$

   We use the scale-form representation of the Laplace distribution:

   $$\phi(\beta|\lambda) = \frac{\lambda}{2}e^{-|\beta|/\lambda}$$

   The defaults for the scale parameters are $\lambda_0 = 0.01$ and $\lambda_1 = 1$. These can be changed by using the `sslaplace(`*#1 #2*`)` option.

Conditions that guarantee variable-selection consistency are considered in Narisetty and He (2014), Narisetty (2022), and Ishwaran and Rao (2005). Specifically, conditions for strong selection consistency require that $\tau_0^2 = o(n^{-1})$ and $\tau_1^2 = O(1 + p^c n^{-1})$, for $c > 2$ and $\theta = O(p^{-1})$, where θ is the hyperparameter of the prior.

The Gibbs sampling for the spike-and-slab lasso model implemented in `bayesselect` is based on a hierarchical representation of the Laplace distribution detailed in Andrews and Mallows (1974) and Park and Casella (2008).

In both spike-and-slab models, the binary indicators γ_j's have independent Bernoulli prior distributions,

$$\gamma_j \sim \text{Bernoulli}(\theta)$$

with a beta distribution with shapes a and b for the hyperparameter θ,

$$\theta \sim \text{Beta}(a, b)$$

The prior on θ controls the sparsity of the regression model.

The defaults for the shape parameters of the beta prior are $a = 1$ and $b = 1$, which corresponds to a uniform on [0,1] prior distribution for θ. You can change these default values by using the `betaprior(`*#1 #2*`)` option. Or you can use the `prior()` option to specify a different prior for θ.

`bayesselect` uses efficient Gibbs sampling for regression coefficients $\boldsymbol{\beta}$, intercept α, latent parameters λ_j's and γ_j's, and hyperparameter θ. An adaptive Metropolis–Hastings sampling is used for σ^2 by default; see *Methods and formulas* of [BAYES] **bayesmh**.

References

Andrews, D. F., and C. L. Mallows. 1974. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, B ser.*, 36: 99–102. https://doi.org/10.1111/j.2517-6161.1974.tb00989.x.

Cadonna, A., S. Frühwirth-Schnatter, and P. Knaus. 2020. Triple the gamma—a unifying shrinkage prior for variance and variable selection in sparse state space and TVP models. *Econometrics* 8, no. 20. https://doi.org/10.3390/econometrics8020020.

Carvalho, C. M., N. G. Polson, and J. G. Scott. 2009. “Handling Sparsity via the Horseshoe”. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, vol. 5: 73–80. Clearwater Beach, FL: Proceedings of Machine Learning Research.

Castillo, I., and A. W. van der Vaart. 2012. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics* 40: 2069–2101. https://doi.org/10.1214/12-AOS1029.

Efron, B. 2008. Microarrays, empirical Bayes and the two-groups model. *Statistical Science* 23: 1–22. https://doi.org/10.1214/07-STS236.

Efron, B., T. J. Hastie, I. M. Johnstone, and R. J. Tibshirani. 2004. Least angle regression. *Annals of Statistics* 32: 407–499. https://doi.org/10.1214/009053604000000067.

George, E. I., and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88: 881–889. https://doi.org/10.2307/2290777.

Griffin, J. E., and P. J. Brown. 2010. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5: 171–188. https://doi.org/10.1214/10-BA507.

Ishwaran, H., and J. S. Rao. 2005. Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics* 33: 730–773. https://doi.org/10.1214/009053604000001147.

Johnstone, I. M., and B. W. Silverman. 2004. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* 32: 1594–1649. https://doi.org/10.1214/009053604000000030.

Mitchell, T. J., and J. J. Beauchamp. 1988. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83: 1023–1032. https://doi.org/10.2307/2290129.

Moran, G. E., V. Ročková, and E. I. George. 2019. Variance prior forms for high-dimensional Bayesian variable selection. *Bayesian Analysis* 14: 1091–1119. https://doi.org/10.1214/19-BA1149.

Narisetty, N. N. 2022. "Theoretical and computational aspects of continuous spike-and-slab priors". In *Handbook of Bayesian Variable Selection*, edited by M. G. Tadesse and M. Vannucci, 57–80. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781003089018.

Narisetty, N. N., and X. He. 2014. Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics* 42: 789–817. https://doi.org/10.1214/14-AOS1207.

Park, T., and G. Casella. 2008. The Bayesian lasso. *Journal of the American Statistical Association* 103: 681–686. https://doi.org/10.1198/016214508000000337.

Polson, N. G., and J. G. Scott. 2011. "Shrink globally, act locally: Sparse Bayesian regularization and prediction". In *Bayesian Statistics: Proceedings of the Ninth Valencia International Meeting, June 3–8, 2010*, edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, vol. 9: 501–538. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199694587.003.0017.

Ročková, V., and E. I. George. 2018. The spike-and-slab lasso. *Journal of the Royal Statistical Society*, B ser., 113: 431–444. https://doi.org/10.1080/01621459.2016.1260469.

Tibshirani, R. J. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B ser., 58: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Also see

[BAYES] **Bayesian postestimation** — Postestimation tools after Bayesian estimation

[BAYES] **bayesmh** — Bayesian models using Metropolis–Hastings algorithm

[BAYES] **Intro** — Introduction to Bayesian analysis

[BMA] **bmaregress** — Bayesian model averaging for linear regression

[LASSO] **lasso** — Lasso for prediction and model selection

[U] 20 Estimation and postestimation commands

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. StataNow and NetCourseNow are trademarks of StataCorp LLC. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2025 StataCorp LLC, College Station, TX, USA. All rights reserved.

For suggested citations, see the FAQ on citing Stata documentation.