Package 'PAGFL'

June 8, 2024

Title Joint Estimation of Latent Groups and Group-Specific Coefficients in Panel Data Models

Version 1.1.0

Maintainer Paul Haimerl <p.haimerl@student.maastrichtuniversity.nl>

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.

License AGPL (>= 3)

Encoding UTF-8

RoxygenNote 7.3.1

LinkingTo Rcpp, RcppArmadillo, RcppParallel

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

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2024-06-08 21:20:02 UTC

Contents

`pagfl` *Pairwise Adaptive Group Fused Lasso*

Description

The pairwise adaptive group fused lasso (*PAGFL*) by Mehrabani (2023) jointly estimates the latent group structure and group-specific slope parameters in a panel data model. It can handle static and dynamic panels, either 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 = 5000,
  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 $Y_{it} = (y_{it}, x_{it}')'$. Conversely, if data is not ordered or not balanced, data must include two index variables, declaring the cross-sectional unit i and the time period t for each observation.
index	a character vector holding two strings specifying the variable names that identify the cross-sectional unit and time period for each observation. The first string denotes the individual unit, while the second string represents the time period. In case of a balanced panel data set that is ordered in the long format, index can be left empty if the 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.
lambda	the tuning parameter. λ governs the strength of the penalty term. Either a single λ or a vector of candidate values can be passed. If a vector is supplied, a BIC-type IC automatically selects the best fitting parameter 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 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. $\boldsymbol{Z}$ is only required when method = "PGMM" is selected. When using "PLS", either pass NULL or $\boldsymbol{Z}$ is disregarded. Default is NULL.
min_group_frac	the minimum group size as a fraction of the total number of individuals N. In case a group falls short of this threshold, a hierarchical classifier allocates its members to the remaining groups. 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 facing a dynamic panel. Default is FALSE.
kappa	the a non-negative weight placed on the adaptive penalty weights. Default is 2.
max_iter	the maximum number of iterations for the *ADMM* estimation algorithm. Default is 5000.
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 0.001.
rho	the tuning parameter balancing the fitness and penalty terms in the IC that determines the penalty parameter λ. 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 ϱ 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.
...	ellipsis
x	of class pagfl.
object	of class pagfl.

Details

Consider the grouped panel data model

$$y_{it} = \gamma_i + \beta_i' x_{it} + \epsilon_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T,$$

where y_{it} is the scalar dependent variable, γ_i is an individual fixed effect, x_{it} is a $p \times 1$ vector of explanatory variables, and ϵ_{it} is a zero mean error. The coefficient vector β_i is subject to the latent group pattern

$$\beta_i = \sum_{k=1}^{K} \alpha_k \mathbf{1}\{i \in G_k\},$$

with $\cup_{k=1}^{K} G_k = \{1, \dots, N\}$, $G_k \cap G_j = \emptyset$ and $\|\alpha_k\| \neq \|\alpha_j\|$ for any $k \neq M$.

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

$$\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \beta_i'\tilde{x}_{it})^2 + \frac{\lambda}{N}\sum_{1\leq i}\sum_{i<j\leq N} \dot{w}_{ij}\|\beta_i - \beta_j\|,$$

where $\tilde{y}_{it}$ is the demeaned scalar dependent variable, $\tilde{x}_{it}$ denotes a $p \times 1$ vector of demeaned weakly exogenous explanatory variables, λ is the penalty tuning parameter and $\dot{w}_{ij}$ 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 solution $\hat{\boldsymbol{\beta}}$ is computed via an iterative alternating direction method of multipliers (*ADMM*) algorithm (see Mehrabani, 2023, sec. 5.1).

PGMM employs a set of instruments $\boldsymbol{Z}$ to control for endogenous regressors. Using *PGMM*, $\boldsymbol{\beta} = (\beta_1', \dots, \beta_N')'$ is estimated by minimizing:

$$\sum_{i=1}^{N} \left[\frac{1}{N} \sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it}) \right]' W_i \left[\frac{1}{T} \sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i' \Delta x_{it}) \right] + \frac{\lambda}{N} \sum_{1 \leq i} \sum_{i < j \leq N} \ddot{w}_{ij} \|\beta_i - \beta_j\|.$$

$\ddot{w}_{ij}$ are obtained by an initial *GMM* estimation. Δ gives the first differences operator $\Delta y_{it} = y_{it} - y_{it-1}$. W_i represents a data-driven $q \times q$ weight matrix. I refer to Mehrabani (2023, eq. 2.10) for more details. $\boldsymbol{\beta}$ is again estimated employing an efficient *ADMM* algorithm (Mehrabani, 2023, sec. 5.2).

Two individuals are assigned to the same group if $\|\hat{\beta}_i - \hat{\beta}_j\| \leq \epsilon_{\text{tol}}$, where ϵ_{tol} is given by `tol_group`. Subsequently, the number of groups follows as the number of distinct elements in $\hat{\boldsymbol{\beta}}$. Given an estimated group structure, it is straightforward to obtain post-Lasso estimates using least squares.

We suggest identifying a suitable λ parameter by passing a logarithmically spaced grid of candidate values with a lower limit of 0 and an upper limit that leads to a fully homogeneous panel. A BIC-type information criterion then selects the best fitting λ 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 $K \times p$ 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 λ, and (iii) the mean squared error,
`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
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 Latent Group Structure*

Description

Construct a static or dynamic, exogenous or endogenous panel data set subject to a latent group structure 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 latent groups K. Default is 3.
group_proportions	a numeric vector of length n_groups indicating the fraction of N each group will contain. 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 a regressor (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	an optional pre-specified $K \times p$ coefficient matrix. 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 following grouped panel data model

$$y_{it} = \gamma_i + \beta_i' x_{it} + u_{it}, \quad i = \{1, \dots, N\}, \quad t = \{1, \dots, T\}.$$

γ_i represents individual fixed effects and x_{it} a $p \times 1$ vector of regressors. The individual slope coefficient vectors β_i are subject to a latent group structure

$$\beta_i = \sum_{k=1}^{K} \alpha_k \mathbf{1}\{i \in G_k\},$$

where $K =$ n_groups. As a consequence, the group-level coefficients $\boldsymbol{\alpha} = (\alpha_1', \dots, \alpha_K')'$ follow the partition $\boldsymbol{G}$ of N cross-sectional units $\boldsymbol{G} = (G_1, \dots, G_K)$ such that $\cup_{k=1}^{K} = \{1, \dots, N\}$ and $G_k \cap G_l = \emptyset,\ \alpha_k \neq \alpha_l$ for any two groups $k \neq l$ (Mehrabani, 2023, sec. 2.1).

If a panel data set with exogenous regressors is generated (set `q = NULL`), the p 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. γ_i and e_i are independent of each other.

In case `alpha_0 = NULL`, the group-level slope parameters α_k are drawn from $\sim U[-2, 2]$.

If a dynamic panel is specified (`dynamic = TRUE`), the AR coefficients β_i^{AR} are drawn from a uniform distribution with support $(-1, 1)$ and $x_{it,j} = e_{it,j}$. The individual fixed effects enter the dependent variable via $(1 - \beta_i^{\text{AR}})\gamma_i$ to account for the autoregressive dependency. I 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 with $\text{E}(u|X) \neq 0$. However, the endogenous regressors can be accounted for by exploiting the q instruments in $\boldsymbol{Z}$, for which $\text{E}(u|Z) = 0$ 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. In case of `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 $\boldsymbol{y} = (y_1, \dots, y_N)'$, $y_i = (y_{i1}, \dots, y_{iT})'$ and the scalar y_{it}.
`X`	a $NT \times p$ matrix of explanatory variables, with $\boldsymbol{X} = (x_1, \dots, x_N)'$, $x_i = (x_{i1}, \dots, x_{iT})'$ and the $p \times 1$ vector x_{it}.
`Z`	a $NT \times q$ matrix of 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. 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 the 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)
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 Latent Group Structure*

Description

Construct a time-varying panel data set subject to a latent group structure.

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,
  DGP = lifecycle::deprecated()
)
```

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 latent 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 the fraction of N each group will contain. 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 a 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.

`DGP`
: **[Deprecated]** the data generating process. Options are

  1 generates a trend only.

  2 simulates a trend and an additional exogenous explanatory variable.

  1 draws a dynamic panel data model with one AR lag.

Details

The scalar dependent variable y_{it} is driven by the following panel data model:

$$y_{it} = \gamma_i + \beta_{it}' x_{it} + u_{it}, \quad i = 1, \ldots, N, \; t = 1, \ldots, T,$$

where y_{it} is the scalar dependent variable, γ_i is an individual fixed effect and x_{it} is a $p \times 1$ vector of explanatory variables. The errors u_{it} feature a iid standard normal distribution. The coefficient vector $\beta_i = \{\beta_{i1}', \ldots, \beta_{iT}'\}'$ is subject to the group pattern

$$\beta_i \left(\frac{t}{T} \right) = \sum_{k=1}^{K} \alpha_k \left(\frac{t}{T} \right) \mathbf{1}\{i \in G_k\},$$

with $K =$`n_groups`, $\cup_{k=1}^{K} G_k = \{1, \ldots, N\}$, $G_k \cap G_j = \emptyset$ and $\|\alpha_k\| \neq \|\alpha_j\|$ for any $k \neq j$.

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

$$y_{it} = \gamma_i + \beta_i'(t/T) x_{it} + u_{it}, \quad i = \{1, \ldots, N\}, \quad t = \{1, \ldots, T\}.$$

γ_i represents individual fixed effects and x_{it} a $p \times 1$ vector of regressors. The individual functional slope coefficient vectors $\beta_i(t/T)$ are subject to a latent group structure $\beta_i(t/T) = \sum_{k=1}^{K} \alpha_k(t/T) \mathbf{1}\{i \in G_k\}$. As a consequence, the group-level coefficients $\boldsymbol{\alpha}(t/T) = (\alpha_1'(t/T), \ldots, \alpha_K'(t/T))'$ follow

the partition $\boldsymbol{G}$ of N cross-sectional units $\boldsymbol{G} = (G_1, \ldots, G_K)$ such that $\cup_{k=1}^{K} = \{1, \ldots, N\}$ and $G_k \cap G_l = \emptyset,\ \alpha_k \neq \alpha_l$ for any two groups $k \neq l$.

The predictors are simulated as:

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

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

In case `locations = NULL`, the location parameters are drawn from $\sim U[0.3, 0.9]$. In case `scales = NULL`, the scale parameters are drawn from $\sim U[0.01, 0.09]$. In case `polynomial_coef = NULL`, the polynomial coefficients are drawn from $\sim U[-20, 20]$ and normalized so that all coefficients of one polynomial sum up to 1. The final coefficient function follows as $\alpha_k(t/T) = 3{*}F(t/T, location, scale) + \sum_{j=1}^{d} a_j (t/T)^j$, 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 $(g_1, \ldots, g_N)$, where $g_i = k$ if $i \in$ group k.
`y`	a $NT \times 1$ vector of the dependent variable, with $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $y_i = (y_{i1}, \ldots, y_{iT})'$ and the scalar y_{it}.
`X`	a $NT \times p$ matrix of explanatory variables, with $\boldsymbol{X} = (x_1, \ldots, x_N)'$, $x_i = (x_{i1}, \ldots, x_{iT})'$ and the $p \times 1$ vector 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 latent group structure
sim <- sim_tv_DGP(N = 20, n_periods = 50, intercept = TRUE, p = 1)
y <- sim$y
```

tv_pagfl *Time-varying Pairwise Adaptive Group Fused Lasso*

Description

The time-varying pairwise adaptive group fused lasso (time-varying *PAGFL*) jointly estimates the latent group structure and group-specific time-varying functional coefficients in a panel data model. The time-varying coefficients are modeled as polynomial B-splines.

Usage

```
tv_pagfl(
  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 = 20000,
  tol_convergence = 1e-10,
  tol_group = 0.001,
  rho = 0.07 * log(N * n_periods)/sqrt(N * n_periods),
  varrho = 1,
  verbose = TRUE,
  parallel = TRUE,
  ...
)

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

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

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

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

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

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

## S3 method for class 'tvpagfl'
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 $\boldsymbol{Y} = (Y_1', \ldots, Y_N')'$, $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ with $Y_{it} = (y_{it}, x_{it}')'$. Conversely, if data is not ordered or not balanced, data must include two index variables, declaring the cross-sectional unit i and the time period t for each observation.
index	a character vector holding two strings specifying the variable names that identify the cross-sectional unit and the time period for each observation. The first string denotes the individual unit, while the second string represents the time period. In case of a balanced panel data set that is ordered in the long format, index can be left empty if the the number of time periods n_periods is supplied. Default is Null.
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. λ governs the strength of the penalty term. Either a single λ or a vector of candidate values can be passed. If a vector is supplied, a BIC-type IC automatically selects the best fitting parameter 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 $M = \text{floor}((NT)^{\frac{1}{7}} - \log(p))$ is used. Note that M does not include the boundary knots.
min_group_frac	the minimum group size as a fraction of the total number of individuals N. In case a group falls short of this threshold, a hierarchical classifier allocates its members to the remaining groups. Default is 0.05.
const_coef	a character vector containing the variable names of explanatory variables that are estimated with time-constant coefficients. All of concerning regressors must be named variables in data.
kappa	the a non-negative weight placed on the adaptive penalty weights. Default is 2.
max_iter	the maximum number of iterations for the *ADMM* estimation algorithm. Default is 20,000.
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 0.001.

rho	the tuning parameter balancing the fitness and penalty terms in the IC that determines the penalty parameter λ. 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 the employed penalized sieve estimation *PSE*, the ϱ value is trivial. 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.
...	ellipsis
object	of class tvpagfl.
x	of class tvpagfl.

Details

Consider the grouped time-varying panel data model

$$y_{it} = \gamma_i + \beta_i'(t/T)x_{it} + \epsilon_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T,$$

where y_{it} is the scalar dependent variable, γ_i is an individual fixed effect, x_{it} is a $p \times 1$ vector of explanatory variables, and ϵ_{it} is a zero mean error. The coefficient vector $\beta_i(t/T)$ is subject to the group pattern

$$\beta_i\left(\frac{t}{T}\right) = \sum_{k=1}^{K} \alpha_k\left(\frac{t}{T}\right)\mathbf{1}\{i \in G_k\},$$

with $\cup_{k=1}^{K} G_k = \{1, \dots, N\}$, $G_k \cap G_j = \emptyset$ and $\|\alpha_k\| \neq \|\alpha_j\|$ for any $k \neq M$. $\beta_i(t/T)$, and $\alpha_k(t/T)$ are estimated as polynomial B-splines using penalized sieve-technique. Let $\boldsymbol{B}(v)$ denote a $M + d + 1$ vector basis functions, where d denotes the polynomial degree and M the number of interior knots. Then, $\beta_i(t/T)$ and $\alpha_i(t/T)$ are approximated as $\beta_i(t/T) = \pi_i'\boldsymbol{B}(t/T)$ and $\alpha_i(t/T) = \xi_i'\boldsymbol{B}(t/T)$, respectively. π_i and ξ_i are $(M+d+1) \times p$ coefficient matrices which weigh the individual basis functions. The explanatory variables are projected onto the spline basis system, which results in the $(M + d + 1) * p \times 1$ vector $z_{it} = x_{it} \otimes \boldsymbol{B}(v)$. Subsequently, the DGP can be reformulated as

$$y_{it} = \gamma_i + z_{it}'\text{vec}(\pi_i) + u_{it},$$

where $u_{it} = \epsilon_{it} + \eta_{it}$ and η_{it} contains the sieve approximation error. I refer to Su et al. (2019, sec. 2) for more details on the sieve technique.

Inspired by Su et al. (2019) and Mehrabani (2023), the time-varying PAGFL estimates the functional coefficients and the group structure by minimizing the criterion:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{z}_{it}'\text{vec}(\pi_i))^2 + \frac{\lambda}{N}\sum_{1\leq i}\sum_{i<j\leq N}\dot{w}_{ij}\|\text{vec}(\pi_i - \pi_j)\|,$$

where $\tilde{y}_{it}$ is the demeaned dependent variable, and $\tilde{z}_{it}$ is likewise demeaned to concentrate out the individual fixed effects γ_i. λ is the penalty tuning parameter and $\dot{w}_{ij}$ denotes adaptive penalty weights which are obtained by a preliminary non-penalized estimation. $\|\cdot\|$ represents the Frobenius norm. The solution $\hat{\boldsymbol{\beta}}$ is computed via the iterative alternating direction method of multipliers

(*ADMM*) algorithm proposed in Mehrabani (2023, sec. 5.1), adapted to accommodate the B-spline coefficient functions.

Two individuals are assigned to the same group if $\|\text{vec}(\hat{\pi}_i - \hat{\pi}_j)\| \leq \epsilon_{\text{tol}}$, where ϵ_{tol} is given by `tol_group`. Subsequently, the number of groups follows as the number of distinct elements in $\hat{\boldsymbol{\beta}}$. Given an estimated group structure, it is straightforward to obtain post-Lasso estimates using least squares.

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

In case of an unbalanced panel data set, the earliest and latest available observations out of the entire panel are employed as the start and end-points of the interval on which the time-varying coefficients are defined.

Value

An object of class `tvpagfl` holding

`model`	a `data.frame` containing the dependent and explanatory variables as well as individual and time indices,
`coefficients`	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. Let $p^{(1)}$ denote the number of time-varying coefficients and $p^{(2)}$ the number of time constant parameters,
`groups`	a `list` containing (i) the total number of groups $\hat{K}$ and (ii) a vector of estimated group memberships $(\hat{g}_1, \ldots, \hat{g}_N)$, 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 λ, and (iii) the mean squared error,
`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 `tvpagfl` 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.

Su, L., Wang, X., & Jin, S. (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, 37(2), 334-349. doi:10.1080/07350015.2017.1340299.

Examples

```
# Simulate a time-varying panel with a trend and a group pattern
set.seed(1)
sim <- sim_tv_DGP(N = 5, n_periods = 20, intercept = TRUE, p = 1)
df <- data.frame(y = c(sim$y))

# Run the time-varying PAGFL with only an intercept
estim <- tv_pagfl(y ~ 1, data = df, n_periods = 20, lambda = 13, max_iter = 100, parallel = FALSE)
summary(estim)
```

Index