Package 'EMVS'

December 14, 2019

Type Package

Title The Expectation-Maximization Approach to Bayesian Variable Selection

Version 1.1

Date 2019-12-13

Author Veronika Rockova [aut,cre], Gemma Moran [aut]

Maintainer Gemma Moran <gm2918@columbia.edu>

Description An efficient expectation-maximization algorithm for fitting Bayesian spike-and-slab regularization paths for linear regression. Rockova and George (2014) <doi:10.1080/01621459.2013.869223>.

License GPL-3

URL https://doi.org/10.1080/01621459.2013.869223

Imports Rcpp (>= 0.12.16), methods

LinkingTo Rcpp, RcppArmadillo

Suggests knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 6.0.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-12-13 23:20:21 UTC

R topics documented:

EMVS 2
EMVSbest 5
EMVSplot 6
EMVSsummary 7

Index **8**

EMVS *Bayesian Variable Selection using EM Algorithm*

Description

EMVS is a fast deterministic approach to identifying sparse high posterior models for Bayesian variable selection under spike-and-slab priors in linear regression. EMVS performs dynamic posterior exploration, which outputs a solution path computed at a grid of values for the spike variance parameter `v0`.

Usage

```
EMVS(Y, X, v0, v1, type = c("betabinomial", "fixed"), independent = TRUE,
      beta_init, sigma_init, epsilon = 10^(-5), temperature, theta, a, b, v1_g,
   direction=c("backward", "forward", "null"), standardize = TRUE, log_v0 = FALSE)
```

Arguments

`Y`	Vector of continuous responses (n x 1). The responses are expected to be centered.
`X`	Matrix of regressors (n x p). Continous predictors are expected to be standardized to have mean zero and standard deviation one.
`v0`	Spike variance parameters. Either a numeric value for a single run or a sequence of increasing values for dynamic posterior exploration.
`v1`	Slab variance parameter. Needs to be greater than v0.
`type`	Type of the prior distribution over the model space: `type="betabinomial"` for the betabinomial prior with shape parameters `a` and `b`, `type="fixed"` for the Bernoulli prior with a fixed inclusion probability `theta`.
`independent`	If TRUE, the regression coefficients and the error variance are taken to be independent a priori (default). If FALSE, a conjugate prior is used as in Rockova and George (2014).
`beta_init`	Vector (p x 1) of initial values for the regression parameters `beta`. If missing, a default vector of starting values obtained as a limiting case of deterministic annealing used $$beta^0 = [X'X + 0.5(1/v1 + 1/v0)I_p]^{-1}X'Y.$$
`sigma_init`	Initial value for the residual variance parameter.
`epsilon`	Convergence margin parameter. The computation at each `v0` is terminated when $$\|\|beta^{k+1} - beta^k\|\|_2 < epsilon.$$
`temperature`	Temperature parameter for deterministic annealing. If missing, a default value `temperature=1` used.
`theta`	Prior inclusion probability for `type="fixed"`.

a,b	Scale parameters of the beta distribution for type="betabinomial".
v1_g	Slab variance parameter value for the g-function. If missing, a default value v1 used.
direction	Direction of the sequential reinitialization in dynamic posterior exploration. The default is direction="backward" - this initializes the first computation at beta_init using the largest value of v0 and uses the resulting output as a warm start for the next largest value v0 in a backward direction (i.e. from the largest to the smallest value of v0). The option direction="forward" proceeds from the smallest value of v0 to the largest value of v0, using the output from the previous solution as a warm start for the next. direction = "null" re-initializes at beta_init for each v0.
standardize	If TRUE (default), the design matrix X is standardized (mean zero and variance n).
log_v0	If TRUE, the v0s are displayed on the log scale in EMVSplot.

Details

An EM algorithm is applied to find posterior modes of the regression parameters in linear models under spike and slab priors. Variable selection is performed by threshodling the posterior modes to obtain models gamma with high posterior probability P(gamma|Y) . The spike variance v0 can be altered to obtain models with various degrees of sparsity. The slab variance is set to a fixed value v1>v0. The thresholding is based on the conditional posterior probabilities of inclusion, which are outputed of the procedure. Variables are included as long as their inclusion probability is above 0.5. Dynamic exploration is achieved by considering a sequence of increasing spike variance parameters v0. For each v0, a candidate model is obtained. For the conjugate prior case, the best model is then picked according to a criterion ("log g-function"), which equals to the log of the posterior model probability up to a constant

$$logg(gamma) = logP(gamma|Y) + C.$$

Independent and sequential initializations are implemented. Sequential initialization uses previously found modes are warm starts in both forward and backward direction of the given sequence of v0 values.

Value

A list object, for which EMVSplot and EMVSbest functions exist.

betas	Matrix of estimated regression coefficients (posterior modal estimates) of dimension (L x p), where L is the length of v0.
log_g_function	Vector (L x 1) of log posterior model probabilities (up to a constant) of subsets found for each v0. (Only available for independent = FALSE).
intersects	Vector (L x 1) of posterior weighted intersection points between spike and slab components.
sigmas	Vector (L x 1) of estimated residual variances.
v1	Slab variance parameter values used.
v0	Spike variance parameter values used.

niters	Vector (L x 1) of numbers of interations until convergence for each v0
prob_inclusion	A matrix (L x p) of conditional inclusion probabilities. Each row corresponds to a single v0 value.
type	Type of the model prior used.
type	Type of initialization used, type="null" stands for the default cold start.
theta	Vector (L x 1) of estimated inclusion probabilities for type="betabinomial".

Author(s)

Veronika Rockova
Maintainer: Veronika Rockova <veronika.rockova@chicagobooth.edu>

References

Rockova, V. and George, E. I. (2014) *EMVS: The EM Approach to Bayesian Variable Selection*, http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2013.869223?journalCode=uasa20#preview
Journal of the American Statistical Association

See Also

EMVSplot, EMVSsummary, EMVSbest

Examples

```
# Linear regression with p>n variables
library(EMVS)

n = 100
p = 1000
X = matrix(rnorm(n * p), n, p)
beta = c(1.5, 2, 2.5, rep(0, p-3))
Y = X[,1] * beta[1] + X[,2] * beta[2] + X[,3] * beta[3] + rnorm(n)

# conjugate prior on regression coefficients and variance
v0 = seq(0.1, 2, length.out = 20)
v1 = 1000
beta_init = rep(1, p)
sigma_init = 1
a = b = 1
epsilon = 10^{-5}

result = EMVS(Y, X, v0 = v0, v1 = v1, type = "betabinomial",
independent = FALSE, beta_init = beta_init, sigma_init = sigma_init,
epsilon = epsilon, a = a, b = b)

EMVSplot(result, "both", FALSE)

EMVSbest(result)
```

```
# independent prior on regression coefficients and variance
v0 = exp(seq(-10, -1, length.out = 20))
v1 = 1
beta_init = rep(1,p)
sigma_init = 1
a = b = 1
epsilon = 10^{-5}

result = EMVS(Y, X, v0 = v0, v1 = v1, type = "betabinomial",
independent = TRUE, beta_init = beta_init, sigma_init = sigma_init,
epsilon = epsilon, a = a, b = b, log_v0 = TRUE)

EMVSplot(result, "both", FALSE)

EMVSbest(result)
```

`EMVSbest` *Select the Best Model with EMVS*

Description

`EMVSbest` outputs indices of the variables included in the model with the highest posterior probability found.

Usage

```
EMVSbest(result)
```

Arguments

`result`	List object outputed by the EMVS procedure

Value

`log_g_function`	The highest log-g-function found along the regularization path
`indices`	The indices of the variables included in the best model found

Author(s)

Veronika Rockova
Maintainer: Veronika Rockova <vrockova@wharton.upenn.edu>

References

Rockova, V. and George, E. I. (2014) *EMVS: The EM Approach to Bayesian Variable Selection*, http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2013.869223?journalCode=uasa20#preview
Journal of the American Statistical Association

See Also

EMVS, EMVSsummary, EMVSplot

EMVSplot *Spike-and-slab Dynamic Posterior Exploration*

Description

EMVSplot procedure plots the solution path of the estimated regression coefficients (posterior modes) for different v0 values.

Usage

```
EMVSplot(result, plot_type=c("both","reg","gf"), omit.zeroes = FALSE)
```

Arguments

result	List object outputed by the EMVS procedure
plot_type	Plot type: "both" for plotting both the regularization path together with the associated log g function, "reg" only for the regulaziation plot, "gf" only for the log g function.
omit.zeroes	Logical: TRUE or FALSE. If TRUE, only the selected coefficients are plotted, the remaining coefficients set to zero

Details

Coefficients that are not thresholded out are depicted in blue, the rest in red. Log g function computed only for models with at most 1 000 predictors.

Author(s)

Veronika Rockova
Maintainer: Veronika Rockova <vrockova@wharton.upenn.edu>

References

Rockova, V. and George, E. I. (2014) *EMVS: The EM Approach to Bayesian Variable Selection*, http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2013.869223?journalCode=uasa20#preview
Journal of the American Statistical Association

See Also

EMVS, EMVSsummary, EMVSbest

EMVSsummary *Select the Best Model with EMVS*

Description

EMVSsummary outputs variable selection indicators of models found together with the log-g-function.

Usage

```
EMVSsummary(result)
```

Arguments

result	List object outputed by the EMVS procedure

Value

log_g_function	The log-g-function computed for all models found along the regularization path
indices	The (L x p) matrix of variable selection indicators after thresholding (1 for selected, 0 for not selected). Each row corresponds to a single v0 value.

Author(s)

Veronika Rockova
Maintainer: Veronika Rockova <vrockova@wharton.upenn.edu>

References

Rockova, V. and George, E. I. (2014) *EMVS: The EM Approach to Bayesian Variable Selection*, http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2013.869223?journalCode=uasa20#preview
Journal of the American Statistical Association

See Also

EMVS, EMVSplot, EMVSbest

Index

∗Topic **Bayesian variable selection**
EMVS, 2
EMVSbest, 5
EMVSplot, 6
EMVSsummary, 7
∗Topic **Spike and slab**
EMVS, 2
EMVSbest, 5
EMVSplot, 6
EMVSsummary, 7

EMVS, 2
EMVSbest, 5
EMVSplot, 6
EMVSsummary, 7