# Package ‘PAGFL’

November 17, 2025

**Title** Joint Estimation of Latent Groups and Group-Specific Coefficients in (Time-Varying) Panel Data Models

**Version** 1.1.4

**Maintainer** Paul Haimerl <paul.haimerl@econ.au.dk>

**Description** Latent group structures are a common challenge in panel data analysis. Disregarding group-level heterogeneity can introduce bias. Conversely, estimating individual coefficients for each cross-sectional unit is inefficient and may lead to high uncertainty.
This package addresses the issue of unobservable group structures by implementing the pairwise adaptive group fused Lasso (PAGFL) by Mehrabani (2023) <doi:10.1016/j.jeconom.2022.12.002>. PAGFL identifies latent group structures and group-specific coefficients in a single step.
On top of that, we extend the PAGFL to time-varying coefficient functions (FUSE-TIME), following Haimerl et al. (2025) <doi:10.48550/arXiv.2503.23165>.

**Depends** R (>= 3.5.0)

**License** AGPL (>= 3)

**Encoding** UTF-8

**RoxygenNote** 7.3.2

**LinkingTo** Rcpp, RcppArmadillo, RcppParallel, RcppThread

**Imports** Rcpp, lifecycle, ggplot2, RcppParallel

**BugReports** `https://github.com/Paul-Haimerl/PAGFL/issues`

**URL** `https://github.com/Paul-Haimerl/PAGFL`

**Suggests** testthat (>= 3.0.0)

**Config/testthat/edition** 3

**NeedsCompilation** yes

**Author** Paul Haimerl [aut, cre] (ORCID: <https://orcid.org/0000-0003-3198-8317>),
Stephan Smeekes [ctb] (ORCID: <https://orcid.org/0000-0002-0157-639X>),
Ines Wilms [ctb] (ORCID: <https://orcid.org/0000-0003-3269-4601>),
Ali Mehrabani [ctb] (ORCID: <https://orcid.org/0000-0002-1848-5582>)

**Repository** CRAN

**Date/Publication** 2025-11-17 21:51:43 UTC

# Contents



---

`fuse_time` *Fused Unobserved group Spline Estimation of TIME varying coefficients*

---

## Description

Estimate a time-varying panel data model subject to a latent group structure using *FUSE-TIME*–Fused Unobserved group Spline Estimation of TIME varying coefficients–by Haimerl et al. (2025). *FUSE-TIME* jointly identifies the latent group structure and group-specific time-varying functional coefficients. The time-varying coefficients are approximated as polynomial B-splines. The function supports both static and dynamic panel data models.

## Usage

```
fuse_time(
  formula,
  data,
  index = NULL,
  n_periods = NULL,
  lambda,
  d = 3,
  M = floor(length(y)^(1/7) - log(p)),
  min_group_frac = 0.05,
  const_coef = NULL,
  kappa = 2,
  max_iter = 50000,
  tol_convergence = 1e-10,
  tol_group = 0.001,
  rho = 0.04 * log(N * n_periods)/sqrt(N * n_periods),
  varrho = 1,
  verbose = TRUE,
  parallel = TRUE,
  ...
)

tv_pagfl(
  formula,
```

```
  data,
  index = NULL,
  n_periods = NULL,
  lambda,
  d = 3,
  M,
  min_group_frac = 0.05,
  const_coef = NULL,
  kappa = 2,
  max_iter = 50000,
  tol_convergence = 1e-10,
  tol_group = 0.001,
  rho,
  varrho = 1,
  verbose = TRUE,
  parallel = TRUE,
  ...
)

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

## S3 method for class 'fusetime'
formula(x, ...)

## S3 method for class 'fusetime'
df.residual(object, ...)

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

## S3 method for class 'fusetime'
coef(object, ...)

## S3 method for class 'fusetime'
residuals(object, ...)

## S3 method for class 'fusetime'
fitted(object, ...)
```

## Arguments

`formula` a formula object describing the model to be estimated.

`data` a `data.frame` or `matrix` holding a panel data set. If no `index` variables are provided, the panel must be balanced and ordered in the long format $\boldsymbol{Y} = (Y_1', \dots, Y_N')'$, $Y_i = (Y_{i1}, \dots, Y_{iT})'$ with $Y_{it} = (y_{it}, \boldsymbol{x}_{it}')'$. Conversely, if `data` is not ordered or not balanced, `data` must include two index variables that declare the cross-sectional unit $i$ and the time period $t$ of each observation.

| | |
|---|---|
| `index` | a character vector holding two strings. The first string denotes the name of the index variable identifying the cross-sectional unit $i$ and the second string represents the name of the variable declaring the time period $t$. The data is automatically sorted according to the variables in `index`, which may produce errors when the time index is a character variable. In case of a balanced panel data set that is ordered in the long format, `index` can be left empty if the number of time periods `n_periods` is supplied. |
| `n_periods` | the number of observed time periods $T$. If an `index` character vector is passed, this argument can be left empty. Default is `NULL`. |
| `lambda` | the tuning parameter determining the strength of the penalty term. Either a single $\lambda$ or a vector of candidate values can be passed. If a vector is supplied, a BIC-type IC automatically selects the best fitting $\lambda$ value. |
| `d` | the polynomial degree of the B-splines. Default is 3. |
| `M` | the number of interior knots of the B-splines. If left unspecified, the default heuristic $M = \text{floor}((NT)^{\frac{1}{7}} - \log(p))$ following Haimerl et al. (2025) is used. |
| `min_group_frac` | the minimum group cardinality as a fraction of the total number of individuals $N$. In case a group falls short of this threshold, each of its members is allocated to one of the remaining groups according to the *MSE*. Default is 0.05. |
| `const_coef` | a character vector containing the variable names of explanatory variables that enter with time-constant coefficients. |
| `kappa` | the a non-negative weight used to obtain the adaptive penalty weights. Default is 2. |
| `max_iter` | the maximum number of iterations for the *ADMM* estimation algorithm. Default is $5 * 10^4$. |
| `tol_convergence` | the tolerance limit for the stopping criterion of the iterative *ADMM* estimation algorithm. Default is $1 * 10^{-10}$. |
| `tol_group` | the tolerance limit for within-group differences. Two individuals are assigned to the same group if the Frobenius norm of their coefficient vector difference is below this threshold. Default is $1 * 10^{-3}$. |
| `rho` | the tuning parameter balancing the fitness and penalty terms in the IC that determines the penalty parameter $\lambda$. If left unspecified, the heuristic $\rho = 0.07 \frac{\log(NT)}{\sqrt{NT}}$ of Haimerl et al. (2025) is used. We recommend the default. |
| `varrho` | the non-negative Lagrangian *ADMM* penalty parameter. For the employed penalized sieve estimation *PSE*, the $\varrho$ value is not very influential. We recommend the default 1. |
| `verbose` | logical. If `TRUE`, helpful warning messages are shown. Default is `TRUE`. |
| `parallel` | logical. If `TRUE`, certain operations are parallelized across multiple cores. Default is `TRUE`. |
| `...` | ellipsis |
| `object` | of class `fusetime`. |
| `x` | of class `fusetime`. |

### Details

Consider the grouped time-varying panel data model subject to a latent group structure

$$y_{it} = \gamma_i^0 + \boldsymbol{\beta}_i^{0\prime}(t/T)\boldsymbol{x}_{it} + \epsilon_{it}, \quad i = 1, \dots, N, \; t = 1, \dots, T,$$

where $y_{it}$ is the scalar dependent variable, $\gamma_i^0$ is an individual fixed effect, $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of explanatory variables, and $\epsilon_{it}$ denotes a zero mean error. The $p$-dimensional coefficient vector $\boldsymbol{\beta}_i^0(t/T)$ contains smooth functions of time and follows the latent group pattern

$$\boldsymbol{\beta}_i^0\left(\frac{t}{T}\right) = \sum_{k=1}^{K} \boldsymbol{\alpha}_k^0\left(\frac{t}{T}\right) \mathbf{1}\{i \in G_k^0\},$$

with $\cup_{k=1}^K G_k^0 = \{1, \dots, N\}$, $G_k^0 \cap G_j^0 = \emptyset$ for any $k \neq j$, $k, j = 1, \dots, K$.

The time-varying coefficient functions are estimated as polynomial B-splines. To this end, let $\boldsymbol{b}(v)$ denote a $M+d+1$ vector of polynomial basis functions with the polynomial degree $d$ and $M$ interior knots. Then, $\boldsymbol{\beta}_i^0(t/T)$ is approximated by forming linear combinations of these basis functions $\boldsymbol{\beta}_i^0(t/T) \approx \boldsymbol{\Pi}_i^{0\prime}\boldsymbol{b}(t/T)$, where $\boldsymbol{\Pi}_i^0$ is a $(M+d+1) \times p$ matrix of spline control points.

To estimate $\boldsymbol{\Pi}_i^0$, we project the explanatory variables onto the spline basis system, resulting in the $(M+d+1)p \times 1$ regressor vector $\boldsymbol{z}_{it} = \boldsymbol{x}_{it} \otimes \boldsymbol{b}(v)$. Subsequently, the DGP can be reformulated as

$$y_{it} = \gamma_i^0 + \boldsymbol{\pi}_i^{0\prime}\boldsymbol{z}_{it} + u_{it},$$

where $\boldsymbol{\pi}_i^0 = \mathrm{vec}(\boldsymbol{\Pi}_i^0)$, and $u_{it} = \epsilon_{it} + \eta_{it}$ collects the idiosyncratic $\epsilon_{it}$ and the sieve approximation error $\eta_{it}$.

Following Haimerl et al. (2025, sec. 2), *FUSE-TIME* jointly estimates the functional coefficients and the group structure by minimizing the criterion

$$F_{NT}(\boldsymbol{\pi}, \lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \boldsymbol{\pi}_i'\tilde{\boldsymbol{z}}_{it})^2 + \frac{\lambda}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\dot{\omega}_{ij}\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\|_2$$

with respect to $\boldsymbol{\pi} = (\boldsymbol{\pi}_i', \dots, \boldsymbol{\pi}_N')'$. $\tilde{a}_{it} = a_{it} - T^{-1}\sum_{t=1}^{T} a_{it}$, $a = \{y, \boldsymbol{z}\}$ to concentrate out the individual fixed effects $\gamma_i^0$ (within-transformation). $\lambda$ is the penalty tuning parameter and $\dot{w}_{ij}$ denotes adaptive penalty weights which are obtained by a preliminary non-penalized estimation. The criterion function is minimized via an iterative alternating direction method of multipliers (*ADMM*) algorithm (Haimerl et al. 2053, Appendix C).

Two individuals are assigned to the same group if $\|\hat{\boldsymbol{\pi}}_i - \hat{\boldsymbol{\pi}}_j\|_2 \leq \epsilon_{\text{tol}}$ (and hence $\hat{\boldsymbol{\xi}}_k = \hat{\boldsymbol{\pi}}_i = \hat{\boldsymbol{\pi}}_j$ for some $k = 1, \dots, \hat{K}$), where $\epsilon_{\text{tol}}$ is determined by `tol_group`. The time-varying coefficients are then retrieved by taking $\hat{\boldsymbol{\beta}}_i(t/T) = \hat{\boldsymbol{\Pi}}_i'\boldsymbol{b}(t/T)$, where $\hat{\boldsymbol{\pi}}_i = \mathrm{vec}(\hat{\boldsymbol{\Pi}}_i)$ (analogously $\hat{\boldsymbol{\alpha}}_k(t/T) = \hat{\boldsymbol{\Xi}}_k'\boldsymbol{b}(t/T)$, using $\hat{\boldsymbol{\xi}}_k = \mathrm{vec}(\hat{\boldsymbol{\Xi}}_k)$).

Subsequently, the estimated number of groups $\hat{K}$ and group structure follow by examining the number of distinct elements in $\hat{\boldsymbol{\pi}}$. Given an estimated group structure, it is straightforward to obtain post-Lasso estimates $\hat{\boldsymbol{\alpha}}_k^p(t/T) = \hat{\boldsymbol{\Xi}}_k^{p\prime}\boldsymbol{b}(t/T)$ for each $k = 1, \dots, \hat{K}$ using group-wise least squares (see `grouped_tv_plm`).

We recommend choosing a $\lambda$ tuning parameter by passing a logarithmically spaced grid of candidate values with a lower limit close to 0 and an upper limit that leads to a fully homogeneous panel. A BIC-type information criterion then automatically selects the best fitting $\lambda$ value.

In case of an unbalanced panel data set, the earliest and latest available observations per group define the start and end-points of the interval on which the group-specific time-varying coefficients are defined.

We refer to Haimerl et al. (2025) for more details.

## Value

An object of class `fusetime` holding

| | |
|---|---|
| `model` | a `data.frame` containing the dependent and explanatory variables as well as cross-sectional and time indices, |
| `coefficients` | let $p^{(1)}$ denote the number of time-varying coefficients and $p^{(2)}$ the number of time constant parameters. A `list` holding (i) a $T \times p^{(1)} \times \hat{K}$ array of the post-Lasso group-specific functional coefficients and (ii) a $K \times p^{(2)}$ matrix of time-constant post-Lasso estimates. |
| `groups` | a `list` containing (i) the total number of groups $\hat{K}$ and (ii) a vector of estimated group memberships $(\hat{g}_1, \dots, \hat{g}_N)$, where $\hat{g}_i = k$ if $i$ is assigned to group $k$, |
| `residuals` | a vector of residuals of the demeaned model, |
| `fitted` | a vector of fitted values of the demeaned model, |
| `args` | a `list` of additional arguments, |
| `IC` | a `list` containing (i) the value of the IC, (ii) the employed tuning parameter $\lambda$, and (iii) the *MSE*, |
| `convergence` | a `list` containing (i) a logical variable if convergence was achieved and (ii) the number of executed *ADMM* algorithm iterations, |
| `call` | the function call. |

An object of class `fusetime` has `print`, `summary`, `fitted`, `residuals`, `formula`, `df.residual`, and `coef` S3 methods.

## Author(s)

Paul Haimerl

## References

Haimerl, P., Smeekes, S., & Wilms, I. (2025). Estimation of latent group structures in time-varying panel data models. *arXiv preprint arXiv:2503.23165*. doi:10.48550/arXiv.2503.23165.

## Examples

```
# Simulate a time-varying panel with a trend and a group pattern
set.seed(1)
sim <- sim_tv_DGP(N = 10, n_periods = 50, intercept = TRUE, p = 1)
df <- data.frame(y = c(sim$y))

# Run FUSE-TIME
estim <- fuse_time(y ~ ., data = df, n_periods = 50, lambda = 10, parallel = FALSE)
summary(estim)
```

grouped_plm *Grouped Panel Data Model*

## Description

Estimate a panel data model subject to an observed group structure. Slope parameters are homogeneous within groups but heterogeneous across groups. This function supports both static and dynamic panel data models, with or without endogenous regressors.

## Usage

```
grouped_plm(
  formula,
  data,
  groups,
  index = NULL,
  n_periods = NULL,
  method = "PLS",
  Z = NULL,
  bias_correc = FALSE,
  rho = 0.07 * log(N * n_periods)/sqrt(N * n_periods),
  verbose = TRUE,
  parallel = TRUE,
  ...
)

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

## S3 method for class 'gplm'
formula(x, ...)

## S3 method for class 'gplm'
df.residual(object, ...)

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

## S3 method for class 'gplm'
coef(object, ...)

## S3 method for class 'gplm'
residuals(object, ...)

## S3 method for class 'gplm'
fitted(object, ...)
```

### Arguments

`formula` a formula object describing the model to be estimated.

`data` a `data.frame` or `matrix` holding a panel data set. If no `index` variables are provided, the panel must be balanced and ordered in the long format $\boldsymbol{Y} = (Y_1', \dots, Y_N')'$, $Y_i = (Y_{i1}, \dots, Y_{iT})'$ with $Y_{it} = (y_{it}, x_{it}')'$. Conversely, if `data` is not ordered or not balanced, `data` must include two index variables that declare the cross-sectional unit $i$ and the time period $t$ of each observation.

`groups` a numerical or character vector of length $N$ that indicates the group membership of each cross-sectional unit $i$.

`index` a character vector holding two strings. The first string denotes the name of the index variable identifying the cross-sectional unit $i$ and the second string represents the name of the variable declaring the time period $t$. The data is automatically sorted according to the variables in `index`, which may produce errors when the time index is a character variable. In case of a balanced panel data set that is ordered in the long format, `index` can be left empty if the number of time periods `n_periods` is supplied.

`n_periods` the number of observed time periods $T$. If an `index` is passed, this argument can be left empty. Default is `NULL`.

`method` the estimation method. Options are

- `"PLS"` for using the penalized least squares (*PLS*) algorithm. We recommend *PLS* in case of (weakly) exogenous regressors (Mehrabani, 2023, sec. 2.2).
- `"PGMM"` for using the penalized Generalized Method of Moments (*PGMM*). *PGMM* is required when instrumenting endogenous regressors, in which case a matrix $\boldsymbol{Z}$ containing the necessary exogenous instruments must be supplied (Mehrabani, 2023, sec. 2.3).

Default is `"PLS"`.

`Z` a $NT \times q$ `matrix` or `data.frame` of exogenous instruments, where $q \geq p$, $\boldsymbol{Z} = (z_1', \dots, z_N')'$, $z_i = (z_{i1}, \dots, z_{iT})'$ and $z_{it}$ is a $q \times 1$ vector. `Z` is only required when `method = "PGMM"` is selected. When using `"PLS"`, the argument can be left empty or it is disregarded. Default is `NULL`.

`bias_correc` logical. If `TRUE`, a Split-panel Jackknife bias correction following Dhaene and Jochmans (2015) is applied to the slope parameters. We recommend using the correction when working with dynamic panels. Default is `FALSE`.

`rho` a tuning parameter balancing the fitness and penalty terms in the IC. If left unspecified, the heuristic $\rho = 0.07 \frac{\log(NT)}{\sqrt{NT}}$ of Mehrabani (2023, sec. 6) is used. We recommend the default.

`verbose` logical. If `TRUE`, helpful warning messages are shown. Default is `TRUE`.

`parallel` logical. If `TRUE`, certain operations are parallelized across multiple cores. Default is `TRUE`.

`...` ellipsis

`x` of class `gplm`.

`object` of class `gplm`.

## Details

Consider the grouped panel data model

$$y_{it} = \gamma_i^0 + \boldsymbol{\beta}_i^{0\prime} \boldsymbol{x}_{it} + \epsilon_{it}, \quad i = 1, \dots, N, \ t = 1, \dots, T,$$

where $y_{it}$ is the scalar dependent variable, $\gamma_i^0$ is an individual fixed effect, $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of (weakly) exogenous explanatory variables, and $\epsilon_{it}$ denotes a zero mean error. The coefficient vector $\boldsymbol{\beta}_i^0$ follows the observed group pattern

$$\boldsymbol{\beta}_i^0 = \sum_{k=1}^{K} \boldsymbol{\alpha}_k^0 \mathbf{1}\{i \in G_k\},$$

with $\cup_{k=1}^{K} G_k = \{1, \dots, N\}$, $G_k \cap G_j = \emptyset$ and $\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_j^0\| \neq 0$ for any $k \neq j$, $k, j = 1, \dots, K$. The group structure $G_1, \dots, G_K$ is determined by the argument `groups`.

Using *PLS*, the group-specific coefficients of group $k$, $k = 1, \dots, K$, are obtained by *OLS*

$$\hat{\boldsymbol{\alpha}}_k = \left( \sum_{i \in G_k} \sum_{t=1}^{T} \tilde{\boldsymbol{x}}_{it} \tilde{\boldsymbol{x}}_{it}' \right)^{-1} \sum_{i \in G_k} \sum_{t=1}^{T} \tilde{\boldsymbol{x}}_{it} \tilde{y}_{it},$$

where $\tilde{a}_{it} = a_{it} - T^{-1} \sum_{t=1}^{T} a_{it}$, $a = \{y, \boldsymbol{x}\}$ to concentrate out the individual fixed effects $\gamma_i^0$ (within-transformation).

In case of *PGMM*, the slope coefficients are derived as

$$\hat{\alpha}_k = \left( \left[ \sum_{i \in G_k} T^{-1} \sum_{t=1}^{T} \boldsymbol{z}_{it} \Delta \boldsymbol{x}_{it} \right]' \boldsymbol{W}_k \left[ \sum_{i \in G_k} T^{-1} \sum_{t=1}^{T} \boldsymbol{z}_{it} \Delta \boldsymbol{x}_{it} \right] \right)^{-1}$$
$$\left[ \sum_{i \in G_k} T^{-1} \sum_{t=1}^{T} \boldsymbol{z}_{it} \Delta \boldsymbol{x}_{it} \right]' \boldsymbol{W}_k \left[ \sum_{i \in G_k} T^{-1} \sum_{t=1}^{T} \boldsymbol{z}_{it} \Delta y_{it} \right],$$

where $\boldsymbol{W}_k$ is a $q \times q$ p.d. symmetric weight matrix and $\Delta$ denotes the first difference operator $\Delta \boldsymbol{x}_{it} = \boldsymbol{x}_{it} - \boldsymbol{x}_{it-1}$ (first-difference transformation).

## Value

An object of class `gplm` holding

| | |
|---|---|
| `model` | a `data.frame` containing the dependent and explanatory variables as well as cross-sectional and time indices, |
| `coefficients` | a $K \times p$ matrix of the group-specific parameter estimates, |
| `groups` | a `list` containing (i) the total number of groups $K$ and (ii) a vector of group memberships $(g_1, \dots, g_N)$, where $g_i = k$ if $i$ is assigned to group $k$, |
| `residuals` | a vector of residuals of the demeaned model, |
| `fitted` | a vector of fitted values of the demeaned model, |
| `args` | a `list` of additional arguments, |
| `IC` | a `list` containing (i) the value of the IC and (ii) the *MSE*, |
| `call` | the function call. |

A `gplm` object has `print`, `summary`, `fitted`, `residuals`, `formula`, `df.residual`, and `coef` S3 methods.

### Author(s)

Paul Haimerl

### References

Dhaene, G., & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3), 991-1030. doi:10.1093/restud/rdv007.

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 235(2), 1464-1482. doi:10.1016/j.jeconom.2022.12.002.

### Examples

```
# Simulate a panel with a group structure
set.seed(1)
sim <- sim_DGP(N = 20, n_periods = 80, p = 2, n_groups = 3)
y <- sim$y
X <- sim$X
groups <- sim$groups
df <- cbind(y = c(y), X)

# Estimate the grouped panel data model
estim <- grouped_plm(y ~ ., data = df, groups = groups, n_periods = 80, method = "PLS")
summary(estim)

# Lets pass a panel data set with explicit cross-sectional and time indicators
i_index <- rep(1:20, each = 80)
t_index <- rep(1:80, 20)
df <- data.frame(y = c(y), X, i_index = i_index, t_index = t_index)
estim <- grouped_plm(
  y ~ .,
  data = df, index = c("i_index", "t_index"), groups = groups, method = "PLS"
)
summary(estim)
```

---

## grouped_tv_plm *Grouped Time-varying Panel Data Model*

---

### Description

Estimate a time-varying panel data model subject to an observed group structure. Coefficient functions are homogeneous within groups but heterogeneous across groups. Time-varying coefficient functions are approximated as polynomial B-splines. The function supports both static and dynamic panel data models.

## Usage

```
grouped_tv_plm(
  formula,
  data,
  groups,
  index = NULL,
  n_periods = NULL,
  d = 3,
  M = floor(length(y)^(1/7) - log(p)),
  const_coef = NULL,
  rho = 0.04 * log(N * n_periods)/sqrt(N * n_periods),
  verbose = TRUE,
  parallel = TRUE,
  ...
)

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

## S3 method for class 'tv_gplm'
formula(x, ...)

## S3 method for class 'tv_gplm'
df.residual(object, ...)

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

## S3 method for class 'tv_gplm'
coef(object, ...)

## S3 method for class 'tv_gplm'
residuals(object, ...)

## S3 method for class 'tv_gplm'
fitted(object, ...)
```

## Arguments

`formula` a formula object describing the model to be estimated.

`data` a `data.frame` or `matrix` holding a panel data set. If no `index` variables are provided, the panel must be balanced and ordered in the long format $\boldsymbol{Y} = (Y_1', \dots, Y_N')'$, $Y_i = (Y_{i1}, \dots, Y_{iT})'$ with $Y_{it} = (y_{it}, \boldsymbol{x}_{it}')'$. Conversely, if `data` is not ordered or not balanced, `data` must include two index variables that declare the cross-sectional unit $i$ and the time period $t$ of each observation.

`groups` a numerical or character vector of length $N$ that indicates the group membership of each cross-sectional unit $i$.

| | |
|---|---|
| `index` | a character vector holding two strings. The first string denotes the name of the index variable identifying the cross-sectional unit $i$, and the second string represents the name of the variable declaring the time period $t$. The data is automatically sorted according to the variables in `index`, which may produce errors when the time index is a character variable. In case of a balanced panel data set that is ordered in the long format, `index` can be left empty if the number of time periods `n_periods` is supplied. |
| `n_periods` | the number of observed time periods $T$. If an `index` character vector is passed, this argument can be left empty. Default is `NULL`. |
| `d` | the polynomial degree of the B-splines. Default is 3. |
| `M` | the number of interior knots of the B-splines. If left unspecified, the default heuristic $M = \text{floor}((NT)^{\frac{1}{7}} - \log(p))$ is used, following Haimerl et al. (2025). Note that $M$ does not include the boundary knots and the entire sequence of knots is of length $M + d + 1$. |
| `const_coef` | a character vector containing the variable names of explanatory variables that enter with time-constant coefficients. |
| `rho` | the tuning parameter balancing the fitness and penalty terms in the IC. If left unspecified, the heuristic $\rho = 0.07\frac{\log(NT)}{\sqrt{NT}}$ of Haimerl et al. (2025) is used. We recommend the default. |
| `verbose` | logical. If `TRUE`, helpful warning messages are shown. Default is `TRUE`. |
| `parallel` | logical. If `TRUE`, certain operations are parallelized across multiple cores. Default is `TRUE`. |
| `...` | ellipsis |
| `object` | of class `tv_gplm`. |
| `x` | of class `tv_gplm`. |

## Details

Consider the time-varying panel data model

$$y_{it} = \gamma_i^0 + \boldsymbol{\beta}_i^{0\prime}(t/T)\boldsymbol{x}_{it} + \epsilon_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T,$$

where $y_{it}$ is the scalar dependent variable, $\gamma_i^0$ is an individual fixed effect, $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of explanatory variables, and $\epsilon_{it}$ is a zero mean error. The $p$-dimensional coefficient vector $\boldsymbol{\beta}_i^0(t/T)$ contains smooth functions of time and follows the observed group pattern

$$\boldsymbol{\beta}_i^0\left(\frac{t}{T}\right) = \sum_{k=1}^{K} \boldsymbol{\alpha}_k^0\left(\frac{t}{T}\right)\mathbf{1}\{i \in G_k\},$$

with $\cup_{k=1}^K G_k = \{1, \ldots, N\}$, $G_k \cap G_j = \emptyset$ for any $k \neq j$, $k, j = 1, \ldots, K$. The group structure $G_1, \ldots, G_K$ is determined by the argument `groups`.

The time-varying coefficient functions in $\boldsymbol{\alpha}_k(t/T)$ and, in turn, $\boldsymbol{\beta}_i(t/T)$ are estimated as polynomial B-splines. To this end, let $\boldsymbol{b}(v)$ denote a $M + d + 1$ vector of polynomial basis functions with the polynomial degree $d$ and $M$ interior knots. $\boldsymbol{\alpha}_k^0(t/T)$ is approximated by forming linear combinations of the basis functions $\boldsymbol{\alpha}_k^0(t/T) \approx \boldsymbol{\Xi}_k^{0\prime}\boldsymbol{b}(t/T)$, where $\boldsymbol{\Xi}_i^0$ is a group-specific $(M+d+1) \times p$ matrix of spline control points.

To estimate $\boldsymbol{\Xi}_k^0$, we project the explanatory variables onto the spline basis system, resulting in the $(M+d+1)p \times 1$ regressor vector $\boldsymbol{z}_{it} = \boldsymbol{x}_{it} \otimes \boldsymbol{b}(v)$. Subsequently, the DGP can be reformulated as

$$y_{it} = \gamma_i^0 + \boldsymbol{\pi}_i^{0\prime} \boldsymbol{z}_{it} + u_{it},$$

where $\boldsymbol{\pi}_i^0 = \text{vec}(\boldsymbol{\Pi}_i^0)$ and $\boldsymbol{\Xi}_k^0 = \boldsymbol{\Pi}_i^0$ if $i \in G_k$. $u_{it} = \epsilon_{it} + \eta_{it}$ collects the idiosyncratic $\epsilon_{it}$ and the sieve approximation error $\eta_{it}$. Then, we obtain $\hat{\boldsymbol{\xi}}_k = \text{vec}(\hat{\boldsymbol{\Xi}}_k)$ by

$$\hat{\xi}_k = \left( \sum_{i \in G_k} \sum_{t=1}^{T} \tilde{\boldsymbol{z}}_{it} \tilde{\boldsymbol{z}}_{it}^{\prime} \right)^{-1} \sum_{i \in G_k} \sum_{t=1}^{T} \tilde{\boldsymbol{z}}_{it} \tilde{y}_{it},$$

with $\tilde{a}_{it} = a_{it} - T^{-1} \sum_{t=1}^{T} a_{it}$, $a = \{y, \boldsymbol{z}\}$ to concentrate out the fixed effect $\gamma_i^0$ (within-transformation). Lastly, $\hat{\boldsymbol{\alpha}}_k(t/T) = \hat{\boldsymbol{\Xi}}_k^{\prime} \boldsymbol{b}(t/T)$. We refer to Haimerl et al. (2025, sec. 2) for more details.

In case of an unbalanced panel data set, the earliest and latest available observations per group define the start and end-points of the interval on which the group-specific time-varying coefficients are defined.

## Value

An object of class `tv_gplm` holding

| | |
|---|---|
| `model` | a `data.frame` containing the dependent and explanatory variables as well as cross-sectional and time indices, |
| `coefficients` | let $p^{(1)}$ denote the number of time-varying and $p^{(2)}$ the number of time constant coefficients. A `list` holding (i) a $T \times p^{(1)} \times K$ array of the group-specific functional coefficients and (ii) a $K \times p^{(2)}$ matrix of time-constant estimates. |
| `groups` | a `list` containing (i) the total number of groups $K$ and (ii) a vector of group memberships $G = (g_1, \dots, g_N)$, where $g_i = k$ if $i$ is part of group $k$, |
| `residuals` | a vector of residuals of the demeaned model, |
| `fitted` | a vector of fitted values of the demeaned model, |
| `args` | a `list` of additional arguments, |
| `IC` | a `list` containing (i) the value of the IC and (ii) the *MSE*, |
| `call` | the function call. |

An object of class `tv_gplm` has `print`, `summary`, `fitted`, `residuals`, `formula`, `df.residual`, and `coef` S3 methods.

## Author(s)

Paul Haimerl

## References

Haimerl, P., Smeekes, S., & Wilms, I. (2025). Estimation of latent group structures in time-varying panel data models. *arXiv preprint arXiv:2503.23165*. doi:10.48550/arXiv.2503.23165.

## Examples

```
# Simulate a time-varying panel with a trend and a group pattern
set.seed(1)
sim <- sim_tv_DGP(N = 10, n_periods = 50, intercept = TRUE, p = 2)
df <- data.frame(y = c(sim$y), X = sim$X)
groups <- sim$groups

# Estimate the time-varying grouped panel data model
estim <- grouped_tv_plm(y ~ ., data = df, n_periods = 50, groups = groups)
summary(estim)
```

---

pagfl *Pairwise Adaptive Group Fused Lasso*

---

## Description

Estimate a panel data model subject to a latent group structure using the pairwise adaptive group fused Lasso (*PAGFL*) by Mehrabani (2023). The *PAGFL* jointly identifies the group structure and group-specific slope parameters. The function supports both static and dynamic panels, with or without endogenous regressors.

## Usage

```
pagfl(
  formula,
  data,
  index = NULL,
  n_periods = NULL,
  lambda,
  method = "PLS",
  Z = NULL,
  min_group_frac = 0.05,
  bias_correc = FALSE,
  kappa = 2,
  max_iter = 10000,
  tol_convergence = 1e-08,
  tol_group = 0.001,
  rho = 0.07 * log(N * n_periods)/sqrt(N * n_periods),
  varrho = max(sqrt(5 * N * n_periods * p)/log(N * n_periods * p) - 7, 1),
  verbose = TRUE,
  parallel = TRUE,
  ...
)

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

```
## S3 method for class 'pagfl'
formula(x, ...)

## S3 method for class 'pagfl'
df.residual(object, ...)

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

## S3 method for class 'pagfl'
coef(object, ...)

## S3 method for class 'pagfl'
residuals(object, ...)

## S3 method for class 'pagfl'
fitted(object, ...)
```

### Arguments

`formula` a formula object describing the model to be estimated.

`data` a `data.frame` or `matrix` holding a panel data set. If no `index` variables are provided, the panel must be balanced and ordered in the long format $\boldsymbol{Y} = (Y_1', \dots, Y_N')'$, $Y_i = (Y_{i1}, \dots, Y_{iT})'$ with $Y_{it} = (y_{it}, \boldsymbol{x}_{it}')'$. Conversely, if `data` is not ordered or not balanced, `data` must include two index variables that declare the cross-sectional unit $i$ and the time period $t$ of each observation.

`index` a character vector holding two strings. The first string denotes the name of the index variable identifying the cross-sectional unit $i$ and the second string represents the name of the variable declaring the time period $t$. The data is automatically sorted according to the variables in `index`, which may produce errors when the time index is a character variable. In case of a balanced panel data set that is ordered in the long format, `index` can be left empty if the number of time periods `n_periods` is supplied.

`n_periods` the number of observed time periods $T$. If an `index` character vector is passed, this argument can be left empty. Default is `NULL`.

`lambda` the tuning parameter determining the strength of the penalty term. Either a single $\lambda$ or a vector of candidate values can be passed. If a vector is supplied, a BIC-type IC automatically selects the best fitting $\lambda$ value.

`method` the estimation method. Options are

- `"PLS"` for using the penalized least squares (*PLS*) algorithm. We recommend *PLS* in case of (weakly) exogenous regressors (Mehrabani, 2023, sec. 2.2).
- `"PGMM"` for using the penalized Generalized Method of Moments (*PGMM*). *PGMM* is required when instrumenting endogenous regressors, in which case a matrix $\boldsymbol{Z}$ containing the necessary exogenous instruments must be supplied (Mehrabani, 2023, sec. 2.3).

Default is `"PLS"`.

Z
: a $NT \times q$ `matrix` or `data.frame` of exogenous instruments, where $q \geq p$, $\boldsymbol{Z} = (z_1', \dots, z_N')'$, $z_i = (z_{i1}, \dots, z_{iT})'$ and $z_{it}$ is a $q \times 1$ vector. `Z` is only required when `method = "PGMM"` is selected. When using `"PLS"`, the argument can be left empty or it is disregarded. Default is `NULL`.

min_group_frac
: the minimum group cardinality as a fraction of the total number of individuals $N$. In case a group falls short of this threshold, each of its members is allocated to one of the remaining groups according to the *MSE*. Default is 0.05.

bias_correc
: logical. If `TRUE`, a Split-panel Jackknife bias correction following Dhaene and Jochmans (2015) is applied to the slope parameters. We recommend using the correction when working with dynamic panels. Default is `FALSE`.

kappa
: the a non-negative weight used to obtain the adaptive penalty weights. Default is 2.

max_iter
: the maximum number of iterations for the *ADMM* estimation algorithm. Default is $1 * 10^4$.

tol_convergence
: the tolerance limit for the stopping criterion of the iterative *ADMM* estimation algorithm. Default is $1 * 10^{-8}$.

tol_group
: the tolerance limit for within-group differences. Two individuals $i$, $j$ are assigned to the same group if the Frobenius norm of their coefficient vector difference is below this threshold. Default is $1 * 10^{-3}$.

rho
: the tuning parameter balancing the fitness and penalty terms in the IC that determines the penalty parameter $\lambda$. If left unspecified, the heuristic $\rho = 0.07 \frac{\log(NT)}{\sqrt{NT}}$ of Mehrabani (2023, sec. 6) is used. We recommend the default.

varrho
: the non-negative Lagrangian *ADMM* penalty parameter. For *PLS*, the $\varrho$ value is trivial. However, for *PGMM*, small values lead to slow convergence. If left unspecified, the default heuristic $\varrho = \max(\frac{\sqrt{5NTp}}{\log(NTp)} - 7, 1)$ is used.

verbose
: logical. If `TRUE`, helpful warning messages are shown. Default is `TRUE`.

parallel
: logical. If `TRUE`, certain operations are parallelized across multiple cores. Default is `TRUE`.

...
: ellipsis

x
: of class `pagfl`.

object
: of class `pagfl`.

## Details

Consider the panel data model

$$y_{it} = \gamma_i^0 + \boldsymbol{\beta}_i^{0\prime} \boldsymbol{x}_{it} + \epsilon_{it}, \quad i = 1, \dots, N, \; t = 1, \dots, T,$$

where $y_{it}$ is the scalar dependent variable, $\gamma_i^0$ is an individual fixed effect, $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of weakly exogenous explanatory variables, and $\epsilon_{it}$ is a zero mean error. The coefficient vector $\boldsymbol{\beta}_i^0$ follows the latent group pattern

$$\boldsymbol{\beta}_i^0 = \sum_{k=1}^{K} \boldsymbol{\alpha}_k^0 \boldsymbol{1}\{i \in G_k^0\},$$

with $\cup_{k=1}^K G_k^0 = \{1,\dots,N\}$, $G_k^0 \cap G_j^0 = \emptyset$ and $\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_j^0\| \neq 0$ for any $k \neq j$, $k, j = 1,\dots,K$.

The *PLS* method jointly estimates the latent group structure and group-specific coefficients by minimizing the criterion

$$Q_{NT}(\boldsymbol{\beta},\lambda) = \frac{1}{T}\sum_{i=1}^N\sum_{t=1}^T(\tilde{y}_{it} - \boldsymbol{\beta}_i'\tilde{\boldsymbol{x}}_{it})^2 + \frac{\lambda}{N}\sum_{i=1}^{N-1}\sum_{j=i}^N \dot{\omega}_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$$

with respect to $\boldsymbol{\beta} = (\boldsymbol{\beta}_1',\dots,\boldsymbol{\beta}_N')'$. $\tilde{a}_{it} = a_{it} - T^{-1}\sum_{t=1}^T a_{it}$, $a = \{y, \boldsymbol{x}\}$ to concentrate out the individual fixed effects $\gamma_i^0$ (within-transformation). $\lambda$ is the penalty tuning parameter and $\dot{\omega}_{ij}$ reflects adaptive penalty weights (see Mehrabani, 2023, eq. 2.6). $\|\cdot\|$ denotes the Frobenius norm. The adaptive weights $\dot{w}_{ij}$ are obtained by a preliminary individual least squares estimation. The criterion function is minimized via an iterative alternating direction method of multipliers (*ADMM*) algorithm (see Mehrabani, 2023, sec. 5.1).

*PGMM* employs a set of instruments `Z` to control for endogenous regressors. Using *PGMM*, $\boldsymbol{\beta}$ is estimated by minimizing

$$Q_{NT}(\boldsymbol{\beta},\lambda) = \sum_{i=1}^N\left[\frac{1}{N}\sum_{t=1}^T \boldsymbol{z}_{it}(\Delta y_{it} - \boldsymbol{\beta}_i'\Delta\boldsymbol{x}_{it})\right]' \boldsymbol{W}_i \left[\frac{1}{T}\sum_{t=1}^T \boldsymbol{z}_{it}(\Delta y_{it} - \boldsymbol{\beta}_i'\Delta\boldsymbol{x}_{it})\right]$$
$$+\frac{\lambda}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^N \ddot{\omega}_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|.$$

$\ddot{\omega}_{ij}$ are obtained by an initial *GMM* estimation. $\Delta$ gives the first differences operator $\Delta y_{it} = y_{it} - y_{it-1}$. $\boldsymbol{W}_i$ represents a data-driven $q \times q$ weight matrix. I refer to Mehrabani (2023, eq. 2.10) for more details. Again, the criterion function is minimized using an efficient *ADMM* algorithm (Mehrabani, 2023, sec. 5.2).

Two individuals are assigned to the same group if $\|\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\| \leq \epsilon_{\text{tol}}$ (and hence $\hat{\boldsymbol{\alpha}}_k = \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_j$ for some $k = 1,\dots,\hat{K}$), where $\epsilon_{\text{tol}}$ is determined by `tol_group`. Subsequently, the estimated number of groups $\hat{K}$ and group structure follows by examining the number of distinct elements in $\hat{\boldsymbol{\beta}}$. Given an estimated group structure, it is straightforward to obtain post-Lasso estimates using group-wise least squares or *GMM* (see `grouped_plm`).

We recommend identifying a suitable $\lambda$ parameter by passing a logarithmically spaced grid of candidate values with a lower limit close to 0 and an upper limit that leads to a fully homogeneous panel. A BIC-type information criterion then selects the best fitting $\lambda$ value.

## Value

An object of class `pagfl` holding

| | |
|---|---|
| `model` | a `data.frame` containing the dependent and explanatory variables as well as cross-sectional and time indices, |
| `coefficients` | a $\hat{K} \times p$ matrix of the post-Lasso group-specific parameter estimates, |
| `groups` | a `list` containing (i) the total number of groups $\hat{K}$ and (ii) a vector of estimated group memberships $(\hat{g}_1,\dots,\hat{g}_N)$, where $\hat{g}_i = k$ if $i$ is assigned to group $k$, |
| `residuals` | a vector of residuals of the demeaned model, |
| `fitted` | a vector of fitted values of the demeaned model, |

`args` a `list` of additional arguments,

`IC` a `list` containing (i) the value of the IC, (ii) the employed tuning parameter $\lambda$, and (iii) the *MSE*,

`convergence` a `list` containing (i) a logical variable indicating if convergence was achieved and (ii) the number of executed *ADMM* algorithm iterations,

`call` the function call.

A `pagfl` object has `print`, `summary`, `fitted`, `residuals`, `formula`, `df.residual`, and `coef` S3 methods.

## Author(s)

Paul Haimerl

## References

Dhaene, G., & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3), 991-1030. doi:10.1093/restud/rdv007.

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 235(2), 1464-1482. doi:10.1016/j.jeconom.2022.12.002.

## Examples

```
# Simulate a panel with a group structure
set.seed(1)
sim <- sim_DGP(N = 20, n_periods = 80, p = 2, n_groups = 3)
y <- sim$y
X <- sim$X
df <- cbind(y = c(y), X)

# Run the PAGFL procedure
estim <- pagfl(y ~ ., data = df, n_periods = 80, lambda = 0.5, method = "PLS")
summary(estim)

# Lets pass a panel data set with explicit cross-sectional and time indicators
i_index <- rep(1:20, each = 80)
t_index <- rep(1:80, 20)
df <- data.frame(y = c(y), X, i_index = i_index, t_index = t_index)
estim <- pagfl(
  y ~ .,
  data = df, index = c("i_index", "t_index"), lambda = 0.5, method = "PLS"
)
summary(estim)
```

sim_DGP *Simulate a Panel With a Group Structure in the Slope Coefficients*

## Description

Construct a static or dynamic, exogenous or endogenous panel data set subject to a group structure in the slope coefficients with optional $AR(1)$ or $GARCH(1,1)$ innovations.

## Usage

```
sim_DGP(
  N = 50,
  n_periods = 40,
  p = 2,
  n_groups = 3,
  group_proportions = NULL,
  error_spec = "iid",
  dynamic = FALSE,
  dyn_panel = lifecycle::deprecated(),
  q = NULL,
  alpha_0 = NULL
)
```

## Arguments

`N` the number of cross-sectional units. Default is 50.

`n_periods` the number of simulated time periods $T$. Default is 40.

`p` the number of explanatory variables. Default is 2.

`n_groups` the number of groups $K$. Default is 3.

`group_proportions` a numeric vector of length `n_groups` indicating size of each group as a fraction of $N$. If `NULL`, all groups are of size $N/K$. Default is `NULL`.

`error_spec` options include

- `"iid"` for $iid$ errors.
- `"AR"` for an $AR(1)$ error process with an autoregressive coefficient of 0.5.
- `"GARCH"` for a $GARCH(1,1)$ error process with a 0.05 constant, a 0.05 ARCH and a 0.9 GARCH coefficient.

Default is `"iid"`.

`dynamic` Logical. If `TRUE`, the panel includes one stationary autoregressive lag of $y_{it}$ as an explanatory variable (see sec. Details for more information on the $AR$ coefficient). Default is `FALSE`.

`dyn_panel` **[Deprecated]** deprecated and replaced by `dynamic`.

`q` the number of exogenous instruments when a panel with endogenous regressors is to be simulated. If panel data set with exogenous regressors is supposed to be generated, pass `NULL`. Default is `NULL`.

`alpha_0` a $K \times p$ matrix of group-specific coefficients. If `dynamic = TRUE`, the first column represents the stationary $AR$ coefficient. If `NULL`, the coefficients are drawn randomly (see sec. Details). Default is `NULL`.

## Details

The scalar dependent variable $y_{it}$ is generated according to the panel data model

$$y_{it} = \gamma_i + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + u_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T.$$

$\gamma_i$ represents individual fixed effects and $\boldsymbol{x}_{it}$ a $p \times 1$ vector of regressors. The individual slope coefficient vectors $\boldsymbol{\beta}_i$ follow the group structure

$$\boldsymbol{\beta}_i = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \mathbf{1}\{i \in G_k\},$$

with $\cup_{k=1}^{K} G_k = \{1, \dots, N\}$, $G_k \cap G_j = \emptyset$ and $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_j\| \neq 0$ for any $k \neq j$, $k, j = 1, \dots, K$. The total number of groups $K$ is determined by `n_groups`.

If a panel data set with exogenous regressors is generated (set `q = NULL`), the explanatory variables are simulated according to

$$x_{it,j} = 0.2\gamma_i + e_{it,j}, \quad \gamma_i, e_{it,j} \sim i.i.d. N(0, 1), \quad j = 1, \dots, p,$$

where $e_{it,j}$ denotes a series of innovations. $\gamma_i$ and $e_i$ are independent of each other.

In case `alpha_0 = NULL`, the group-level slope parameters $\boldsymbol{\alpha}_k$ are drawn from $\sim Unif[-2, 2]$.

If a dynamic panel is specified (`dynamic = TRUE`), the $AR$ coefficients $\beta_i^{\text{AR}}$ are drawn from a uniform distribution with support $(-1, 1)$ and $x_{it,j} = e_{it,j}$. Moreover, the individual fixed effects enter the dependent variable via $(1 - \beta_i^{\text{AR}})\gamma_i$ to account for the autoregressive dependency. We refer to Mehrabani (2023, sec 6) for details.

When specifying an endogenous panel (set `q` to $q \geq p$), the $e_{it,j}$ correlate with the cross-sectional innovations $u_{it}$ by a magnitude of 0.5 to produce endogenous regressors ($\text{E}(u|\boldsymbol{X}) \neq 0$). However, the endogenous regressors can be accounted for by exploiting the $q$ instruments in $\boldsymbol{Z}$, for which $\text{E}(u|\boldsymbol{Z}) = 0$ holds. The instruments and the first stage coefficients are generated in the same fashion as $\boldsymbol{X}$ and $\boldsymbol{\alpha}$ when `q = NULL`.

The function nests, among other, the DGPs employed in the simulation study of Mehrabani (2023, sec. 6).

## Value

A list holding

`alpha` the $K \times p$ matrix of group-specific slope parameters. If `dynamic = TRUE`, the first column holds the $AR$ coefficient.

`groups` a vector indicating the group memberships $(g_1, \dots, g_N)$, where $g_i = k$ if $i \in$ group $k$.

| | |
|---|---|
| y | a $NT \times 1$ vector of the dependent variable, with $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_N)'$, $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$ and the scalar $y_{it}$. |
| X | a $NT \times p$ matrix of explanatory variables, with $\boldsymbol{X} = (\boldsymbol{X}_1', \dots, \boldsymbol{X}_N')'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT})'$ and the $p \times 1$ vector $\boldsymbol{x}_{it}$. |
| Z | a $NT \times q$ matrix of instruments , where $q \geq p$, $\boldsymbol{Z} = (\boldsymbol{Z}_1', \dots, \boldsymbol{Z}_N')'$, $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \dots, \boldsymbol{z}_{iT})'$ and $\boldsymbol{z}_{it}$ is a $q \times 1$ vector. In case a panel with exogenous regressors is generated (q = NULL), $\boldsymbol{Z}$ equals NULL. |
| data | a $NT \times (p+1)$ data.frame of the outcome and $p$ explanatory variables. |

## Author(s)

Paul Haimerl

## References

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics*, 235(2), 1464-1482. doi:10.1016/j.jeconom.2022.12.002.

## Examples

```
# Simulate DGP 1 from Mehrabani (2023, sec. 6)
set.seed(1)
alpha_0_DGP1 <- matrix(c(0.4, 1, 1.6, 1.6, 1, 0.4), ncol = 2)
DGP1 <- sim_DGP(
  N = 50, n_periods = 20, p = 2, n_groups = 3,
  group_proportions = c(.4, .3, .3), alpha_0 = alpha_0_DGP1
)
```

---

# sim_tv_DGP *Simulate a Time-varying Panel With a Group Structure in the Slope Coefficients*

---

## Description

Construct a time-varying panel data set subject to a group structure in the slope coefficients with optional $AR(1)$ innovations.

## Usage

```
sim_tv_DGP(
  N = 50,
  n_periods = 40,
  intercept = TRUE,
  p = 1,
  n_groups = 3,
  d = 3,
  dynamic = FALSE,
```

```
  group_proportions = NULL,
  error_spec = "iid",
  locations = NULL,
  scales = NULL,
  polynomial_coef = NULL,
  sd_error = 1
)
```

## Arguments

`N` the number of cross-sectional units. Default is 50.

`n_periods` the number of simulated time periods $T$. Default is 40.

`intercept` logical. If `TRUE`, a time-varying intercept is generated.

`p` the number of simulated explanatory variables

`n_groups` the number of groups $K$. Default is 3.

`d` the polynomial degree used to construct the time-varying coefficients.

`dynamic` Logical. If `TRUE`, the panel includes one stationary autoregressive lag of $y_{it}$ as a regressor. Default is `FALSE`.

`group_proportions` a numeric vector of length `n_groups` indicating size of each group as a fraction of $N$. If `NULL`, all groups are of size $N/K$. Default is `NULL`.

`error_spec` options include

- `"iid"` for $iid$ errors.
- `"AR"` for an $AR(1)$ error process with an autoregressive coefficient of 0.5.

Default is `"iid"`.

`locations` a $p \times K$ matrix of location parameters of a logistic distribution function used to construct the time-varying coefficients. If left empty, the location parameters are drawn randomly. Default is `NULL`.

`scales` a $p \times K$ matrix of scale parameters of a logistic distribution function used to construct the time-varying coefficients. If left empty, the location parameters are drawn randomly. Default is `NULL`.

`polynomial_coef` a $p \times d \times K$ array of coefficients for the polynomials used to construct the time-varying coefficients. If left empty, the location parameters are drawn randomly. Default is `NULL`.

`sd_error` standard deviation of the cross-sectional errors. Default is 1.

## Details

The scalar dependent variable $y_{it}$ is generated according to the time-varying panel data model

$$y_{it} = \gamma_i + \boldsymbol{\beta}_i'(t/T)\boldsymbol{x}_{it} + u_{it}, \quad i = 1, \dots, N, \; t = 1, \dots, T,$$

where $\gamma_i$ is an individual fixed effect and $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of explanatory variables. The $p \times 1$ coefficient vector $\boldsymbol{\beta}_i(t/T)$ follows the group pattern

$$\boldsymbol{\beta}_i\left(\frac{t}{T}\right) = \sum_{k=1}^{K} \boldsymbol{\alpha}_k\left(\frac{t}{T}\right) \mathbf{1}\{i \in G_k\},$$

with $\cup_{k=1}^{K} G_k = \{1, \dots, N\}$ and $G_k \cap G_j = \emptyset$ for any $k \neq j$, $k, j = 1, \dots, K$. The total number of groups $K$ is determined by `n_groups`.

The predictors are simulated as:

$$x_{it,j} = 0.2\gamma_i + e_{it,j}, \quad \gamma_i, e_{it,j} \sim i.i.d.N(0,1), \quad j = \{1, \dots, p\},$$

where $e_{it,j}$ denotes a series of innovations. $\gamma_i$ and $e_i$ are independent of each other.

The errors $u_{it}$ feature a $iid$ standard normal distribution.

In case `locations = NULL`, the location parameters are drawn from $\sim Unif[0.3, 0.9]$. In case `scales = NULL`, the scale parameters are drawn from $\sim Unif[0.01, 0.09]$. In case `polynomial_coef = NULL`, the polynomial coefficients are drawn from $\sim Unif[-20, 20]$ and normalized so that all coefficients of one polynomial sum up to 1. The final coefficient function follows as $\boldsymbol{\alpha}_k(t/T) = 3 * F(t/T, location, scale) + \sum_{j=1}^{d} a_j(t/T)^j$, where $F(\cdot, location, scale)$ denotes a cumulative logistic distribution function and $a_j$ reflects a polynomial coefficient.

## Value

A list holding

| | |
|---|---|
| `alpha` | a $T \times p \times K$ array of group-specific time-varying parameters |
| `beta` | a $T \times p \times N$ array of individual time-varying parameters |
| `groups` | a vector indicating the group memberships $(g_1, \dots, g_N)$, where $g_i = k$ if $i \in$ group $k$. |
| `y` | a $NT \times 1$ vector of the dependent variable, with $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_N)'$, $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$ and the scalar $y_{it}$. |
| `X` | a $NT \times p$ matrix of explanatory variables, with $\boldsymbol{X} = (\boldsymbol{X}_1', \dots, \boldsymbol{X}_N')'$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT})'$ and the $p \times 1$ vector $\boldsymbol{x}_{it}$. |
| `data` | a $NT \times (p+1)$ data.frame of the outcome and the explanatory variables. |

## Author(s)

Paul Haimerl

## Examples

```
# Simulate a time-varying panel subject to a time trend and a group structure
set.seed(1)
sim <- sim_tv_DGP(N = 20, n_periods = 50, p = 1)
y <- sim$y
X <- sim$X
```

# Index