Package: PEPBVS (via r-universe)

October 13, 2024

Type Package

Title Bayesian Variable Selection using Power-Expected-Posterior Prior

Version 1.0

Date 2023-09-14

Maintainer Konstantina Charmpi <xarmpi.kon@gmail.com>

Description Performs Bayesian variable selection under normal linear models for the data with the model parameters following as prior either the power-expected-posterior (PEP) or the intrinsic (a special case of the former) (Fouskakis and Ntzoufras (2022) <doi:10.1214/21-BA1288>, Fouskakis and Ntzoufras (2020) <doi:10.3390/econometrics8020017>). The prior distribution on model space is the uniform on model space or the uniform on model dimension (a special case of the beta-binomial prior). The selection can be done either with full enumeration of all possible models or using the Markov Chain Monte Carlo Model Composition (MC3) algorithm (Madigan and York (1995) <doi:10.2307/1403615>). Complementary functions for making predictions, as well as plotting and printing the results are also provided.

License GPL (>= 2)

Imports Matrix, Rcpp (>= 1.0.9)

LinkingTo Rcpp, RcppArmadillo, RcppGSL

SystemRequirements GNU GSL

Encoding UTF-8

RoxygenNote 7.2.1

Depends R (>= 2.10)

LazyData true

NeedsCompilation yes

Author Konstantina Charmpi [aut, cre], Dimitris Fouskakis [aut], Ioannis Ntzoufras [aut]

Repository CRAN

Date/Publication 2023-09-19 16:40:02 UTC

Contents

`PEPBVS-package` *Bayesian variable selection using power-expected-posterior prior*

Description

Performs Bayesian variable selection under normal linear models for the data with the model parameters following as prior either the PEP or the intrinsic (a special case of the former). The prior distribution on model space is the uniform on model space or the uniform on model dimension (a special case of the beta-binomial prior). Posterior model probabilities and marginal likelihoods can be derived in closed-form expressions under this setup. The selection can be done either with full enumeration of all possible models (for small–to–moderate model spaces) or using the MC3 algorithm (for large model spaces). Complementary functions for making predictions, as well as plotting and printing the results are also available.

References

Fouskakis, D. and Ntzoufras, I. (2022) Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models. Bayesian Analysis, 17(4): 1073-1099. doi:10.1214/21BA1288

Fouskakis, D. and Ntzoufras, I. (2020) Bayesian Model Averaging Using Power-Expected-Posterior Priors. Econometrics, 8(2): 17. doi:10.3390/econometrics8020017

`full_enumeration_pep` *Bayesian variable selection through exhaustive search*

Description

Given a response vector and an input data matrix, performs Bayesian variable selection using full enumeration of the model space. Normal linear models are assumed for the data with the prior distribution on the model parameters (beta coefficients and error variance) being the PEP or the intrinsic. The prior distribution on the model space can be the uniform on the model space or the uniform on the model dimension (special case of the beta-binomial prior). The model space consists of all possible models including an intercept term.

Usage

```
full_enumeration_pep(
  x,
  y,
  intrinsic = FALSE,
  reference.prior = TRUE,
  beta.binom = TRUE,
  ml_constant.term = FALSE
)
```

Arguments

`x` A matrix of numeric (of size nxp), input data matrix. This matrix contains the values of the p explanatory variables without an intercept column of 1's.

`y` A vector of numeric (of length n), response vector.

`intrinsic` Boolean, indicating whether the PEP (`FALSE`) or the intrinsic - which is a special case of it - (`TRUE`) should be used as prior on the regression parameters. Default value=`FALSE`.

`reference.prior`
Boolean, indicating whether the reference prior (`TRUE`) or the dependence Jeffreys prior (`FALSE`) is used as baseline. Default value=`TRUE`.

`beta.binom` Boolean, indicating whether the beta-binomial distribution (`TRUE`) or the uniform distribution (`FALSE`) should be used as prior on the model space. Default value=`TRUE`.

`ml_constant.term`
Boolean, indicating whether the constant (marginal likelihood of the null/intercept-only model) should be included in computing the marginal likelihood of a model (`TRUE`) or not (`FALSE`). Default value=`FALSE`.

Details

The function works when p<=n-2 where p is the number of explanatory variables and n is the sample size.

It is suggested to use this function (i.e. enumeration of the model space) when p is up to 20.

The reference model is the null model (i.e. intercept-only model).

The case of missing data (i.e. presence of `NA`'s either in the input data matrix or the response vector) is not currently supported.

All models considered (i.e. model space) include an intercept term.

If p>1, the input data matrix needs to be of full rank.

The reference prior as baseline corresponds to hyperparameter values d0=0 and d1=0, while the dependence Jeffreys prior corresponds to model-dependent-based values for the hyperparameters d0 and d1, see Fouskakis and Ntzoufras (2022) for more details.

For computing the marginal likelihood of a model, Equation 16 of Fouskakis and Ntzoufras (2022) is used.

When `ml_constant.term=FALSE` then the log marginal likelihood of a model in the output is shifted by -logC1 (logC1: log marginal likelihood of the null model).

When the prior on the model space is beta-binomial (i.e. `beta.binom=TRUE`), the following special case is used: uniform prior on model dimension.

Value

`full_enumeration_pep` returns an object of class `pep`, i.e. a list with the following elements:

`models` A matrix containing information about the models examined. In particular, in row i after representing the model i with variable inclusion indicators, its marginal likelihood (in log scale), the R2, its dimension (including the intercept), the corresponding Bayes factor, posterior odds and its posterior probability are contained. The models are sorted in decreasing order of the posterior probability. For the Bayes factor and the posterior odds, the comparison is done to the model with the largest posterior probability.

`inc.probs` A named vector with the posterior inclusion probabilities of the explanatory variables.

`x` The input data matrix (of size nxp).

`y` The response vector (of length n).

`intrinsic` Whether the prior on the model parameters was PEP or intrinsic.

`reference.prior`
Whether the baseline prior was the reference prior or the dependence Jeffreys prior.

`beta.binom` Whether the prior on the model space was beta-binomial or uniform.

References

Fouskakis, D. and Ntzoufras, I. (2022) Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models. Bayesian Analysis, 17(4): 1073-1099. doi:10.1214/21BA1288

See Also

`mc3_pep`

Examples

```
data(UScrime_data)
y <- UScrime_data[,"y"]
X <- UScrime_data[,-15]
res <- full_enumeration_pep(X,y)
resu <- full_enumeration_pep(X,y,beta.binom=FALSE)
resi <- full_enumeration_pep(X,y,intrinsic=TRUE)
resiu <- full_enumeration_pep(X,y,intrinsic=TRUE,beta.binom=FALSE)
resj <- full_enumeration_pep(X,y,reference.prior=FALSE)
```

image.pep *Heatmap for top models*

Description

Generates a heatmap where the rows correspond to the (top) models and the columns to the input/explanatory variables. The value depicted in cell (i,j) corresponds to the posterior inclusion probability of variable i if this is included in model j and 0 otherwise.

Usage

```
## S3 method for class 'pep'
image(x, n.models = 20, ...)
```

Arguments

`x`	An object of class pep (e.g. output of `full_enumeration_pep` or `mc3_pep`).
`n.models`	Positive integer, number of models to be shown on the heatmap. Default value=20.
`...`	Additional parameters to be passed to `heatmap`.

Details

The number of models to be displayed on the heatmap is computed as the minimum between the number asked by the user and the number of models present in the object `x`.

The color code is as follows: the darker the blue in the figure, the higher the posterior inclusion probability is, while white means that the variable is not included in the model.

In the special case of no explanatory variables, no heatmap is produced and a message is printed.

Value

No return value, used for generating a heatmap.

See Also

`plot.pep`

Examples

```
data(UScrime_data)
y <- UScrime_data[,"y"]
X <- UScrime_data[,-15]
set.seed(123)
resu <- mc3_pep(X,y,beta.binom=FALSE,itermc3=5000)
image(resu)
image(resu,n.models=10)
```

mc3_pep *Bayesian variable selection with MC3 algorithm*

Description

Given a response vector and an input data matrix, performs Bayesian variable selection using the MC3 algorithm. Normal linear models are assumed for the data with the prior distribution on the model parameters (beta coefficients and error variance) being the PEP or the intrinsic. The prior distribution on the model space can be the uniform on the model space or the uniform on the model dimension (special case of the beta-binomial prior).

Usage

```
mc3_pep(
  x,
  y,
  intrinsic = FALSE,
  reference.prior = TRUE,
  beta.binom = TRUE,
  ml_constant.term = FALSE,
  burnin = 1000,
  itermc3 = 11000
)
```

Arguments

`x`	A matrix of numeric (of size nxp), input data matrix. This matrix contains the values of the p explanatory variables without an intercept column of 1's.
`y`	A vector of numeric (of length n), response vector.
`intrinsic`	Boolean, indicating whether the PEP (`FALSE`) or the intrinsic - which is a special case of it - (`TRUE`) should be used as prior on the regression parameters. Default value=`FALSE`.
`reference.prior`	Boolean, indicating whether the reference prior (`TRUE`) or the dependence Jeffreys prior (`FALSE`) is used as baseline. Default value=`TRUE`.
`beta.binom`	Boolean, indicating whether the beta-binomial distribution (`TRUE`) or the uniform distribution (`FALSE`) should be used as prior on the model space. Default value=`TRUE`.
`ml_constant.term`	Boolean, indicating whether the constant (marginal likelihood of the null/intercept-only model) should be included in computing the marginal likelihood of a model (`TRUE`) or not (`FALSE`). Default value=`FALSE`.
`burnin`	Non-negative integer, the burnin period for the MC3 algorithm. Default value=1000.
`itermc3`	Positive integer (larger than `burnin`), the (total) number of iterations for the MC3 algorithm. Default value=11000.

Details

The function works when p<=n-2 where p is the number of explanatory variables and n is the sample size.

It is suggested to use this function (i.e. MC3 algorithm) when p is larger than 20.

The reference model is the null model (i.e. intercept-only model).

The case of missing data (i.e. presence of `NA`'s either in the input matrix or the response vector) is not currently supported.

The intercept term is included in all models.

If p>1, the input matrix needs to be of full rank.

The reference prior as baseline corresponds to hyperparameter values d0=0 and d1=0, while the dependence Jeffreys prior corresponds to model-dependent-based values for the hyperparameters d0 and d1, see Fouskakis and Ntzoufras (2022) for more details.

The MC3 algorithm was first introduced by Madigan and York (1995) while its current implementation is described in the Appendix of Fouskakis and Ntzoufras (2022).

When `ml_constant.term=FALSE` then the log marginal likelihood of a model in the output is shifted by -logC1 (logC1: marginal likelihood of the null model).

When the prior on the model space is beta-binomial (i.e. `beta.binom=TRUE`), the following special case is used: uniform prior on model size.

Value

`mc3_pep` returns an object of class `pep`, as this is described in detail in `full_enumeration_pep`. The difference is that here the number of rows of the first list element is not 2^p but the number of unique models 'visited' by the MC3 algorithm. Further, the posterior probability of a model corresponds to the estimated posterior probability as this is computed by the relative Monte Carlo frequency of the 'visited' models by the MC3 algorithm.

References

Fouskakis, D. and Ntzoufras, I. (2022) Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models. Bayesian Analysis, 17(4): 1073-1099. doi:10.1214/21BA1288

Madigan, D. and York, J. (1995) Bayesian Graphical Models for Discrete Data. International Statistical Review, 63(2): 215–232. doi:10.2307/1403615

See Also

`full_enumeration_pep`

Examples

```
data(UScrime_data)
y <- UScrime_data[,"y"]
X <- UScrime_data[,-15]
set.seed(123)
res <- mc3_pep(X,y,itermc3=3000)
resu <- mc3_pep(X,y,beta.binom=FALSE,itermc3=3000)
```

```
resj <- mc3_pep(X,y,reference.prior=FALSE,burnin=500,itermc3=2200)
```

plot.pep *Plots for object of class pep*

Description

Generates four plots related to an object of class pep. In particular, the first one is a plot of the residuals against fitted values under Bayesian model averaging. The second plots the cumulative posterior probability of the top models (those with cumulative posterior probability larger than 0.99). The third plot depicts the marginal likelihood (in log scale) of a model against its dimension while the fourth plot shows the posterior inclusion probabilities of the explanatory variables (with those exceeding 0.5 marked in red).

Usage

```
## S3 method for class 'pep'
plot(x, ...)
```

Arguments

x	An object of class pep (e.g. output of `full_enumeration_pep` or `mc3_pep`).
...	Additional graphical parameters to be passed to plotting functions.

Details

Let k be the number of models with cumulative posterior probability up to 0.99. Then, the second plot depicts the cumulative posterior probability of the top (k+1) models.

In the special case of no explanatory variables, the fourth plot with the posterior inclusion probabilities is not generated.

Value

No return value, used for generating a figure.

See Also

image.pep

Examples

```
data(UScrime_data)
y <- UScrime_data[,"y"]
X <- UScrime_data[,-15]
res <- full_enumeration_pep(X,y)
plot(res)
```

predict.pep *Prediction under PEP approach*

Description

Computes predicted or fitted values under the PEP approach. Predictions can be based on Bayesian model averaging, maximum a posteriori model or median probability model. For the Bayesian model averaging, a subset of the top models (either based on explicit number or on their cumulative probability) can be used for prediction.

Usage

```
## S3 method for class 'pep'
predict(
  object,
  xnew,
  estimator = "BMA",
  n.models = NULL,
  cumul.prob = 0.99,
  ...
)
```

Arguments

`object`	An object of class pep (e.g. output of `full_enumeration_pep` or `mc3_pep`).
`xnew`	A matrix of numeric (with p columns), the new data to be used for prediction. This matrix contains the values of the explanatory variables without an intercept column of 1's, i.e. the number of its columns coincides with the number of columns of `object$x`. If omitted, fitted values are computed.
`estimator`	A character, the type of prediction. One of "BMA" (Bayesian model averaging, default), "MAP" (maximum a posteriori model) or "MPM" (median probability model).
`n.models`	Positive integer, the number of (top) models that prediction is based on or `NULL`. Relevant for `estimator="BMA"`. Default value=`NULL`.
`cumul.prob`	Numeric between 0 and 1, cumulative probability of top models to be used for prediction. Relevant for `estimator="BMA"`. Default value=0.99.
`...`	Additional parameters to be passed, currently none.

Details

When `xnew` is missing or `xnew=object$x` then fitted values are computed (and returned).

For prediction, Equation 9 of Fouskakis and Ntzoufras (2020) is used.

The case of missing data (i.e. presence of NA's) in the new data matrix is not currently supported.

Let k be the number of models with cumulative posterior probability up to the given value of `cumul.prob`. Then, for Bayesian model averaging the prediction is based on the top (k+1) models if they exist, otherwise on the top k models.

When both `n.models` and `cumul.prob` are provided - once specifying the number of models for the given cumulative probability as described above - the minimum between the two numbers is used for prediction.

Value

`predict` returns a vector with the predicted (or fitted) values for the different observations.

References

Fouskakis, D. and Ntzoufras, I. (2022) Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models. Bayesian Analysis, 17(4): 1073-1099. doi:10.1214/21BA1288

Fouskakis, D. and Ntzoufras, I. (2020) Bayesian Model Averaging Using Power-Expected-Posterior Priors. Econometrics, 8(2): 17. doi:10.3390/econometrics8020017

Examples

```
data(UScrime_data)
y <- UScrime_data[,"y"]
X <- UScrime_data[,-15]
set.seed(123)
res <- mc3_pep(X[1:45,],y[1:45],intrinsic=TRUE,itermc3=4000)
resf <- predict(res)
resf2 <- predict(res,estimator="MPM")
resp <- predict(res,xnew=X[46:47,])
```

`print.pep` *Printing object of class pep*

Description

For each of the top models (shown in columns), the following information is printed: the model representation using variable inclusion indicators, its marginal likelihood (in log scale), the R2, the model dimension, the Bayes factor, posterior odds and posterior probability. An additional column with the posterior inclusion probabilities of the explanatory variables is also printed.

Usage

```
## S3 method for class 'pep'
print(
  x,
  n.models = 5,
  actual.PO = FALSE,
  digits = max(3L, getOption("digits") - 3L),
  ...
)
```

Arguments

`x` An object of class pep (e.g. output of `full_enumeration_pep` or `mc3_pep`).

`n.models` Positive integer, the number of top models for which information is provided. Default value=5.

`actual.PO` Boolean, relevant for the MC3 algorithm. If `TRUE` then apart from the estimated posterior odds, the actual posterior odds of the top models (i.e. ratios based on the marginal likelihood times prior probability) are also printed - which could be used as a convergence indicator of the algorithm. Default value=`FALSE`.

`digits` Positive integer, the number of digits for printing numbers. Default value=`max(3L, getOption("digits") - 3L)`.

`...` Additional parameters to be passed to `print.default`.

Details

The number of models for which information is provided, is computed as the minimum between the number asked by the user and the number of models present in the object `x`.

Value

No return value, used for printing the results on the R console.

Examples

```
data(UScrime_data)
y <- UScrime_data[,"y"]
X <- UScrime_data[,-15]
res <- full_enumeration_pep(X,y)
print(res)
```

`UScrime_data` *US Crime Data*

Description

The dataset has been borrowed from the MASS R package and describes the effect of punishment regimes on crime rates. One explanatory variable (indicator variable for a Southern state) was removed since it was binary.

Format

This data frame contains the following columns:

`M` percentage of males aged 14–24.

`Ed` mean years of schooling.

`Po1` police expenditure in 1960.

`Po2` police expenditure in 1959.

`LF` labour force participation rate.

`M.F` number of males per 1000 females.

`Pop` state population.

`NW` number of non-whites per 1000 people.

`U1` unemployment rate of urban males 14–24.

`U2` unemployment rate of urban males 35–39.

`GDP` gross domestic product per head.

`Ineq` income inequality.

`Prob` probability of imprisonment.

`Time` average time served in state prisons.

`y` rate of crimes in a particular category per head of population.

Source

Data from the R package MASS

Index