Package 'GPvam'

December 12, 2024

Type Package

Title Maximum Likelihood Estimation of Multiple Membership Mixed Models Used in Value-Added Modeling

Version 3.2-0

Date 2024-12-05

Description An EM algorithm, Karl et al. (2013) <doi:10.1016/j.csda.2012.10.004>, is used to estimate the generalized, variable, and complete persistence models, Mariano et al. (2010) <doi:10.3102/1076998609346967>. These are multiple-membership linear mixed models with teachers modeled as ``G-side'' effects and students modeled with either ``G-side'' or ``R-side'' effects.

Depends R (>= 3.2.0), Matrix

Imports numDeriv, rlang, Rcpp (>= 0.11.2), graphics, grDevices, methods, stats, utils, ggplot2, patchwork, MASS

LinkingTo Rcpp, RcppArmadillo

ByteCompile yes

NeedsCompilation yes

LazyData yes

License GPL-2

Author Andrew Karl [cre, aut] (<https://orcid.org/0000-0002-5933-8706>),
Yan Yang [aut],
Sharon Lohr [aut]

Maintainer Andrew Karl <akarl@asu.edu>

Repository CRAN

Date/Publication 2024-12-12 10:10:06 UTC

Contents

`GPvam-package` *Maximum Likelihood Estimation of Multiple Membership Mixed Models Used in Value-Added Modeling*

Description

An EM algorithm, Karl et al. (2013) <doi:10.1016/j.csda.2012.10.004>, is used to estimate the generalized, variable, and complete persistence models, Mariano et al. (2010) <doi:10.3102/1076998609346967>. These are multiple-membership linear mixed models with teachers modeled as "G-side" effects and students modeled with either "G-side" or "R-side" effects.

Details

Package:	GPvam
Type:	Package
Version:	3.1-1
Date:	2024-10-14
License:	GPL-2

Author(s)

Andrew Karl, Yan Yang, and Sharon Lohr

Maintainer: Andrew Karl <akarl@asu.edu>

References

Karl, A., Yang, Y. and Lohr, S. (2013) Efficient Maximum Likelihood Estimation of Multiple Membership Linear Mixed Models, with an Application to Educational Value-Added Assessments *Computational Statistics & Data Analysis* **59**, 13–27.

Karl, A., Yang, Y. and Lohr, S. (2014) Computation of Maximum Likelihood Estimates for Multiresponse Generalized Linear Mixed Models with Non-nested, Correlated Random Effects *Computational Statistics & Data Analysis* **73**, 146–162.

Karl, A., Yang, Y. and Lohr, S. (2014) A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects *Journal of Educational and Behavioral Statistics* **38**, 577–603.

Karl, A., Zimmerman, D. (2021) A diagnostic for bias in linear mixed model estimators induced by dependence between the random effects and the corresponding model matrix *Journal of Statistical Planning and Inference* **211**, 107–118.

Lockwood, J., McCaffrey, D., Mariano, L., Setodji, C. (2007) Bayesian Methods for Scalable Multivariate Value-Added Assesment. *Journal of Educational and Behavioral Statistics* **32**, 125–150.

Mariano, L., McCaffrey, D. and Lockwood, J. (2010) A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling. *Journal of Educational and Behavioral Statistics* **35**, 253–279.

McCaffrey, D. and Lockwood, J. (2011) Missing Data in Value-Added Modeling of Teavher Effects, *Annals of Applied Statistics* **5**, 773–797

`bias.test.custom` *Permutation Tests for Fixed Effects Bias Assessment*

Description

Performs permutation tests on fixed effects within a linear mixed model to assess the bias of fixed effect parameters or contrasts. The function allows for both standard basis vectors and custom vectors to define the effects being tested. See Karl and Zimmerman (2021) <doi:10.1016/j.jspi.2020.06.004>.

Usage

```
bias.test.custom(result,
                 k_vectors = NULL,
                 n_perms = 1e5)
```

Arguments

`result` An object containing GPvam results, including the fixed effects matrix (`X`), random effects design matrix (`Z`), inverse variance matrix (`vinv`), estimated random effects (`eta.hat`), variance components matrix (`G`), number of teachers per group (`num.teach`), and persistence type (`persistence`). The object must contain the following components:

- `X` Fixed effects matrix.
- `Z` Random effects design matrix.
- `vinv` Inverse variance matrix.
- `eta.hat` Estimated random effects.
- `G` Variance components matrix for random effects.
- `num.teach` Vector indicating the number of teachers (random effects) per group.
- `persistence` Persistence type, must be either `"CP"` or `"VP"` or `"ZP"`.

k_vectors (Optional) A list of numeric vectors specifying custom k vectors for combined fixed effects. Each vector should be the same length as the number of fixed effects in the model. If NULL, the function generates standard basis vectors (one-hot vectors) to perform permutation tests for each fixed effect individually.

n_perms (Optional) The number of permutations to perform for each k vector. A higher number of permutations increases the accuracy of the p-value estimates but also increases computation time. Default is 1e5.

Value

A list containing:

permutation_results
A data frame with the following columns:
- Fixed_Effect Name of the fixed effect or custom contrast tested.
- Nu_Prime_Eta The observed value of $\nu'\hat{\eta}$.
- Permutation_P_Value Permutation p-value for the test of the fixed effect bias.

plot_list A list of ggplot2 objects for the permutation histograms.

Examples

```
## Not run:
# Assuming 'result' is your GPvam object

# Perform bias test for all fixed effects
test_results <- bias.test.custom(result)

# Perform bias test including a custom contrast
k_custom <- c(1, -1, 0, 0)  # Contrast between first and second fixed effects
test_results <- bias.test.custom(result, k_vectors = list(k_custom))

## End(Not run)
```

GP.csh *Internal G-side effects function*

Description

An internal function

Usage

```
GP.csh(Z_mat, fixed_effects, control)
```

Arguments

Z_mat data frame

fixed_effects formula specifying fixed effects to be included in model

control a list

GP.un — *Internal R-side effects function*

Description

An internal function

Usage

```
GP.un(Z_mat, fixed_effects, control)
```

Arguments

Z_mat	data frame
fixed_effects	formula specifying fixed effects to be included in model
control	a list

GPvam — *Fitting the Generalized and Variable Persistence Models*

Description

An EM algorithm, Karl et al. (2013) <doi:10.1016/j.csda.2012.10.004>, is used to estimate the generalized, variable, and complete persistence models, Mariano et al. (2010) <doi:10.3102/1076998609346967>. These are multiple-membership linear mixed models with teachers modeled as "G-side" effects and students modeled with either "G-side" or "R-side" effects.

Usage

```
GPvam(vam_data, fixed_effects = formula(~as.factor(year) + 0),
   student.side = "R", persistence="GP", max.iter.EM = 1000, tol1 = 1e-07,
   hessian = FALSE, hes.method = "simple", REML = FALSE, verbose = TRUE)
```

Arguments

vam_data	a data frame that contains at least a column "y" containing the student scores, a column "student" containing unique student ID's, a column "teacher" containing the teacher ID's, and a column "year" which contains the year (or semester, etc.) of the time period. The "y" and "year" variables needs to be numeric. If other variables are to be included as fixed effects, they should also be included in vam_data. See 'Note' for further discussion.
fixed_effects	an object of class `formula` describing the structure of the fixed effects. Categorical variables should be wrapped in an `as.factor` statement.
student.side	a character. Choices are "G" or "R". See section 'Details'.

`persistence`	a character. Choices are "GP", "rGP", "VP", "CP", or "ZP". Only "GP" is currently compatible with student.side="G". See section 'Details'.
`max.iter.EM`	the maximum number of EM iterations
`tol1`	convergence tolerance for EM algorithm. The convergence criterion is specified under 'Details'.
`hessian`	logical indicating whether the Hessian of the variance parameters (and persistence parameters for persistence="VP") should be calculated after convergence of the EM algorithm. Standard errors for the fixed and EBLUPs are calculated by default.
`hes.method`	a character string indicating the method of numerical differentiation used to calculate the Hessian of the variance parameters. Options are "simple" or "richardson".
`REML`	logical indicating whether REML estimation should be used instead of ML estimation. Only currently compatible with persistence = CP, VP, or ZP.
`verbose`	logical. If TRUE, model information will be printed at each iteration.

Details

The design for the random teacher effects according to the generalized persistence model of Mariano et al. (2010) is built into the function. The model includes correlated current- and future-year effects for each teacher. By setting `student.side="R"`, the intra-student correlation is modeled via an unstructured, block-diagonal error covariance matrix, as specified by Mariano et al. (2010). Setting `student.side="G"` keeps the same teacher structure, but models intra-student correlation via random student effects. This is similar to the model used by McCaffrey and Lockwood (2011), and is appropriate when the testing scale is the same across years. In this case, the error covariance matrix is diagonal, although a separate variance is calculated for each year. From a computational perspective, the model estimating the R-side student effects has better scalability properties, although the G-side function is faster (Karl et al. 2012).

The `persistence` option determines the type of persistence effects that are modeled. The generalized persistence model ("GP") is described above. When `student.side="R"`, other models for teacher persistence are available. The reduced GP model ("rGP", Karl et al. 2012) combines each teacher's future year effects from the GP model into a single effect. The variable persistence model ("VP") assumes that teacher effects in future years are multiples of their effect in the current year (Lockwood et al. 2007). The multipliers in the VP model are called persistence parameters, and are estimated. By contrast, the complete ("CP") and zero ("ZP") persistence models fix the persistence parameters at 1 and 0, respectively (Lockwood et al. 2007).

Convergence is declared when $(l_k - l_{k-1})/l_k < 1E-07$, where l_k is the log-likelihood at iteration k.

The model is estimated via an EM algorithm. For details, see Karl et al. (2012). The model was estimated through Bayesian computation in Mariano et al. (2010).

Note: When student.side="R" is selected, the first few iterations of the EM algorithm will take longer than subsequent iterations. This is a result of the hybrid gradient-ascent/Newton-Raphson method used in the M-step for the R matrix in the first two iterations (Karl et al. 2012).

Program run time and memory requirements: The data file GPvam.benchmark that is included with the package contains runtime and peak memory requirements for different persistence settings, using simulated data sets with different values for number of years, number of teachers per year, and number of students per teacher. These have been multiplied to show the total number of teachers in the data set, as well as the total number of students. With student.side="R", the persistence="GP" model is most sensitive to increases in the size of the data set. With student.side="G", the memory requirements increase exponentially with the number of students and teachers, and that model should not be considered scalable to extremely large data sets.

All of these benchmarks were performed with Hessian=TRUE. Calculation of the Hessian accounts for anywhere from 20% to 75% of those run times. Unless the standard errors of the variance components are needed, leaving Hessian=FALSE will lead to a faster run time with smaller memory requirements.

Value

GPvam returns an object of class `GPvam`

An object of class `GPvam` is a list containing the following components:

`loglik`	the maximized log-likelihood at convergence of the EM algorithm
`teach.effects`	a data frame containing the predicted teacher effects and standard errors
`parameters`	a matrix of estimated model parameters and standard errors
`Hessian`	if requested, the Hessian of the variance parameters
`R_i`	(only when `student_side` is set to 'R') a matrix containing the error covariance matrix of a student
`teach.cov`	a list containing the unique blocks of the covariance matrix of teacher effects
`mresid`	a vector of the raw marginal residuals
`cresid`	a vector of the raw conditional residuals
`sresid`	a vector of the scaled conditional residuals
`yhat`	a vector of the predicted values

The function `summary` provides a summary of the results. This includes the estimated model parameters and standard errors, along with the correlation matrices corresponding to the estimated correlation matrices. Summary information about scaled and raw residuals is reported.

Note

The model assumes that each teacher teaches only one year. If, for example, a teacher teaches in years 1 and 2, his/her first year performance is modeled independently of the second year performance. To keep these effects separate, the progam appends "(year i)" to each teacher name, where i is the year in which the teacher taught.

The `fixed_effects` argument of `GPvam` utilizes the functionality of R's `formula` class. In the statement `fixed_effects=formula(~as.factor(year)+cont_var+0))`, `as.factor(year)` identifies `year` as a categorical variable. `+0` indicates that no intercept is to be fitted, and `+cont_var` indicates that a seperate effect is to be fitted for the continuous variable "cont_var." An interaction between "year" and "cont_var" could be specified by `~as.factor(year)*cont_var+0`, or equivalently, `~as.factor(year)+cont_var+as.factor(year):cont_var+0`. See `formula` for more details.

When applied to an object of class GPvam, `plot.GPvam` returns a caterpillar plot for each effect, as well as residual plots.

Author(s)

Andrew Karl <akarl@asu.edu>, Yan Yang, Sharon Lohr

References

Karl, A., Yang, Y. and Lohr, S. (2013) Efficient Maximum Likelihood Estimation of Multiple Membership Linear Mixed Models, with an Application to Educational Value-Added Assessments *Computational Statistics & Data Analysis* **59**, 13–27.

Karl, A., Yang, Y. and Lohr, S. (2014) Computation of Maximum Likelihood Estimates for Multiresponse Generalized Linear Mixed Models with Non-nested, Correlated Random Effects *Computational Statistics & Data Analysis* **73**, 146–162.

Karl, A., Yang, Y. and Lohr, S. (2014) A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects *Journal of Educational and Behavioral Statistics* **38**, 577–603.

Lockwood, J., McCaffrey, D., Mariano, L., Setodji, C. (2007) Bayesian Methods for Scalable Multivariate Value-Added Assesment. *Journal of Educational and Behavioral Statistics* **32**, 125–150.

Mariano, L., McCaffrey, D. and Lockwood, J. (2010) A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling. *Journal of Educational and Behavioral Statistics* **35**, 253–279.

McCaffrey, D. and Lockwood, J. (2011) Missing Data in Value-Added Modeling of Teacher Effects," *Annals of Applied Statistics* **5**, 773–797

See Also

plot.GPvam, summary.GPvam, vam_data

Examples

```
data(vam_data)
GPvam(vam_data,student.side="R",persistence="CP",
fixed_effects=formula(~as.factor(year)+cont_var+0),verbose=TRUE,max.iter.EM=1)

result <- GPvam(vam_data,student.side="R",persistence="VP",
fixed_effects=formula(~as.factor(year)+cont_var+0),verbose=TRUE)
 summary(result)

 plot(result)
```

GPvam.benchmark *Benchmarks of the program using simulated data.*

Description

The data file GPvam.benchmark that is included with the package contains runtime and peak memory requirements for different persistence settings, using simulated data sets with different values for number of years, number of teachers per year, and number of students per teacher. These have been multiplied to show the total number of teachers in the data set, as well as the total number of students. With student.side="R", the persistence="GP" model is most sensitive to increases in the size of the data set. With student.side="G", the memory requirements increase exponentially with the number of students and teachers, and that model should not be considered scalable to extremely large data sets.

All of these benchmarks were performed with Hessian=TRUE. Calculation of the Hessian accounts for anywhere from 20% to 75% of those run times. Unless the standard errors of the variance components are needed, leaving Hessian=FALSE will lead to a faster run time with smaller memory requirements.

Usage

```
data(vam_data)
```

Examples

```
data(GPvam.benchmark)
print(GPvam.benchmark[1,])
```

---

plot *Plot method for GPvam*

Description

Plot teacher effects and residuals. The caterpillar plots use a modified version of the plotCI function from R package gplots. According to that package, "Original version [of plotCI] by Bill Venables wvenable@attunga.stats.adelaide.edu.au posted to r-help on Sep. 20, 1997. Enhanced version posted to r-help by Ben Bolker ben@zoo.ufl.edu on Apr. 16, 2001. This version was modified and extended by Gregory R. Warnes greg@warnes.net. Additional changes suggested by Martin Maechler maechler@stat.math.ethz.ch integrated on July 29, 2004."

Usage

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

Arguments

x	an object of class GPvam
...	other arguments
alpha	the significance level for the caterpillar plots

Value

Requires user to click window or press "enter" to progress through plots. Returns caterpillar plots (via the package gplots) and residual plots.

Author(s)

Andrew Karl <akarl@asu.edu> Yan Yang Sharon Lohr

Other authors as listed above for the caterpillar plots.

References

Karl, A., Yang, Y. and Lohr, S. (2013) Efficient Maximum Likelihood Estimation of Multiple Membership Linear Mixed Models, with an Application to Educational Value-Added Assessments *Computational Statistics & Data Analysis* **59**, 13–27.

Karl, A., Yang, Y. and Lohr, S. (2014) Computation of Maximum Likelihood Estimates for Multiresponse Generalized Linear Mixed Models with Non-nested, Correlated Random Effects *Computational Statistics & Data Analysis* **73**, 146–162.

Karl, A., Yang, Y. and Lohr, S. (2014) A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects *Journal of Educational and Behavioral Statistics* **38**, 577–603.

Lockwood, J., McCaffrey, D., Mariano, L., Setodji, C. (2007) Bayesian Methods for Scalable Multivariate Value-Added Assesment. *Journal of Educational and Behavioral Statistics* **32**, 125–150.

Mariano, L., McCaffrey, D. and Lockwood, J. (2010) A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling. *Journal of Educational and Behavioral Statistics* **35**, 253–279.

McCaffrey, D. and Lockwood, J. (2011) Missing Data in Value-Added Modeling of Teavher Effects, *Annals of Applied Statistics* **5**, 773–797

See Also

summary.GPvam

Examples

```
data(vam_data)

GPvam(vam_data,student.side="R",persistence="VP",
fixed_effects=formula(~as.factor(year)+cont_var+0),verbose=TRUE,max.iter.EM=1)

result <- GPvam(vam_data,student.side="R",persistence="VP",
fixed_effects=formula(~as.factor(year)+cont_var+0),verbose=TRUE)
 summary(result)

 plot(result)
```

print *Print*

Description

Prints names of elements in GPvam object.

Usage

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

Arguments

x	object of class GPvam
...	other arguments to be passed to summary

rGP.un *Internal R-side effects function for reduced GP model*

Description

An internal function

Usage

```
rGP.un(Z_mat, fixed_effects, control)
```

Arguments

Z_mat	data frame
fixed_effects	formla specifying fixed effects to be included in model
control	a list

summary *Summary*

Description

Prints summary information for object of class GPvam

Usage

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

Arguments

object	object of class GPvam
...	other arguments to be passed to summary

Author(s)

Andrew Karl <akarl@asu.edu> Yan Yang Sharon Lohr

See Also

plot.GPvam

Examples

```
## Not run:
data(vam_data)
result<-GPvam(vam_data)
summary(result)

## End(Not run)
```

vam_data *Simulated Data*

Description

A simulated data set used to illustrate the functionality of the package. The data are simulated according to the VP model, and demonstrate the stability of the program in the presence of perfectly correlated future year effects.

Usage

```
data(vam_data)
```

Format

A data frame with 3750 observations on 1250 students over 3 years, with 50 teachers in each year. The data set contains the following 5 variables.

- `y` a numeric vector representing the student score
- `student` a numeric vector
- `year` a numeric vector
- `teacher` a numeric vector
- `cont_var` a numeric vector representing a continuous covariate

Details

The data set may be reproduced with the following code.

```
set.seed(0)
years<-3
#teacher in each year
teachers<-50
#students in each class
students<-25
alpha<-.4
eta.stu<-rnorm(students*teachers,0,5)
z1<-rep(1:teachers,each=students)
z2<-sample(rep(1:teachers,each=students))
z3<-sample(rep(1:teachers,each=students))
cont_var1<-rnorm(students*teachers,0,4)
cont_var2<-rnorm(students*teachers,0,4)
cont_var3<-rnorm(students*teachers,0,4)
gam1<- rnorm(teachers,0,5)
gam2<- rnorm(teachers,0,5)
gam3<- rnorm(teachers,0,5)
eps1<- rnorm(students*teachers,0,5)
eps2<- rnorm(students*teachers,0,5)
eps3<- rnorm(students*teachers,0,5)
y1<-eta.stu+gam1[z1]+cont_var1+eps1
y2<-eta.stu+gam1[z1]*alpha+gam2[z2]+cont_var2+eps2
y3<-eta.stu+gam1[z1]*alpha+gam2[z2]*alpha+gam3[z3]+cont_var3+eps3
student<-1:(students*teachers)
teacher<-c(z1,z2,z3)
cont_var<-c(cont_var1,cont_var2,cont_var3)
year<-c(rep(1:3,each=students*teachers))
y<-c(y1,y2,y3)
vam_data<-as.data.frame(cbind(student,teacher,year,y,cont_var))
```

Examples

```
data(vam_data)
print(vam_data[1,])
```

`VP.CP.ZP.un` *Internal R-side effects function for the variable persistence model.*

Description

An internal function

Usage

```
VP.CP.ZP.un(Z_mat, fixed_effects, control)
```

Arguments

`Z_mat`	data frame
`fixed_effects`	formula specifying fixed effects to be included in model
`control`	a list

Index