Package 'bplsr'

November 12, 2024

Title Bayesian partial least squares regression

Version 1.0.0

Maintainer Szymon Urbas <szymon.urbas@mu.ie>

Description Fits the Bayesian partial least squares regression model introduced in Urbas et al. (2024) <doi:10.1214/24-AOAS1947>. Suitable for univariate and multivariate regression with high-dimensional data.

License GPL (>= 3)

Encoding UTF-8

Language eng

RoxygenNote 7.3.2

Depends R (>= 2.10)

LazyData true

Imports coda, progress, statmod, stats, utils

NeedsCompilation no

Author Szymon Urbas [aut, cre],
Pierre Lovera [ctb],
Robert Daly [ctb],
Alan O'Riordan [ctb],
Donagh Berry [ctb],
Isobel Claire Gormley [ctb]

Repository CRAN

Date/Publication 2024-11-12 14:50:02 UTC

Contents

bplsr *Run the BPLS regression model*

Description

Posterior inference of the Bayesian partial least squares regression model using a Gibbs sampler. There are three types of models available depending on the assumed prior structure on the model parameters (see details).

Usage

```
bplsr(
  X,
  Y,
  Xtest = NULL,
  Prior = NULL,
  Qs = NULL,
  N_MCMC = 20000,
  BURN = ceiling(0.3 * N_MCMC),
  Thin = 1,
  model.type = "standard",
  scale. = TRUE,
  center. = TRUE,
  PredInterval = 0.95
)
```

Arguments

X	Matrix of predictor variables.
Y	Vector or matrix of responses.
Xtest	Matrix of predictor variables to predict for.
Prior	List of hyperparameters specifying the parameter prior distributions. If left NULL, a generic set of priors will be generated.
Qs	Upper limit on the number of latent components. If NULL it is chosen automatically.
N_MCMC	Number of iterations to run the Markov chain Monte Carlo algorithm.
BURN	Number of iteration to be discarded as the burn-in.
Thin	Thinning procedure for the MArkov chain. Thin = 1 results in no thinning. Only use for long chains to reduce memory.
model.type	Type of BPLS model to use; one of standard, ss (spike-and-slab), or LASSO (see details).
scale.	Logical; if TRUE then the data variables will be scale to have unit variance.
center.	Logical; if TRUE then the data variables will be zero-centred.
PredInterval	Coverage of prediction intervals if Xtest is provided; 0.95 by default.

Details

The number of latent variables is inferred using the multiplicative gamma process prior (Bhattacharya and Dunson, 2011). Posterior samples from the fitted model are stored as a list. There are three types of parameter prior structures resulting in three different model types:

- **BPLS**: No additional structure assumed; set `model.type=standard`. This model mimics the standard partial least squares regression (PLS; Wold, 1973).
- **ss-BPLS**: A spike-and-slab variant introducing additonal column-wise sparsity to the loading matrix relating to the response variables `Y`; set `model.type=ss`. This approach mimics the Two-way Orthogonal PLS regression (O2PLS; Trygg and Wold, 2003).
- **L-BPLS**: A LASSO variant introducing additonal element-wise sparsity to the loading matrix relating to the response variables `Y`; set `model.type=LASSO`. This approach mimics the sparse PLS regression (sPLS; Chun and Keles, 2010).

Empirical comparisons in Urbas et al. (2024) suggest that the LASSO variant is the best at point predictions and prediction interval coverage when applied to spectral data.

Value

A list of:

`chain`	A Markov chain of samples from the parameter posterior.
`X`	Original set of predictor variables.
`Y`	Original set of response variables.
`Xtest`	Original set of predictor variables to predict from; if `Xtest` is provided.
`Ytest`	Point predictions for new responses; if `Xtest` is provided.
`Ytest_PI`	Prediction intervals for new responses (by default 0.95 coverage); if `Xtest` is provided.
`Ytest_dist`	Posterior predictive distributions for new responses; if `Xtest` is provided.
`diag`	Additional diagnostics for assessing chain convergence.

References

Bhattacharya, A. and Dunson, D. B. (2011) Sparse Bayesian infinite factor models, *Biometrika*, 98(2): 291–306

Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Trygg, J. and Wold, S. (2003). O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, 17(1):53–64.

Urbas, S., Lovera, P., Daly, R., O'Riordan, A., Berry, D., and Gormley, I. C. (2024). "Predicting milk traits from spectral data using Bayesian probabilistic partial least squares regression." *The Annals of Applied Statistics*, 18(4): 3486-3506.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate analysis–III*, pages 383–407. Elsevier.

Examples

```
# data(milk_MIR)
X = milk_MIR$xMIR
Y = milk_MIR$yTraits[, c('Casein_content','Fat_content')]

set.seed(1)
# fit model to 25% of data and predict on remaining 75%
idx = sample(seq(nrow(X)),floor(nrow(X)*0.25),replace = FALSE)

Xtrain = X[idx,];Ytrain = Y[idx,]
Xtest = X[-idx,];Ytest = Y[-idx,]

# fit the model (for default MCMC settings leave Qs and N_MCMC blank; can take longer)
bplsr_Fit = bplsr(Xtrain,Ytrain, Qs = 10, N_MCMC = 5000)

# generate predictions
bplsr_pred = bplsr.predict(model = bplsr_Fit, newdata = Xtest)

# point predictions
head(bplsr_pred$Ytest)

# lower and upper limits of prediction interval
head(bplsr_pred$Ytest_PI)

# plot of predictive posterior distribution for single test sample
hist(bplsr_pred$Ytest_dist[1,'Casein_content',], freq = FALSE,
     main = 'Posterior predictive density', xlab = 'Casein_content')
```

`bplsr.predict` *Predict from a fitted BPLS regression model*

Description

Generates predictions from the fitted BPLS regression model using Monte Carlo simulation.

Usage

```
bplsr.predict(model, newdata, PredInterval = 0.95)
```

Arguments

`model`	Output of `bplsr`.
`newdata`	Matrix of predictor variables to predict for.
`PredInterval`	Intended coverage of prediction intervals (between 0 and 1). Setting the value to 0 only produces point predictions without prediction intervals.

Details

Predictions of the responses are generated from the posterior predictive distribution, marginalising out the model parameters; see Section 3.5 of Urbas et al. (2024).

Value

A list of:

`Ytest`	Point predictions for new responses; if `Xtest` is provided.
`Ytest_PI`	Prediction intervals for new responses (by default 0.95 coverage); if `Xtest` is provided.
`Ytest_dist`	Posterior predictive distributions for new responses; if `Xtest` is provided.

References

Urbas, S., Lovera, P., Daly, R., O'Riordan, A., Berry, D., and Gormley, I. C. (2024). "Predicting milk traits from spectral data using Bayesian probabilistic partial least squares regression." *The Annals of Applied Statistics*, 18(4): 3486-3506.

Examples

```
# data(milk_MIR)
X = milk_MIR$xMIR
Y = milk_MIR$yTraits[, c('Casein_content','Fat_content')]

set.seed(1)
# fit model to 25% of data and predict on remaining 75%
idx = sample(seq(nrow(X)),floor(nrow(X)*0.25),replace = FALSE)

Xtrain = X[idx,];Ytrain = Y[idx,]
Xtest = X[-idx,];Ytest = Y[-idx,]

# fit the model (for default MCMC settings leave Qs and N_MCMC blank; can take longer)
bplsr_Fit = bplsr(Xtrain,Ytrain, Qs = 10, N_MCMC = 5000)

# generate predictions
bplsr_pred = bplsr.predict(model = bplsr_Fit, newdata = Xtest)

# point predictions
head(bplsr_pred$Ytest)

# lower and upper limits of prediction interval
head(bplsr_pred$Ytest_PI)

# plot of predictive posterior distribution for single test sample
hist(bplsr_pred$Ytest_dist[1,'Casein_content',], freq = FALSE,
     main = 'Posterior predictive density', xlab = 'Casein_content')
```

`milk_MIR` *Milk traits and corresponding mid-infrared spectra*

Description

Data containing spectral measurements for 431 milk samples with various chemical and technological traits measures. Details can be found in Visentin et al. (2015) and McDermot et al. (2016).

Usage

```
data(milk_MIR)
```

Format

Data of 431 dairy milk samples with variables split into a list of 3 data frames:

`info` Information variables on the samples.

`xMIR` 531 spectral measurements in the mid-infrared region (predictors). Noisy water-regions have been removed.

`yTraits` 45 chemical and technological traits (responses); contains missing values.

References

McDermott, A., Visentin, G., De Marchi, M., Berry, D., Fenelon, M., O'connor, P., Kenny, O., and McParland, S. (2016). Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. *Journal of Dairy Science, 99(4):3171–3182.*

Visentin, G., McDermott, A., McParland, S., Berry, D., Kenny, O., Brodkorb, A., Fenelon, M., and De Marchi, M. (2015). Prediction of bovine milk technological traits from midinfrared spectroscopy analysis in dairy cows. *Journal of Dairy Science*, 98(9):6620–6629.

Examples

```
data(milk_MIR, package="bplsr")
```

Index