Package 'vagam'

January 9, 2019

Type Package

Title Variational Approximations for Generalized Additive Models

Version 1.0

Date 2019-1-03

Depends R (>= 3.4.0), mgcv, gamm4, Matrix, mvtnorm, truncnorm

LazyLoad yes

LazyData yes

Maintainer Han Lin Shang <hanlin.shang@anu.edu.au>

Description Fits generalized additive models (GAMs) using a variational approximations (VA) framework. In brief, the VA framework provides a fully or at least closed to fully tractable lower bound approximation to the marginal likelihood of a GAM when it is parameterized as a mixed model (using penalized splines, say). In doing so, the VA framework aims offers both the stability and natural inference tools available in the mixed model approach to GAMs, while achieving computation times comparable to that of using the penalized likelihood approach to GAMs. See Hui et al. (2018) <doi:10.1080/01621459.2018.1518235>.

License GPL-3

NeedsCompilation no

Author Han Lin Shang [aut, cre, cph] (<https://orcid.org/0000-0003-1769-6430>), Francis K.C. Hui [aut] (<https://orcid.org/0000-0003-0765-3533>)

Repository CRAN

Date/Publication 2019-01-09 17:00:03 UTC

R topics documented:

`vagam-package` *Variational approximations for generalized additive models*

Description

Fits generalized additive models (GAMs) using a variational approximations (VA) framework. In brief, the VA framework provides a fully or at least closed to fully tractable lower bound approximation to the marginal likelihood of a GAM when it is parameterized as a mixed model (using penalized splines, say). In doing so, the VA framework aims offers both the stability and natural inference tools available in the mixed model approach to GAMs, while achieving computation times comparable to that of using the penalized likelihood approach to GAMs. See Hui et al. (2018) <doi:10.1080/01621459.2018.1518235>.

Author(s)

NA Maintainer: Han Lin Shang <hanlin.shang@anu.edu.au>

References

- Hui, F. K. C., You, C., Shang, H. L., and Mueller, S. (2018). Semiparametric regression using variational approximations, *Journal of the American Statistical Association*, **forthcoming**.

Examples

```
## Please see examples in the help file for the vagam function.
```

`gamsim` *Simulate example datasets from a generalized additive models (GAM).*

Description

This function is a modification from example 7 of the gamSim function available in the mgcv package (Wood, 2017), which is turn is Gu and Wahba 4 univariate example with correlated predictors. Please see the source code for exactly what is simulated. The function is primarily used as the basis for conducting the simulation studies in Hui et al., (2018).

Usage

```
gamsim(n = 400, extra.X = NULL, beta = NULL, dist = "normal", scale = 1, offset = NULL)
```

Arguments

n	Sample size.
extra.X	Extra covariates, including critically an intercept if is to be included in the linear predictor for the GAM.
beta	Regression coefficient estimates.
dist	Currently only the "normal", "poisson" or "binomial" corresponding to the binomial distributions are available.
scale	Scale parameter in the Normal distribution.
offset	This can be used to specify an a-priori known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length equal to n.

Value

A data frame containing information such as the simulated responses, covariates, each of the 4 "truth" smooths, and the overall linear predictor.

Author(s)

NA

References

- Hui, F. K. C., You, C., Shang, H. L., and Mueller, S. (2018). Semiparametric regression using variational approximations, *Journal of the American Statistical Association*, **forthcoming**.
- Wood, S. N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

See Also

vagam for the main fitting function

Examples

```
normal_dat = gamsim(n = 40, dist = "normal",
                    extra.X = data.frame(int = rep(1,40), trt = rep(c(0,1), each = 20)),
                    beta = c(-1, 0.5))

pois_dat = gamsim(n = 40, dist = "poisson",
                  extra.X = data.frame(int = rep(1, 40), trt = rep(c(0,1), each = 20)),
                  beta = c(-1, 0.5))

binom_dat = gamsim(n = 40, dist = "binomial",
                   extra.X = data.frame(int = rep(1, 40), trt = rep(c(0,1), each = 20)),
                   beta = c(0, 0.5))

## Please see examples in the help file for the vagam function.
```

plot.vagam *Basic plots for a fitted generalized additive model (GAMs).*

Description

Takes a fitted vagam object produced by the main vagam function and plots the component smooth functions that make it up, on the scale of the linear predictor.

Usage

```
## S3 method for class 'vagam'
plot(x, n = 100, alpha = 0.05, rug = TRUE, se = TRUE,
xlim = NULL, ylim = NULL, xlab = NULL, ylab = NULL, main = NULL,
select = NULL, ...)
```

Arguments

x	An object of class "vagam".
n	Number of observations used in constructing predictions for plotting.
alpha	Level of significance for the pointwise confidence bands for predictions.
rug	If rug = TRUE, adds a rug representation (1-d plot) of the data to the plot.
se	If se = TRUE, adds lower and upper bounds of 95% pointwise confidence bands.
xlim	Range of plotting for x variable.
ylim	Range of plotting for y variable.
xlab	Label for x variable.
ylab	Label for y variable.
main	Title of plotting.
select	Select which observations to plot.
...	Other plotting arguments.

Details

Currently implements a basic plot for each of the fitted smoothers from a GAM fitted using the main vagam function. This is done by making use of the predict function to construct the fitted smooths. There is also the option of adding pointwise confidence bands based on fitted vagam object. Under the variational approximations framework, the smooths and confidence bands are constructed based on the variational approximation to the posterior distribution of the smoothing coefficients (which are treated as random effects with a normal prior under the mixed model framework). Please see Hui et al., (2018) for more information.

Value

The functions main purpose is its side effect of generating a set of plots.

Author(s)

NA

References

- Hui, F. K. C., You, C., Shang, H. L., and Mueller, S. (2018). Semiparametric regression using variational approximations, *Journal of the American Statistical Association*, **forthcoming**.

See Also

`vagam` for the main fitting function

Examples

```
## Please see examples in the help file for the vagam function.
```

`predict.vagam` *Predictions from a fitted generalized additive model (GAM).*

Description

Takes a fitted vagam object produced by the main vagam function and produces predictions given a new set of values for the model covariates or the original values used for the model fit.

Usage

```
## S3 method for class 'vagam'
predict(object, new.smoothX, new.paraX = NULL, terms = NULL,
alpha = 0.05, type = "link", ...)
```

Arguments

`object`	An object of class "vagam".
`new.smoothX`	A new matrix of covariates, each of which are were entered as additive smooth terms in the fitted GAM.
`new.paraX`	A new matrix of covariates, each of which were be entered as parametric terms in the fitted GAM. Note the predictions will account for the intercept ONLY if new.paraX is supplied.
`terms`	If `terms = NULL`, prediction is made across all smoothing covariates. Else, prediction is made to the specified smoothing covariate, with no intercept added.
`alpha`	Level of significance for the pointwise confidence bands for predictions.
`type`	When `type = "link"` (default) the linear predictor (with associated standard errors) is returned. When `type = "response"` predictions on the scale of the response are returned (with associated standard errors).
`...`	This is currently ignored.

Details

Current implemented a basic method of constructing predictions either for a single smoothing covariate, or across all the smoothing (and parametric if supplied) covariates, based on a GAM fitted using the main vagam function. By default, standard errors and this pointwise confidence bands are also produced based on fitted vagam object. Under the variational approximations framework, the smooths and confidence bands are constructed based on the variational approximation to the posterior distribution of the smoothing coefficients (which are treated as random effects with a normal prior under the mixed model framework). Please see Hui et al., (2018) for more information.

Value

A data frame containing information such as the predicted response, standard errors, and lower and upper bounds of the pointwise confidence bands.

Author(s)

NA

References

- Hui, F. K. C., You, C., Shang, H. L., and Mueller, S. (2018). Semiparametric regression using variational approximations, *Journal of the American Statistical Association*, **forthcoming**.

See Also

`vagam` for the main fitting function

Examples

```
## Please see examples in the help file for the vagam function.
```

`summary.vagam` *Summary of generalized additive model (GAM) fitted using variational approximations (VA).*

Description

A summary of the results from applying `vagam`.

Usage

```
## S3 method for class 'vagam'
summary(object, ...)

## S3 method for class 'vagam'
print(x,...)
```

Arguments

`object`	An object of class "vagam".
`x`	An object of class "vagam".
`...`	Not used.

Value

A list (some of which is printed) containing the following elements:

- call:The matched call.
- para.coeff:The estimated regression coefficients corresponding to the covariates in parametric component of the GAM. This includes the intercept term.
- smooth.coeff:The estimated smoothing coefficients corresponding to the (P-spline bases set up for) covariates in the nonparametric component of the GAM. This corresponds to the mean vector of the variational distribution.
- smooth.param:The estimated smoothing coefficients, or the fixed smoothing parameters if `lambda` was supplied.
- phi:The estimated residual variance when the Gaussian distribution is assumed for the response.
- logLik:The maximized value of the variational log-likelihood.
- family:The assumed distribution for the response.
- smooth.stat:A small table of summary statistics for the nonparametric component of the GAM, including an approximate Wald-type hypothesis test for the significance of each nonparametric covariate.
- para.stat:If `para.se=TRUE`, then a small table containing summary statistics for the estimated parametric component of the GAM, including an approximate Wald-type hypothesis test for the significance of each parameteric covariate.

Author(s)

NA

See Also

`vagam` for the main fitting function

Examples

```
## Please see examples in the help file for the vagam function.
```

vagam *Fitting generalized additive models (GAMs) using variational approximations (VA).*

Description

Follows the variational approximation approach of Hui et al. (2018) for fitted generalized additive models. In this package, the term GAM is taken to be generalized linear mixed model, specifically, the nonparametric component is modeled using a P-splines i.e., cubic B-splines with a first order difference penalty. Because the penalty can be written as a quadratic form in terms of the smoothing coefficients, then it is treated a (degenerate) multivariate normal random effects distribution and a marginal log-likleihood for the resulting mixed model can be constructed.

The VA framework is then utilized to provide a fully or at least closed to fully tractable lower bound approximation to the marginal likelihood of a GAM. In doing so, the VA framework aims offers both the stability and natural inference tools available in the mixed model approach to GAMs, while achieving computation times comparable to that of using the penalized likelihood approach to GAMs.

Usage

```
vagam(y, smooth.X, para.X = NULL, lambda = NULL, int.knots, family = gaussian(),
A.struct = c("unstructured", "block"), offset = NULL, save.data = FALSE,
para.se = FALSE, doIC = FALSE,
control = list(eps = 0.001, maxit = 1000, trace = TRUE, seed.number = 123,
mc.samps = 4000, pois.step.size = 0.01))
```

Arguments

y	A response vector.
smooth.X	A matrix of covariates, each of which are to be entered as additive smooth terms in the GAM.
para.X	An optional matrix of covariates, each of which are to be entered as parametric terms in the GAM. Please note that NO intercept term needs to be included as it is included by default.
lambda	An optional vector of length ncol(smooth.X), where each element corresponds to the smoothing parameter to be applied to the respective covariate in smooth.X. If supplied, then the GAM is fitted with the smoothing parameters held fixed at this values. If lambda=NULL, then smoothing parameters for all covariates to be smoothed are updated automatically as part of the VA algorithm.
int.knots	Either a single number of a vector of length ncol(smooth.X), corresponding to the number of interior knots to be use for the respective covariate iin smooth.X. This argument is passed to the function smooth.construct from the mgcv package (Wood, 2017) in order to construct the P-splines bases. Equally spaced knots based on quantiles are used.

`family` Currently only the `gaussian(link = "identity")`, `poisson(link = "log")`, and `binomial(link = "logit")` corresponding to Bernoulli distributions are available.

`A.struct` The assumed structure of the covariance matrix in the variational distribution of the smoothing coefficients. Currently, the two options are `A.struc = "unstructured"` corresponding to assuming an fully unstructured covariance matrix, and `A.struc = "block"` which assumes a block diagonal structure where the all covariances between different smoothing covariates are assumed to be zero (but the covariance submatrix remains unstructured within the spline basis functions for a selected smoothing covariate). The latter is sub-optimal in the sense that the most appropriate variational distribution should use a completely unstructured covariance matrix , but MAY (but is not guaranteed) save computation time especially when the number of smoothing covariates and/or the number of interior knots is very large.

`offset` This can be used to specify an a-priori known component to be included in the linear predictor during fitting. This should be `NULL` or a numeric vector of length equal to `length(y)`.

`save.data` If `save.date=TRUE`, then the returned vagam object will also include `y, smooth.X, para.X`, and the full matrix of P-spline basis functions.

`para.se` If `para.se=TRUE`, the standard errors based on the VA approach are returned for any covariates in `para.X` that are included as parametric terms. Note that if `para.se=FALSE` then a standard error for the intercept term will not be returned even though an intercept term is included by default.

`doIC` If `doIC=TRUE`, then the AIC and BIC are returned, where the AIC is calculated as $AIC = -2\times$ variation log-likelihood $+2 \times trace(H)$ with $trace(H)$ bring a measure of the degrees of freedom of the model as based on the hat-matrix arising from iterative reweighted least squares, and the BIC replaces the 2 with `log(length(y))` for the model complexity penalty; please see Wood (2017) for more details. Note however that this out is largely mute as the VA approach provides an automatic method of selecting the smoothing parameters, meaning an external approach such as information criteria is not required.

`control` A list controlling the finer details of the VA approach for fitting GAMs. These include:

- mc.samps:This controls Monte Carlo samples for calculating variational observed information matrix using Louis' method
- seed:This controls seed for starting values of the fitting algorithm in general
- pois.step.size:This controls step size for penalized iterative reweighted least squares (P-IRLS) portion of the VA approach when `family=poisson()`. This may be tweaked to use smaller step sizes as the approach here can be a tad unstable especially if there is possible overdispersion.

Details

Please note that the package is still in its early days, and only a very basic form of GAMs with purely additive terms and P-splines is fitted. The function borrows heavily from the excellent software available in the mgcv package (Wood, 2017), in the sense that it uses the `smooth.construct`

function with `bs = "ps"` to set up the matrix of P-splines bases (so cubic B-splines with a first order difference penalty matrix) along with imposing the necessary centering constraints. With these ingredients, it then maximizes the variational log-likelihood by iteratively updating the model and variational parameters. The variational log-likelihood is obtained by proposing a variational distribution for the smoothing coefficients (in this case, a multivariate normal distribution between unknown mean vector and covariance matrix), and then minimizing the Kullback-Leibler distance between this variational distribution and the true posterior distribution of the smoothing coefficients. In turn, this is designed to be (closed to) fully tractable lower bound approximation to the true marginal log-likelihood for the GAM, which for non-normal responses does not possess a tractable form. Note that in contrast to the marginal log-likelihood or many approximations such the Laplace approximation and adaptive quadrature, the variational approximation typically presents a tractable form that is relatively straightforward to maximize. At the same time, because it takes views the GAM as a mixed model, then it also possesses nice inference tools such as an approximate posterior distribution of the smoothing coefficients available immediately from maximizing the VA log-likelihood, and automatic choice of the smoothing parameters. We refer to readers to Wood (2017) and Ruppert et al. (2003) for detailed introductions to GAMs and how many of them can be set up as mixed models; Eilers and Marx (1996) for the seminal text on P-splines, and Hui et al. (2018) for the text on which this package is based.

Value

An object of vagam class containing one or more of the following elements:

- call:The matched call.
- kappa:The estimated regression coefficients corresponding to the covariates in `para.X`. This includes the intercept term.
- a:The estimated smoothing coefficients corresponding to the (P-spline bases set up for) covariates in `smooth.X`. This corresponds to the mean vector of the variational distribution.
- A:The estimated posterior covariance of the smoothing coefficients corresponding to the (P-spline bases set up for) covariates in `smooth.X`. This corresponds to the covariance matrix of the variational distribution.
- lambda:The estimated smoothing parameters, or the fixed smoothing parameters if `lambda` was supplied.
- IC:A vector containing the calculated values of and AIC and BIC if `doIC=TRUE`. Note this is largely a mute output.
- phi:The estimated residual variance when `family=gaussian()`.
- linear.predictors:The estimated linear predictor i.e., the parametric plus nonparametric component.
- logL:The maximized value of the variational log-likelihood.
- no.knots:The number of interior knots used, as per `int.knots`.
- index.cov:A vector indexing which covariate each column in the final full matrix P-spline bases belongs to.
- basis.info:A list with length equal to `ncol(smooth.X)`, with each element being the output from an application of `smooth.construct` to construct the P-spline for a selected covariate in `smooth.X`.

- y, para.X, smooth.X, Z:Returned in `save.data=TRUE`. Note critically that `Z` final full matrix P-spline basis functions.
- smooth.stat:A small table of summary statistics for the nonparametric component of the GAM, including an approximate Wald-type hypothesis test for the significance of each nonparametric covariate.
- para.stat:If `para.se=TRUE`, then a small table containing summary statistics for the estimated parametric component of the GAM, including an approximate Wald-type hypothesis test for the significance of each parameteric covariate.
- obs.info:If `para.se=TRUE`, then the estimated variational observed information matrix for the parameteric component of the GAM; please see Hui et al. (2018) for more information.

Author(s)

NA

References

- Eilers, P. H. C., and Marx, B. D. (1996) Flexible Smoothing with B-splines and Penalties. Statistical Science, 11: 89-121
- Hui, F. K. C., You, C., Shang, H. L., and Mueller, S. (2018). Semiparametric regression using variational approximations, *Journal of the American Statistical Association*, **forthcoming**.
- Ruppert, D., Wand M. P., and Carroll, R. J. (2003) Semiparametric Regression. Cambridge University Press.
- Wood, S. N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.

See Also

`summary.vagam` for a basic summary of the fitted model; `plot.vagam` for basic plotting the component smooths; `predict.vagam` for basic prediction

Examples

```
## Example 1: Application to wage data
data(wage_data)

south_code <- gender_code <- race_code <- union_code <- vector("numeric", nrow(wage_data))
union_code[wage_data$union == "member"] <- 1
south_code[wage_data$south == "yes"] <- 1
gender_code[wage_data$gender == "female"] <- 1
race_code[wage_data$race == "White"] <- 1
para.X <- data.frame(south = south_code, gender = gender_code, race = race_code)

fit_va <- vagam(y = union_code, smooth.X = wage_data[,c("education", "wage", "age")],
                      para.X = para.X,
                      int.knots = 8, save.data = TRUE,
                      family = binomial(),
                      para.se = TRUE)
summary(fit_va)
```

```
a <- 1
par(mfrow = c(1, 3), las = 1, cex = a, cex.lab = a+0.2, cex.main = a+0.5, mar = c(5,5,3,2))
plot(fit_va, ylim = c(-2.7, 2.7), select = 1,
                       xlab = "Education", ylab = "Smooth of Education", lwd = 3)
plot(fit_va, ylim = c(-2.7, 2.7), select = 2,
                    xlab = "Wage", ylab = "Smooth of Wage", main = "Plots from VA-GAM", lwd = 3)
plot(fit_va, ylim = c(-2.7, 2.7), select = 3,
                       xlab = "Age", ylab = "Smooth of Age", lwd = 3)


## Not run:
## Example 2: Simulated data with size = 50 and compare how GAMs can be fitted
## in VA and mgcv (which uses penalized quasi-likelihood)
choose_k <- 5 * ceiling(50^0.2)
true_beta <- c(-1, 0.5)

poisson_dat <- gamsim(n = 50, dist = "poisson", extra.X = data.frame(intercept = rep(1,50),
                       treatment = rep(c(0,1), each = 50/2)), beta = true_beta)

## GAM using VA
fit_va <- vagam(y = poisson_dat$y, smooth.X = poisson_dat[,2:5],
         para.X = data.frame(treatment = poisson_dat$treatment),
                          int.knots = choose_k, save.data = TRUE, family = poisson(),
                          para.se = TRUE)
summary(fit_va)

## GAM using mgcv with default options
fit_mgcv1 <- gam(y ~ treatment + s(x0) + s(x1) + s(x2) + s(x3),
                             data = poisson_dat, family = poisson())


## GAM using mgcv with P-splines and preset knots;
## this is equivalent to VA in terms of the splines bases functions
fit_mgcv2 <- gam(y ~ treatment + s(x0, bs = "ps", k = round(choose_k/2) + 2, m  = c(2,1)) +
                             s(x1, bs = "ps", k = round(choose_k/2) + 2, m  = c(2,1)) +
                             s(x2, bs = "ps", k = round(choose_k/2) + 2, m  = c(2,1)) +
                             s(x3, bs = "ps", k = round(choose_k/2) + 2, m  = c(2,1)),
                             data = poisson_dat, family = poisson())

## End(Not run)
```

---

wage_data *Union membership data set*

Description

1985 North American population survey containing information on union membership and various worker's attributes.

Usage

```
data("wage_data")
```

Format

A data frame with 534 observations on the following 11 variables.

`education` a numeric vector.

`south` a factor with levels `no` `yes`.

`gender` a factor with levels `female` `male`.

`experience` a numeric vector.

`union` a factor with levels `member` `not_member`.

`wage` a numeric vector.

`age` a numeric vector.

`race` a factor with levels `Hispanic` `Other` `White`.

`occupation` a factor with levels `Clerical` `Management` `Other` `Professional` `Sales` `Service`.

`section` a factor with levels `Construction` `Manufacturing` `Other`.

`marital` a factor with levels `Married` `Unmarried`.

Details

The data consist of $n = 534$ observations, with the response being a Bernoulli variable of whether they were a member of union (1 = yes; 0 = no), and six covariates: gender (1 = female, 0 = male), race (1 = white; 0 = other), an indicator variable for whether the worker lives in the south (1 = yes; 0 = no), age in years, hourly wage, and number of years in education.

One of the aims of the survey is to uncover associations between workers' characteristics and their probability of union membership. The dataset is used in Ruppert et al., (2003) and Hui et al. (2018), among others, to illustrate the application of Semiparametric regression, as it is believed that union membership may vary non-linearly with the three continuous variables (age, wage, education).

Source

http://mldata.org/repository/data/viewslug/statlib-20050214-cps_85_wages/

References

- Berndt, E. (1991). The Practice of Econometrics: Classic and Contemporary. Addison-Wesley Publishing Company, Reading, Massachusetts.
- Hui, F. K. C., You, C., Shang, H. L., and Mueller, S. (2018). Semiparametric regression using variational approximations, *Journal of the American Statistical Association*, **forthcoming**.
- Ruppert, D., Wand, M. P., and Carroll, R. (2003). Semiparametric Regression. Cambridge University Press, New York.

Examples

```
data(wage_data)

## Please see examples in the help file for the vagam function.
```

Index